

Adaptation, Not Translation: An Architecture and Evaluation Framework for Long-Form Cross-Cultural Story Localization

Taaha Kazi, Aditya Sanjiv Kanade, Ivory Yang, Mohammad Saifullah, Vasu Sharma, Bhavesh Pant, Kunal Bhatia, Abhinav Prakash, Abdur Rahman, Varun Agarwal, Sarjita Jain, Viveca Pavate, Nishita Maithul

Pocket FM, Bengaluru, India

{taaha.kazi, kanade.aditya, ivory.yang, mohammad.saifullah, vasu.sharma, bhavesh.pant, kunal.bhatia, abhinav.prakash, rahman.abdur, varun.agarwal, sarjita.jain, viveca.pavate, nishita.maithul}@pocketfm.com

Abstract

Translating a story renders its words into a new *language*; adapting it transposes its world into a new *culture*. Large Language Models (LLMs) translate and adapt individual passages effectively, but long-form novels spanning ~ 100 chapters suffer model failures. We formalize *long-form cross-cultural story adaptation* as a generation task parameterized by source and target language-culture pairs, with three structural axes: **A1** identity coherence across narrative arc, **A2** cultural-system alignment within scenes, and **A3** dependency-chain integrity among related entities. This is addressed through a five-stage agentic adaptation framework comprising entity registry, target-culture plan, dependency-aware batched localization, validation and rewrite. Evaluation is conducted through ablations across three language-culture pairs and three model families; the proposed system outperforms across all three axes, reducing identity errors by 65%, dependency errors by 64%, and cultural misalignment errors by 27% against competitive baselines. Further validation through a production A/B test on a commercial storytelling platform demonstrates listener retention comparable to that for human-adapted content, reflecting viability of the framework.

1 Introduction

When a Chinese family drama is adapted for an American audience, two operations are conducted. *Translation* renders the words in English: “*During Spring Festival, the family pasted up spring couplets, wrapped dumplings, watched the Spring Gala, and the children received red envelopes.*” *Adaptation* transposes the cultural world; Spring Festival becomes Christmas Day, couplets become a wreath, wrapping dumplings change to baking cookies, Spring Gala turns into a holiday parade,



Figure 1: Translation vs. Adaptation on a Chinese New Year passage. Translation preserves source-culture items (Spring Festival, couplets, etc) while adaptation swaps each for target-culture analogue that preserves narrative function (Christmas Day, wreath, etc).

and receiving red envelopes becomes hanging up stockings. Translation changes the language, adaptation changes its cultural world (Fig. 1).

Which of these two operations a story receives is a production decision with measurable consequences. The target deployment is a serialized audio-fiction platform that runs two distinct production pipelines: a translation-only pipeline for language pairs with sufficient cultural proximity, and the cultural-adaptation pipeline described in this paper for pairs with large cultural distance (e.g., Mandarin-origin serialized fiction prepared for US English-speaking listeners). Internal engagement data shows that culturally distant content produced via translation alone yields significantly

Five-Stage Localization Pipeline

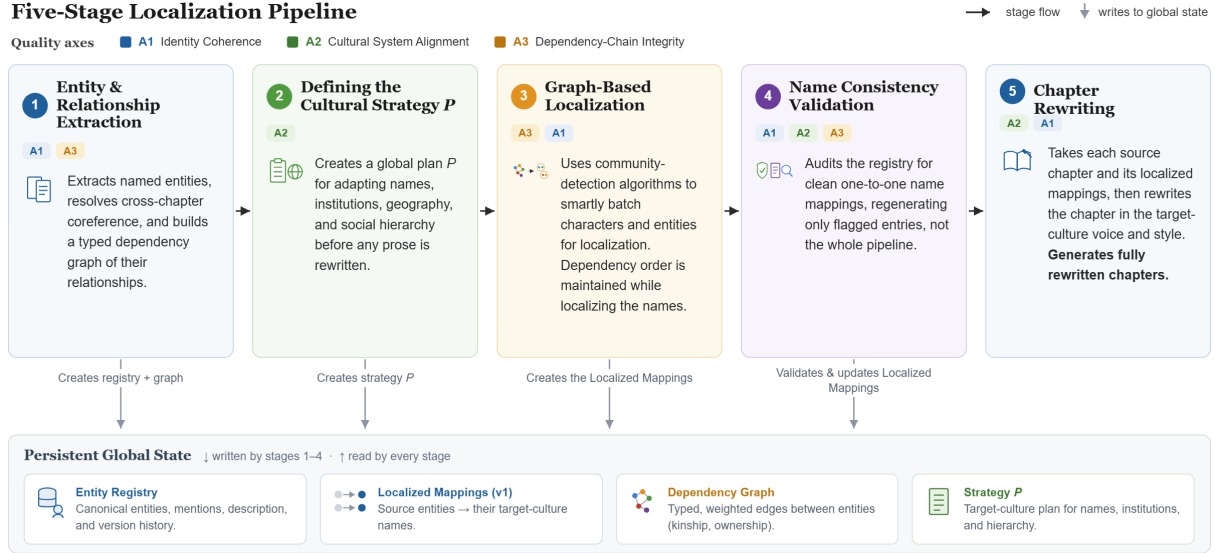


Figure 2: **Five-stage localization pipeline** Stages 1-4 progressively build a persistent global state: an entity registry, a dependency graph, and a target-culture strategy P . Every stage reads and updates the global state, keeping entities, culture, and dependencies coherent (axes A1-A3) across the entire story. Stage 5 then rewrites each chapter in the target-culture voice using the validated mappings and strategy P .

higher listener drop-off than adapted content; that operational gap motivates the distinction this paper formalizes.

Large Language Models (LLMs) translate and adapt effectively for individual passages (Yao et al., 2024; Singh et al., 2024), yet long-form novels spanning ~ 100 chapters exhibit recurring plot inconsistencies despite larger context windows or stronger decoders. We formalize *long-form cross-cultural story adaptation* as a generation task parameterized by source language-culture pair (L_S, C_S) and target language-culture pair (L_T, C_T) across three structural axes: **A1** identity coherence, **A2** cultural-system alignment, and **A3** dependency-chain integrity. A five-stage agentic adaptation framework externalizes narrative state through entity tracking, cultural planning, dependency-aware localization, validation and rewriting. Ablations across three language-culture pairs and three model families show reduction in identity errors by 65%, dependency errors by 64%, and cultural misalignment errors by 27% against competitive baselines. Production A/B test on three Mandarin-to-English-USA serialized shows demonstrates listener retention comparable to that for human-adapted content, reflecting viability of real-world deployment.

2 Related Work

Translation studies has long recognized adaptation beyond linguistic transfer, reflected in Skopos

theory (Reiss and Vermeer, 1984), domestication and foreignization (Yang, 2010), and transcreation (Pedersen, 2014). Machine Translation has improved consistency beyond sentence level through document context (Karpinska and Iyyer, 2023), coreference modeling (Vu et al., 2024), document-level evaluation (Jiang et al., 2022). Long-context training (Gao et al., 2025) extends the usable window further; however, these approaches target translation fidelity rather than systemic cross-cultural transformation, and treat consistency as a contextual retrieval problem.

Growing research on cross-cultural adaptation and localization (Herscovich et al., 2022; Liu et al., 2025) includes culturally grounded translation (Anik et al., 2025), content adaptation (Singh et al., 2024; Cao et al., 2024), and cross-cultural evaluation (AlKhamissi et al., 2024). However, existing approaches operate at passage or document level (Singh et al., 2024; Yao et al., 2024), and do not address the long-range coordination needed for consistency across narratives spanning ~ 100 chapters. The most directly relevant prior work operates at *atomic* scale: Cao et al. (2024) adapt individual recipes cross-culturally, Khanuja et al. (2024) extend adaptation to individual images for cross-cultural relevance, and Khanuja et al. (2025) propose automatic evaluation metrics for image transcreation. These artifacts share a defining property: each is self-contained, with no recurring identity, no dependent sibling entities, and no narrative

continuity, so the techniques do not lift to long-form output (detailed positioning in Appendix C). To our knowledge, this is the first work to formalize *long-form cross-cultural story adaptation* as a generation task, with explicit structural axes and an aligned architectural decomposition.

3 Long-Form Cross-Cultural Adaptation

Long-form cross-cultural story adaptation maps a source chapter sequence to a target chapter sequence, conditioned on source and target language-culture pairs. $S = (c_1, \dots, c_N)$ denotes a source story of N chapters (typically $N \gtrsim 10^2$) of source language-culture (L_S, C_S) , for target language-culture (L_T, C_T) . Long-form cross-cultural story adaptation seeks a target story

$$S' = (c'_1, \dots, c'_N) = f(S, L_S, C_S, L_T, C_T), \quad (1)$$

that preserves narrative while rendering the story native to the target culture. Equation (1) defines a chapter-level *ordering contract*: source chapter c_i maps to target chapter c'_i , preserving the narrative arc across all N chapters. It does not constrain within-chapter scene structure or prose phrasing to a one-to-one rendering; each target chapter is fully re-generated, conditioned on the cultural strategy and the localized entity registry, and the model is free to rephrase within the chapter. Preservation includes plot structure, character arcs, and causal dependencies; adaptation includes culturally appropriate entities, institutions, settings, and social conventions. The three axes are jointly necessary.

A1 – Identity coherence. Localized entities e' in S' must remain consistent across the narrative arc $c'_1, \dots, c'_N$, including names, aliases, honorifics, and references. Identity coherence is evaluated globally across chapters rather than passages.

A2 – Cultural-system alignment. Cultural artifacts a appearing in target-culture scenes S' must be coherent with target culture unless explicitly presented as foreign, ensuring adaptation beyond linguistic translation to cultural localization. We evaluate a domestication-oriented localization regime; foreignization-preserving adaptation is valid but outside this metric.

A3 – Dependency-chain integrity. Semantically related entities must be adapted consistently.

Changes to a parent entity (e.g., a family name, institution, or location) must propagate to dependent entities that inherit or reference it.

4 System Architecture

We propose a five-stage pipeline (see Fig. 2) that maintains story-level state through shared entity registries, relationship graphs, and a global adaptation plan.

Stage 1: Entity and Relationship Extraction. Before adaptation, the system extracts named entities, resolves coreferences, and constructs a relationship graph over the entire story. The resulting entity registry and dependency graph provide a persistent narrative state, avoiding common document-level inconsistency issues, improving identity coherence (A1) and supporting dependency-aware localization (A3).

Stage 2: Defining the Cultural Strategy P . The pipeline then constructs a global cultural strategy P , specifying how names, institutions, locations, and social structures are localized to target culture. This plan anchors the narrative to a coherent target culture (A2) and guides subsequent localization and rewriting decisions.

Stage 3: Graph-Based Localization. Entities are localized in topological order over the dependency graph from Stage 1. Related characters are grouped through Louvain community detection (Blondel et al., 2008), allowing the model to adapt closely tied characters (like a family) in the same batch, which outperforms simple topological sorting or greedy batching.

Stage 4: Name Consistency Validation. Before rewriting, the localized entity registry is audited to ensure one-to-one mappings between source and target entities, and to detect duplicate localized names. Conflicts are resolved by regenerating only the affected mappings, preserving consistency without rerunning the entire pipeline. The one-to-one constraint is a consistency constraint on the *entity registry*, not a restriction on lexical choice in the generated prose: (i) each source entity receives exactly one canonical target-side form, so the same character is never rendered under multiple names across chapters, and (ii) no two distinct source entities are assigned the same target name, so different characters cannot be conflated. Source cultural concepts may admit several

plausible target analogues; the strategy P selects among them by *narrative-functional equivalence* (a family-reunion Mid-Autumn Festival scene may map to Thanksgiving, whereas a story trading on nocturnal, supernatural atmosphere may map the same festival to Halloween), and the selected mapping is then applied consistently for the remainder of the work. A worked conflict-detection example is given in Appendix A.11.

Stage 5: Chapter Rewriting. Lastly, the pipeline rewrites each chapter to match the target culture’s voice and style. The model is provided with the original chapter, localized entity mappings from Stage 3, and cultural strategy P . Localized names are injected directly into the prompt to prevent reverting to source names or hallucinations. Here, the actual cultural alignment takes shape, extending prior work on atomic-scale (recipe, menu, passage, image) adaptation (Cao et al., 2024; Zhang et al., 2024; Yao et al., 2024; Singh et al., 2024; Khanuja et al., 2024, 2025) to long-form narratives (A2).

Error containment. Stage 1 errors cannot silently propagate: Stage 4 audits the registry before any rewriting, detecting conflicts (including inconsistently localized related entities) and regenerating only the affected mappings, and the validators of §5.1 act as a production safety barrier, surfacing only below-threshold chapters to a human reviewer with each flagged issue’s location and category. An extraction error thus stops at a Stage 4 correction.

5 Evaluation Framework

5.1 Automated Evaluation

The automated evaluation scores the three structural axes defined in Section 3 using LLM-as-a-judge pipelines. For each variant v , an initial extraction step maps how recurring characters, named entities, and dependent entities are realized in the adapted text. A validation step then scores these mappings for identity stability (m_{A1}) and coherent dependency localization (m_{A3}). To measure cultural alignment (m_{A2}), an LLM judge reads the adapted scenes and identifies out-of-place source-culture artifacts, reporting the number of confirmed cultural errors per chapter. The judge pipeline is validated against independent native-speaker annotation on a 50-chapter subset, reaching 92% precision and 97% recall against the

annotator consensus (Appendix A.5).

Per-variant counts. For variant v , let I_v be the number of validated recurring character/entity items, $C_v^{(1)}$ the number with stable identity mappings across chapters, $C_v^{(3)}$ the number with coherent related-entity localization (i.e., dependent entities localized consistently with their heads), $F_v^{(2)}$ the number of cultural misalignment errors (out-of-place source-culture artifacts), and H_v the number of evaluated chapters.

Axis scores. The three axes are reported as

$$\begin{aligned} m_{A1}(v) &= \frac{I_v - C_v^{(1)}}{I_v}, \\ m_{A2}(v) &= \frac{F_v^{(2)}}{H_v}, \\ m_{A3}(v) &= \frac{I_v - C_v^{(3)}}{I_v}. \end{aligned}$$

These are the identity error rate, cultural-misalignment rate per chapter, and dependency error rate respectively; *lower is better* for all three.

Plot Fidelity (m_{A0}). We verify that the adaptation process preserves the core plot. For each source and adapted chapter pair, an LLM judge conducts a beat-level analysis and produces several sub-scores (*beat recall*, *beat precision*, *sequence fidelity*, *character fidelity*), culminating in a holistic plot-fidelity score in $[0, 1]$. Variant-level score is the unweighted chapter mean,

$$m_{A0}(v) = \frac{1}{H_v} \sum_{h=1}^{H_v} s_{v,h}^{(0)},$$

where $s_{v,h}^{(0)}$ is the holistic score of chapter h under variant v . This auxiliary metric ensures that improvements in cultural adaptation do not come at the expense of plot fidelity.

5.2 Human Evaluation: Production A/B Test

The A/B test provides a real-world benchmark by comparing our system directly against a professional human adapter. We deployed three Mandarin-to-English (US) web novels on a commercial platform. For each show, half of the readers received a version drafted by our pipeline (Variant A), while the other received a version drafted end-to-end by a human professional (Variant B). Both versions underwent the platform’s standard

human validation step before publication, but Variant A began with the P_{full} output and received a lightweight human review guided by validator flags, while Variant B was drafted by a professional adapter. Listener retention was measured hourly, and the results were reported descriptively given the limited scope of the study (Section 6.4).

6 Experiments

6.1 Datasets

Evaluations are conducted across three language-culture adaptation pairs: Mandarin→English (US), English→German, and English→Hindi. The set contains three Mandarin→English shows and two English-source shows adapted into German and Hindi; each show spans ~ 100 chapters.

6.2 Experiment Variants

We compare four experiment variants: each receives the same source story and target language-culture pair, but differs in how cross-chapter localization state is represented during adaptation.

B1: plan-then-adapt. B1 is a non-pipeline planning baseline. The model reads the full story and produces a global target-culture strategy specifying how names, places, institutions, social roles, and recurring cultural references are adapted. As this strategy is fixed before rewriting, chapters are adapted independently and in parallel from the chapter text conditioned on the global plan. B1 is thus the strongest single-workflow alternative (same Stage 2 plan schema; sample in Appendix A.10) but keeps no per-entity state: an alias absent from the plan has no canonical record at rewrite time, and the $>300\text{K}$ -token source exceeds reliable single-prompt state management.

B2: rolling summary. B2 is a sequential memory baseline. The model adapts the story chapter by chapter while maintaining a bounded running summary of prior plot events and localization choices. At chapter k , the model receives the current source chapter together with the summary accumulated from chapters $1, \dots, k-1$.

S1: early architecture variant. S1 is an intermediate variant of the architecture in Section 4. It extracts recurring source entities, records explicit source-to-target entity decisions, and conditions chapter rewriting on those decisions. It therefore tests whether persistent identity state before rewriting improves over prose-only context, without the strategy step or the dependency-aware

graph-localization component of P_{full} .

P_{full} : final localization-sheet pipeline. P_{full} is the full system evaluated in this paper. It instantiates the five-stage architecture in Section 4: entity and relationship extraction, global cultural strategy P , graph-based localization, consistency validation, and chapter rewriting with localized names and related forms injected into the prompt.

Together, these variants separate three levels of adaptation state. B1 tests whether a global cultural plan alone is sufficient; B2 tests whether sequential prose memory can carry localization decisions across chapters; S1 tests whether explicit entity state before rewriting improves over prose-only context; and P_{full} tests the complete architecture with persistent entity state, dependency-aware localization, and validation before rewrite.

6.3 Models

All headline results (§7) use Gemini 3.1 Pro as the generator for every stage and baseline. The judges of §5.1 are fixed: Claude Sonnet 4.5 scores m_{A1}/m_{A3} and Gemini 3 Flash scores m_{A2}/m_{A0} , unchanged across all variants and generator families, and validated against native-speaker annotation (Appendix A.5). To test model-independence we replicate the four-variant ablation on GPT-5 and Claude Opus 4.5 across all three directions (Table 2; per-pair breakdown in Appendix A.9). Since the judges span two families and are fixed, any self-preference applies uniformly and cannot favor P_{full} .

6.4 Production A/B Test Details

The A/B test was conducted on three Mandarin-to-English (US) adapted shows. Platform users were randomly assigned at first listen to either the AI-drafted arm (Variant A: P_{full} plus validator-guided human review) or the professional human-drafted arm (Variant B), with variant identity hidden. Allocation was balanced: AB1 assigned 2,135 users to Variant A and 2,082 to Variant B; AB2 assigned 769 to Variant A and 796 to Variant B; and AB3 assigned 2,684 to Variant A and 2,700 to Variant B, yielding 5,588 users in Variant A and 5,578 in Variant B (11,166 total). We tracked hour-level listener retention (R_h) at $h \in \{6, 7, 8, 9, 10, 15, 20, 25, 30\}$ and report the per-show delta $\Delta_h = R_h^B - R_h^A$ descriptively.

7 Results

Overall Performance. Table 1 presents aggregate and per-language-pair performance for the proposed pipeline and baselines, evaluated across all four automated axes. P_{full} achieves the lowest pooled error rates across all three structural metrics: $m_{A1} = 0.081$, $m_{A2} = 0.083$, and $m_{A3} = 0.105$. Compared to the non-pipeline baselines, this represents an error reduction of 65% on identity (m_{A1}) and 64% on dependency (m_{A3}) versus B1. Against the B2 baseline, P_{full} maintains a 54% (m_{A1}) and 53% (m_{A3}) error reduction. Gains in cultural alignment (m_{A2}) are similarly robust, with the pipeline reducing misalignment errors by 27% compared to B1 and 26% compared to B2. Crucially, these structural improvements come at a negligible cost in plot fidelity: m_{A0} drops by only 0.007 (0.980 pooled vs. B1’s 0.987), so P_{full} achieves strong entity and dependency continuity without meaningful narrative drift.

Per-pair patterns. Results reveal that the dependency mechanism (m_{A3}) is universally robust, with P_{full} achieving the lowest error rate across all three pairs. Identity coherence (m_{A1}) is exceptionally strong for Mandarin→English and English→German. However, English→Hindi adaptation proves more challenging; the rolling-summary baseline (B2) outperforms the pipeline on identity coherence. Cultural alignment (m_{A2}) shows more variance; P_{full} decisively wins the English→German setting (0.055), significantly improves upon B1 and S1 in Mandarin→English (37 errors over 298 chapters, $m_{A2} = 0.124$) but trails B2 by a single error (36 errors over 298 chapters, $m_{A2} = 0.121$). Finally, English→Hindi slightly favors S1 on cultural alignment.

Cross-model generalization. Table 2 replicates the four-variant ablation on GPT-5 and Claude Opus 4.5 (aggregated over all three directions). P_{full} achieves the lowest error on all three axes in both families, reducing identity errors by 64% (GPT-5) and 31% (Opus 4.5) and cultural-misalignment errors by 28% in both relative to B1, while its plot fidelity stays ≥ 0.996 . The consistent $-27/-28\%$ m_{A2} reduction across all three families shows the cultural-planning stage is model-independent, and the cross-family pattern establishes that the gains are architectural (per-pair breakdowns in Appendix A.9).

	$m_{A1} \downarrow$	$m_{A2} \downarrow$	$m_{A3} \downarrow$	$m_{A0} \uparrow$
<i>Pooled (all pairs)</i>				
B1	0.233	0.114	0.291	0.987
B2	0.174	0.112	0.223	0.980
S1	0.151	0.124	0.201	0.970
P_{full}	0.081	0.083	0.105	0.980
<i>Mandarin→English</i>				
B1	0.235	0.178	0.311	0.980
B2	0.205	0.121	0.249	0.971
S1	0.157	0.215	0.196	0.955
P_{full}	0.075	0.124	0.112	0.981
<i>English→German</i>				
B1	0.120	0.060	0.131	0.992
B2	0.182	0.065	0.186	0.992
S1	0.130	0.070	0.203	0.990
P_{full}	0.014	0.055	0.018	0.993
<i>English→Hindi</i>				
B1	0.323	0.070	0.391	0.994
B2	0.111	0.146	0.207	0.982
S1	0.173	0.041	0.220	0.975
P_{full}	0.147	0.050	0.166	0.965

Table 1: Per-variant scores under the $m_{A1}/m_{A2}/m_{A3}/m_{A0}$ definitions of §5.1. Lower is better for $m_{A1}/m_{A2}/m_{A3}$; higher is better for m_{A0} . Bold marks the best value in each (pair, axis) cell.

	$m_{A1} \downarrow$	$m_{A2} \downarrow$	$m_{A3} \downarrow$	$m_{A0} \uparrow$
<i>GPT-5</i>				
B1	0.162	0.105	0.170	0.999
B2	0.157	0.115	0.160	0.992
S1	0.151	0.112	0.233	0.987
P_{full}	0.059	0.076	0.097	0.999
vs. B1	−64%	−28%	−43%	—
<i>Claude Opus 4.5</i>				
B1	0.121	0.098	0.139	0.999
B2	0.190	0.107	0.222	0.943
S1	0.119	0.102	0.131	0.988
P_{full}	0.084	0.071	0.101	0.996
vs. B1	−31%	−28%	−27%	—

Table 2: Cross-model replication: aggregate scores over all three language directions ($\downarrow$ better for $m_{A1}/m_{A2}/m_{A3}$; $\uparrow$ for m_{A0}). P_{full} attains the lowest error on every axis in both families, as with the primary backbone (Table 1).

Production A/B Test Results. Figure 3 plots hour-level A/B retention deltas for three deployed Mandarin→English(US) shows. Both arms (Variant A: drafted by P_{full} ; Variant B: human-drafted end-to-end) undergo identical downstream human review. The goal is to establish the commercial viability of P_{full} ’s long-form narratives. Across all three shows, retention delta (Δ_h) oscillates closely around the zero-line, with absolute deviations typically constrained to single digits. Notably, during the “deep retention” phase (hours 15–30), where baseline LLM adaptations can lose listeners due to identity drift (A1) and broken dependencies (A3), our pipeline-generated shows remain highly competitive. For instance, Variant A significantly outperforms the human arm in AB1 at hr15 (−7.21 pp)

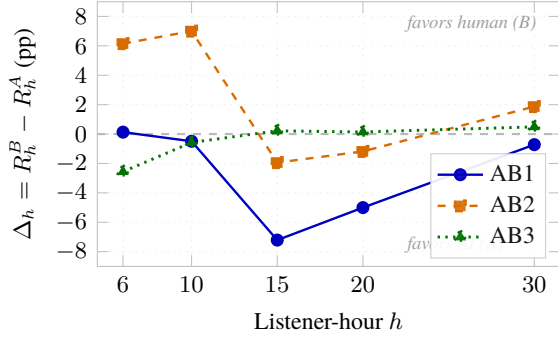


Figure 3: Production A/B retention for three Mandarin→English (US) shows. $\Delta_h = R_h^B - R_h^A$ in percentage points, where Variant A is drafted by P_{full} and Variant B by a professional human adapter; positive values favor the human arm. The dashed line marks no-effect.

and ties in AB3. This is the desired operating point. Stages 1–5 ran fully automatically; a validator-guided review pass then used Stage 4 outputs to flag specific chapters for targeted human revision, in which reviewers corrected validator-flagged errors and performed a single sanity read—24 review hours (~ 3 reviewer-days) for 10 hours of content. Across multiple generations, the strongest baseline averaged $\sim 1.8\times$ the errors of P_{full} and required $\sim 1.5\times$ more reviewer time, reflecting P_{full} ’s upstream consistency guarantees. Deployed shows remained comparable to the human-drafted arm. P_{full} achieves this parity at $\sim 2.9\times$ lower per-show cost ($\sim \$990$ vs. $\sim \$2,880$ end-to-end human; Table 5). The structural consistency generated by P_{full} effectively translates to real-world listener retention at human-parity levels, with limited human assistance.

8 Discussion

To contextualize the quantitative gains (Section 7), Table 7 (Appendix E) contrasts characteristic baseline failures with P_{full} ’s resolution; the gains on A1, A2, and A3 stem from the pipeline’s architectural state.

A1: Resolving Identity Drift via Persistent State. Standard generation paradigms invariably leak identity state over long narrative arcs. Without a globally anchored target-culture surface form, B1’s parallel generation splinters a single Chinese character into nine distinct Westernized variants; B2’s lossy rolling summaries fracture a single university into ten different institutional identities across a 100-chapter arc. P_{full} bypasses these

failures by treating identity as explicit, persistent state: the entity registry is extracted upfront and its locked forms are injected into every rewriter prompt, guaranteeing global consistency.

A2: Enforcing Cultural Alignment. Without a persistent target-culture plan tied to the entity registry, baselines retain source-culture terminology inside scenes that require adaptation. As shown in Table 7, B1 retains 真气, conventionally rendered as *True Qi* or internal energy, creating cultural residue that interrupts the adaptation’s sci-fi biometric register. P_{full} instead maps it to *somatic current*, preserving its narrative function within the target ontology of bio-energy, and renders 基础丹田/丹田 as *core somatic nexus* rather than *Dantian*, matching the adapted system’s scientific vocabulary. Localizing recurring entities at registry level thus mitigates A2 errors.

A3: Maintaining Dependency-Chain Integrity. Dependent entities must co-localize with their heads, particularly when the dependent lexically inherits from the head; baselines without explicit relational state routinely sever these possessive and institutional chains. For 罗十六的母亲 (*Luo Shiliu’s mother*), B2 drifts the dependent between disconnected proper nouns like *Margaret Hayes* and detached supernatural descriptors. By batching dependencies via Louvain community detection (Stage 3), P_{full} forces heads and dependents into the same localization window, keeping lexical and relational chains intact across hundreds of chapters and stably mapping the pair to *Ethan* and *Ethan’s mother*.

9 Conclusion

We formalize long-form cross-cultural story adaptation along three structural axes (A1, A2, A3) and instantiate a five-stage pipeline that externalizes cross-chapter state, reducing identity, dependency, and cultural-misalignment errors by 65%, 64%, and 27% over the non-pipeline baseline. A production A/B test on three deployed Mandarin→English shows demonstrates listener retention comparable to professional human adaptation at $\sim 2.9\times$ lower per-show cost. Persistent global state, not larger context, is the operative architectural unit for long-form adaptation.

Limitations

We find that the primary bottleneck of our current architecture lies in its monotonic commitment

policy: all entity decisions are frozen globally at the end of Stage 2. This strict global anchoring is highly effective for linguistically proximal pairs like English→German, yielding near-zero identity and dependency errors. It proves overly brittle, however, for culturally divergent settings like English→Hindi, where it creates systematic regressions. Furthermore, we observe similar struggles with task-level relational ambiguities, such as the shared-surname collisions observed in E.2. In both scenarios, a top-down global commitment prematurely truncates the optimization space before sufficient downstream context can be integrated. We leave the design of a dynamic entity-representation stage that incorporates feedback from the downstream rewriter as a future direction.

References

- Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. [MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2598–2637.
- Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. [Investigating cultural alignment of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422.
- Mahfuz Ahmed Anik, Abdur Rahman, Azmine Tushik Wasi, and Md Manjurul Ahsan. 2025. [Preserving cultural identity with context-aware translation through multi-agent ai systems](#). In *Proceedings of the 1st Workshop on Language Models for Under-served Communities (LM4UC 2025)*, pages 51–60.
- Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. [Fast unfolding of communities in large networks](#). *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.
- Yong Cao, Yova Kementchedjheva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. [Cultural adaptation of recipes](#). *Transactions of the Association for Computational Linguistics*, 12:80–99.
- Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. [Art or artifice? large language models and the false promise of creativity](#). In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI ’24)*, pages 1–34. Association for Computing Machinery.
- Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. [Humans or LLMs as the judge? a study on judgement bias](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327.
- Radina Dobрева, Jie Zhou, and Rachel Bawden. 2020. [Document sub-structure in neural machine translation](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 3657–3667.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *First Conference on Language Modeling (COLM)*.
- Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. 2025. [How to train long-context language models \(effectively\)](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7376–7399.
- Daniel Hershcovich, Stella Frank, Heather Lent, Miryam De Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. [Challenges and strategies in cross-cultural nlp](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013.
- Mete Ismayilzada, Claire Stevenson, and Lonneke van der Plas. 2025. Evaluating creative short story generation in humans and large language models. In *Proceedings of the 16th International Conference on Computational Creativity (ICCC)*.
- Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. [Blonde: An automatic evaluation metric for document-level machine translation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565.
- Linghao Jin, Li An, and Xuezhe Ma. 2024. [Towards chapter-to-chapter context-aware literary translation via large language models](#). *arXiv preprint arXiv:2407.08978*.
- Marzena Karpinska and Mohit Iyyer. 2023. [Large language models effectively leverage document-level context for literary translation, but critical errors persist](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451.

- Simran Khanuja, Vivek Iyer, Xiaoyu He, and Graham Neubig. 2025. [Towards automatic evaluation for image transcreation](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7034–7047.
- Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. [An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10258–10279. Best Paper Award.
- Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. [Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art](#). *Transactions of the Association for Computational Linguistics*, 13:652–689.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-Eval: NLG evaluation using GPT-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.
- Carmen Mangiron and Minako O’Hagan. 2006. Game localisation: Unleashing imagination with ‘restricted’ translation. *The Journal of Specialised Translation*, (6):10–21.
- Minako O’Hagan and Carme Mangiron. 2013. *Game Localization: Translating for the Global Digital Entertainment Industry*. John Benjamins, Amsterdam.
- Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. [LLM evaluators recognize and favor their own generations](#). In *Advances in Neural Information Processing Systems 37 (NeurIPS)*.
- Daniel Pedersen. 2014. Exploring the concept of transcreation: transcreation as “more than translation”? *Cultus*, 7:57–71.
- Katharina Reiss and Hans J. Vermeer. 1984. *Grundlegung einer allgemeinen Translationstheorie*. Niemeyer. English: Towards a General Theory of Translational Action: Skopos Theory Explained, trans. Christiane Nord, Routledge, 2013.
- Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. [Verbosity bias in preference labeling by large language models](#). *arXiv preprint arXiv:2310.10076*.
- Friedrich Schleiermacher. 1813. On the different methods of translating. Lecture delivered 24 June 1813 at the Royal Academy of Sciences, Berlin. English translation by Susan Bernofsky in Lawrence Venuti (ed.), *The Translation Studies Reader*, 3rd ed., Routledge, 2012, pp. 43–63.
- Beatriz Silva, Helena Wu, Yan Jingxuan, Vera Cabarrão, Helena Moniz, Sara Guerreiro de Sousa, João Almeida, Malene Sjørøsløv Søholm, Ana Farinha, and Paulo Dimas. 2024. [Cultural transcreation with LLMs as a new product](#). In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 57–58.
- Pushpdeep Singh, Mayur Patidar, and Lovekesh Vig. 2024. [Translating across cultures: LLMs for intralingual cultural adaptation](#). In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 400–418.
- Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. [Exploring document-level literary machine translation with parallel paragraphs from world literature](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902.
- Gideon Toury. 1995. *Descriptive Translation Studies – and Beyond*. John Benjamins.
- Lawrence Venuti. 2017. *The translator’s invisibility: A history of translation*. Routledge.
- Huy Hien Vu, Hidetaka Kamigaito, and Taro Watanabe. 2024. [Context-aware machine translation with source coreference explanation](#). *Transactions of the Association for Computational Linguistics*, 12:856–874.
- Longyue Wang, Siyou Liu, Chenyang Lyu, Wenxiang Jiao, Xing Wang, Jiahao Xu, Zhaopeng Tu, Yan Gu, Weiyu Chen, Minghao Wu, Liting Zhou, Philipp Koehn, Andy Way, and Yulin Yuan. 2024. [Findings of the wmt 2024 shared task on discourse-level literary translation](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 699–700.
- Minghao Wu, George Foster, Lizhen Qu, and Gholamreza Haffari. 2023. [Document flattening: Beyond concatenating context for document-level neural machine translation](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 448–462.
- Wenfeng Yang. 2010. [Brief study on domestication and foreignization in translation](#). *Journal of Language Teaching and Research*, 1(1):77–80.
- Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. [Benchmarking machine translation with cultural awareness](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096.
- Ran Zhang, Wei Zhao, and Steffen Eger. 2025. [How good are LLMs for literary translation, really? literary translation evaluation with humans and LLMs](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10961–10988.

Zhonghe Zhang, Xiaoyu He, Vivek Iyer, and Alexandra Birch. 2024. [Cultural adaptation of menus: A fine-grained approach](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1258–1271.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems 36 (NeurIPS)*.

A Evaluation details

This appendix explains how the automatic metrics in Section 5 are produced. A1, A2, and A3 are the three adaptation axes: identity coherence, cultural-system alignment, and dependency-chain integrity. The reported metrics m_{A1} , m_{A2} , and m_{A3} are error rates for those axes, so lower is better. The evaluators below are intentionally narrow: each one turns a concrete audit artifact into counts used by the equations in Section 5.

A.1 A1/A3 Mention Mapping

Vetted recurring-item list. For A1/A3 evaluation, each story has a human-vetted list of recurring source items. The item may be a character, institution, object, place, title, or other recurring entity. The list stores a persistent item ID, source names or aliases, and optional metadata such as chapter numbers, gender, relationships, or a short description. It is fixed once per story and reused across all evaluated variants. The list does not provide the expected adapted name; it provides the source-side inventory that the evaluator should try to map into each adapted manuscript.

The A1/A3 evaluator reads paired source and adapted chapters together with this vetted list. It first keeps only listed items that appear in the evaluated source chapters. These grouped, observed items are the validated items I_v used in Section 5. This prevents unused list entries from inflating the denominator, while also preventing a weak adaptation from lowering its denominator by omitting difficult recurring entities.

Mention-mapping artifact. For each story-variant pair, the mapper writes an audit workbook. The vetted list tells the mapper which source items to look for and which aliases should be treated as the same item. One workbook view records each observed source mention and the adapted text span that realizes it. A second view groups those rows

by vetted item ID. The grouped view is the metric-bearing view: a character with many aliases contributes one item to the denominator, not one item per spelling occurrence.

Identity-consistency checks. The A1 check asks whether each grouped source item has one coherent adapted identity across the evaluated chapters. A grouped item is consistent when the adaptation uses a stable adapted form for the same source identity, allowing ordinary titles, inflection, and spelling variation. It is inconsistent when one source identity is split into incompatible adapted identities, when unrelated source identities collapse into the same adapted identity without narrative justification, or when the adapted mention contradicts the vetted item’s role. These checks produce the consistency count $C_v^{(1)}$, which Section A.4 converts into m_{A1} .

Dependency-coherence checks. The A3 check uses the same grouped source items, but asks whether related items still fit together after adaptation. A failure is counted when an adapted form no longer coheres with the character, family, institution, place, or artifact it depends on. Examples include an estate name whose family association is broken, a school or company whose naming pattern conflicts with its affiliated characters, or a title whose adapted form no longer matches the social relation it encodes. These checks produce the dependency-consistency count $C_v^{(3)}$.

A.2 A2 Cultural-Residue Detection

The A2 evaluator measures cultural alignment from the perspective of a reader of the adapted story. It runs in two steps: a high-recall detector first proposes possible cultural problems, and a confirmation pass then keeps only supported issues for the metric. The detector reads adapted chapters together with culture labels and prior adapted-story context. A culture label names the intended movement of the adaptation: the source culture is the cultural setting the story came from, and the target culture is the setting the adaptation is meant to read as native to. Prior context is a compact record of earlier adapted chapters, used to decide whether the story has already explained a cultural detail.

The main target is *source-culture residue*: a detail whose logic still depends on the source culture or original setting after the adaptation has moved the scene into a target-culture setting. A fluent

chapter can still fail A2 if it leaves such a detail unexplained.

Chunk inputs. Detection is run over chapter chunks. For each chunk, the judge receives three kinds of information: a synthesized prior-context packet, a window of prior adapted chapters, and the current adapted chapters to analyze. The prior context helps the judge decide whether a cultural detail has already been explained by the story. It is reference material only. A reported A2 error must quote evidence from the current analyzed chapters.

Candidate findings. The detector writes candidate findings with the chapter number, a short quote from the adapted text, an error type, a Hall level, a priority, and an explanation. This first pass is high recall: it is designed to catch possible cultural residue even when some candidates may later be rejected.

Completeness gates. A run is accepted for headline evaluation only when all intended chapters have context packets and all detection chunks complete. These gates matter because m_{A2} is a per-chapter rate: missing chapters would otherwise make the adaptation look better than it is.

A.3 A2 Decision Rule and Confirmation

A candidate becomes an A2 issue only if it passes a concrete localization test. The text must establish or strongly imply a target setting, institution, social context, or character background. The quoted detail must be treated as normal in that setting. The detail must depend on source-culture logic, source-language convention, or the original setting. Finally, the surrounding story must not give a clear explanation for why that detail belongs.

This rule prevents A2 from becoming a generic “foreignness” detector. Non-target-culture details are allowed when they are explained by the story: immigrant or heritage background, travel, religious practice, imported goods, a restaurant setting, genre worldbuilding, historical setting, or an explicitly multicultural scene. Such details are counted only when the passage asks the reader to treat them as ordinary target-culture defaults.

Hall levels. Each confirmed issue is assigned one Hall-triad level. **Technical/Visible** issues involve explicit cultural objects or labels, such as food, currency, clothing, holidays, place names, honorifics, school systems, or administrative institutions. **Formal/Semi-visible** issues involve

learned routines and social procedures, such as etiquette, courtship scripts, family-address norms, workplace hierarchy, school discipline, or ritual behavior. **Informal/Invisible** issues involve deeper assumptions about authority, obligation, kinship, status, shame, privacy, gender roles, or social risk. The assigned level describes the depth of the mismatch; it is not a separate metric.

Confirmation pass. Because the detector is high recall, a separate confirmation pass reviews every candidate. A confirmed issue must state the target-culture expectation, the source-culture logic retained by the adapted text, and why prior story context does not justify that retention. Candidates that are weak, speculative, or plausibly explained by context are excluded from the metric. The headline m_{A2} numerator uses only confirmed issues.

Diagnostic fields. Error type and priority are kept for analysis but do not change the score. Priority records reader-facing severity, and error type records the kind of cultural problem, such as a food or ritual mismatch, an institutional mismatch, an honorific or title mismatch, a place-name mismatch, or a deeper social-norm mismatch. m_{A2} counts all confirmed issues equally.

A.4 Metric Mapping

For A1/A3, let I_v denote the number of validated grouped source items for variant v . Let $C_v^{(1)}$ be the number with stable identity mappings, and let $C_v^{(3)}$ be the number whose dependent or related adapted forms remain coherent. The displayed error rates are therefore

$$m_{A1}(v) = 1 - \frac{C_v^{(1)}}{I_v}, \quad m_{A3}(v) = 1 - \frac{C_v^{(3)}}{I_v}.$$

This is the complement of the consistency-rate fields emitted by the validator. Using the complement keeps all three automatic metrics in the same direction: lower is better.

For MA2, let $F_v^{(2)}$ be the confirmed Hall-triad issue count for variant v , and let H_v be the number of evaluated adapted chapters for the same run. The displayed score is

$$m_{A2}(v) = \frac{F_v^{(2)}}{H_v}.$$

The numerator is read from the confirmed-count artifact, not from the raw high-recall detector output.

The denominator is the target chapter count after the completeness gates in Appendix A.2.

For pair-level and pooled tables, numerators and denominators are summed over the included story-variant cells before converting to rates. We therefore sum counts before dividing, rather than averaging per-story rates.

A.5 Evaluation framework validation against native annotators

We validated the evaluation framework against independent human annotation using an internal evaluation tool that pairs the LLM judge with three native-speaker annotators. On a 50-chapter held-out subset covering all three structural axes, the annotators independently labeled identity, dependency, and Hall-triad cultural errors, then deliberated to reach a consensus list of 37 confirmed errors. The judge, run on the same chapters, flagged 39 issues. Of these, 36 overlapped with the annotator consensus; 3 were false positives (flagged by the judge but not confirmed by annotators), and 1 annotator-confirmed error was absent from the judge’s list. This yields a precision of 92% and a recall of 97% against human judgment across all error types. The 50 chapters were drawn from two Mandarin→English shows held out from all experiments in this paper (a drama-romance series and a fantasy series), so the judge was validated on material it was never tuned or evaluated on. Annotators labeled independently before deliberating to the consensus set; pre-deliberation inter-annotator agreement was not logged, and we report only consensus-level precision and recall. The judge prompts were iteratively refined with annotators in the loop prior to this final validation pass.

A.6 Dataset characteristics

The five source stories on which the headline metrics are computed are commercially-produced serialized chapters of roughly 2,000–4,000 source characters each, with ~ 100 chapters per story. The two English-source stories are adapted into both German and Hindi from a shared source manuscript.

A.7 Competing method configurations

Tables 3 and 4 summarize the four headline variants under two complementary views: which pipeline mechanisms each variant runs, and what each generation call sees at the input.

Component	B1	B2	S1	P_{full}
Cultural plan P	✓	—	—	✓
Entity registry	—	—	✓	✓
Cross-chapter coreference	—	—	✓	✓
Relationship graph	—	—	—	✓
Louvain batching	—	—	—	✓
Validation (Stage 4)	—	—	—	✓
Canonical-name injection	—	—	✓	✓
Chapter rewrite	✓	✓	✓	✓

Table 3: Pipeline-component matrix for the four headline variants. ✓: component active; —: absent. B2 substitutes a rolling natural-language summary for the structured registry/plan (see Table 4). This matrix is the source of truth; the prose definitions in §6.2 and below are aligned with it.

B1 – plan-then-adapt. B1 comprises one whole-novel planning call followed by parallel per-chapter rewrite calls. The planning call receives the entire source novel and produces a target-culture plan in the same six-section schema as Stage 2 of P_{full} : cultural mappings, character localization rules, place localization rules, narrative-voice strategy, genre framing, and naming conventions. Each rewrite call receives one source chapter together with the plan and produces that chapter’s adaptation.

B2 – rolling-summary monolith. B2 processes the source novel chapter-by-chapter. Each chapter’s prompt comprises (L_T, C_T) , the source chapter, and a running natural-language summary of the entities introduced and the localization decisions made in prior chapters. A subsequent generator call after each chapter updates the summary under a 4,000-token budget. The summary is empty prior to chapter 1.

S1 – in-context rolling NER. S1 replaces Stage 1’s cross-chapter coreference and relationship-graph construction with per-chapter NER and a flat running entity table whose rows record source name, localized name, first chapter of appearance, and cumulative mention count. Cross-chapter coreference is disabled and no relationship graph is constructed; localization decisions for newly encountered entities are conditioned only on the entity table and the current chapter. Stage 2 (cultural plan P) is disabled and Stage 3’s dependency-aware graph-localization is replaced by chapter-co-occurrence batching (no relationship graph). Stage 4 (validation) and Stage 5 (chapter rewrite with canonical-name injection) operate at their P_{full} defaults.

Variant	Upfront pass over the novel	Per-chapter rewrite input	Cross-chapter state carried forward
B1	Full novel $\rightarrow$ cultural plan (one call)	Source Chapter + plan $\rightarrow$ full adaptation in multiple parallel calls	Plan only; no per-entity state
B2	—	Source chapter + rolling summary ($\leq 4k$ tokens)	Running natural-language summary
S1	Full novel processed in chunks (per-chapter NER and cross chapter coreference) $\rightarrow$ registry	Source chapter + registry slice	Flat entity table (name pairs, mention counts); no plan, no graph
P_{full}	Full novel processed in chunks $\rightarrow$ registry, dependency graph, plan P (Stages 1–4)	Source chapter + registry slice	Structured state: registry, dependency graph, plan P

Table 4: Per-call context budget and chapter access by variant. The contrast B1 vs. P_{full} shows that the headline pipeline’s rewrite call receives strictly *less* raw context than B1’s; the contrast S1 vs. P_{full} holds chapter access fixed and varies only the structured-state content carried into the rewriter, isolating the contribution of plan P plus registry beyond chapter-by-chapter access.

A.8 Production A/B trial details

The production A/B trial runs on a commercial serialized-fiction platform. For each of three Mandarin-source serialized shows, two adapted English-language scripts are produced—Variant A, the P_{full} output with a minimal validator-guided human review pass, and Variant B, an end-to-end professional human adaptation under the platform’s standard editorial process. At first listen for a given show, the platform randomly allocates each new listener to Variant A or Variant B; allocation is at the user level within show, and listeners are not informed of their assigned variant. Variant identity is concealed in the listener-facing interface. The primary outcome is hour-level retention R_h for $h \in \{6, 7, 8, 9, 10, 15, 20, 25, 30\}$, and the secondary outcome is listener-day active users (LDAU). The three shows contribute 11,166 allocated users in total, split 5,588/5,578 across Variant A/Variant B, with 963/998 LDAU. Operationally, the AI-drafted arm required 3 days of validator-guided human review for 10 hours of content; the professional writer workflow took 6 writer days and included story-level changes from the base draft.

Cost comparison. We estimate the per-show cost of each arm for a typical ~ 100 -chapter Mandarin $\rightarrow$ English production (median chapter $\approx 3,200$ source tokens). The P_{full} + verification arm aggregates LLM tokens for the five-stage pipeline (Gemini 3.1 Pro generator; $\sim \$18$, with Stage 5’s 100 per-chapter rewrites contributing roughly two-thirds) and for the evaluation framework (Claude Sonnet 4.5 and Gemini 3 Flash

Lane	LLM	Hours	Rate	Total
P_{full} + verification	\$29	24	\$40/hr	\$989
Full human adaptation	—	48	\$60/hr	\$2,880

Table 5: Ballpark per-show cost for one ~ 100 -chapter Mandarin $\rightarrow$ English production. LLM tokens priced at Claude Sonnet 4.5 rates (\$3/\$15 per 1M input/output), the costliest model used, giving an upper bound. Human time normalized to an 8-hour workday.

judges; $\sim \$10$, dominated by Pass 2’s sliding-window detector and the per-chapter m_{A0} evaluator), plus 3 days of validator-guided human review at \$40/hr. The full-human arm is 6 writer-days at \$60/hr. All LLM tokens are priced at Claude Sonnet 4.5 published rates (\$3 / \$15 per 1M input / output tokens), the costliest model in the stack, so the LLM figures are upper bounds (Gemini 3.1 Pro: \$2 / \$12 per 1M); a workday is normalized to 8 hours. Table 5 summarizes the comparison.

The pipeline lane is roughly $2.9\times$ cheaper, saving $\sim \$1,890$ per show ($\sim 66\%$ reduction), and LLM tokens account for only $\sim 3\%$ of the P_{full} total.

A.9 Cross-model replication: per-language-pair breakdown

Table 2 in the main text reports aggregate scores for the four-variant ablation replicated on GPT-5 and Claude Opus 4.5; each aggregate is the mean of the three language directions. Table 6 gives the per-direction values for the identity and dependency axes (m_{A1} , m_{A3}). The Claude Opus 4.5 English $\rightarrow$ Hindi P_{full} cell was re-scored with the

same fixed judges, so its three-direction mean differs from Table 2 by rounding.

	$m_{A1} \downarrow$	$m_{A3} \downarrow$
<i>GPT-5 — Mandarin→English</i>		
B1	0.055	0.055
B2	0.041	0.041
S1	0.050	0.167
P_{full}	0.054	0.167
<i>GPT-5 — English→German</i>		
B1	0.155	0.155
B2	0.171	0.171
S1	0.191	0.217
P_{full}	0.052	0.052
<i>GPT-5 — English→Hindi</i>		
B1	0.276	0.299
B2	0.260	0.268
S1	0.213	0.315
P_{full}	0.071	0.071
<i>Claude Opus 4.5 — Mandarin→English</i>		
B1	0.131	0.159
B2	0.027	0.114
S1	0.064	0.068
P_{full}	0.075	0.085
<i>Claude Opus 4.5 — English→German</i>		
B1	0.060	0.069
B2	0.221	0.230
S1	0.190	0.207
P_{full}	0.147	0.147
<i>Claude Opus 4.5 — English→Hindi</i>		
B1	0.173	0.189
B2	0.323	0.323
S1	0.102	0.118
P_{full}	0.035	0.081

Table 6: Cross-model replication by language direction ($\downarrow$ better). Bold marks the best value among the four variants for that (direction, axis).

A.10 Sample Stage-2 plan output (B1 baseline)

B1’s single whole-novel planning call emits a structured plan in the same six-section schema as Stage 2 of P_{full} : cultural mappings, character localization rules, place localization rules, narrative-voice strategy, genre framing, and naming conventions. The excerpt below is verbatim output from a B1 run on one Mandarin→English show (a contemporary romance); the platform name is redacted and the naming-conventions list is truncated (9 of 17 entries shown).

```
### 1. Cultural mappings
* The Guzheng & Musical Factions: The traditional Chinese Guzheng is mapped to the Classical Violin. The "Yun Guzheng Faction" becomes an elite, historic American string methodology known as the "Quinn Bowing Technique." Regional factions (Shaanxi, Zhejiang) become rival American conservatories (e.g., the New York Juilliard elite vs. the Chicago String Society).
* Antique Instruments: Famous Guzheng brands (Dunhuang, Moon Crescent) are mapped to elite violin makers. A standard high-end violin is a "Modern Cremona," while the priceless antiques are "Stradivarius" or "Guarneri" models.
* Food & Culinary Customs: "Biangbiang noodles" and "Hulu chicken" are mapped to authentic New England Clam Chowder and Artisanal Roast Pheasant. The night
```

```
snacks (宵夜) Chang Su cooks become late-night gourmet charcuterie boards or artisanal midnight pastas.
* Web Novelist Career: Bai He's career as a struggling web novelist is mapped to a [platform] indie romance author who is constantly stressing over her [platform] page reads and Amazon algorithm rankings.
* The Will & Inheritance: The grandfather's will requiring marriage and a child to inherit the estate is mapped to an Old Money New England Trust Fund stipulation, a common trope in Western billionaire romance.
```

2. Character localization rules

```
* Julian Quinn (覃小津 - Qin Xiaojin): The aloof, prodigy violinist. He speaks in highly formal, precise, and slightly archaic English. He never uses contractions or modern slang, reflecting his isolated, elite upbringing.
* Harper White (白荷 - Bai He): The resilient, pragmatic indie author. She uses contemporary millennial/Gen Z slang, sarcasm, and self-deprecating humor. Her dialogue should feel chaotic but grounded.
[...]
```

3. Place / setting localization rules

```
* City Setting: "Yun City" is mapped to Boston, Massachusetts, a city with a deep-rooted classical music scene and old-money families.
* The Estate: "Lanhua Wu" (Blue Jacaranda Estate) is mapped to Blue Wisteria Manor, a sprawling, historic estate located in the wealthy, wooded suburbs of Massachusetts (e.g., Wellesley or Concord).
[...]
```

6. Naming conventions

```
* Qin Xiaojin -> Julian Quinn
* Bai He -> Harper White
* Xiang Qing -> Chloe Shaw
* Chang Su -> Sam Carter
* Qin Shanhai -> Arthur Quinn
* Qin Hu -> Helen Quinn
* Sang Professor -> Professor Beatrice Quinn
* Gong Yi -> Gabriel Vance
* Liu Lang / Qin Lang -> Leo Quinn
[...]
```

A.11 Worked conflict-detection example (Stage 4)

We illustrate the registry-level containment described in §4 with a logged Stage 4 instance from a Mandarin→English run. Stage 3 had localized two unrelated characters, 牛婉兮 (*Annie Beaumont*) and 李庆丰 (*James Beaumont*), to the same surname. Before any chapter was rewritten, the duplicate-name audit raised:

```
Character 'char_107' (original: '牛婉兮') shares localized last name 'Beaumont' with unrelated character 'char_72' (original: '李庆丰'). Their original surnames are different, confirming they are not family. Any reference to 'Beaumont' is ambiguous.
```

Because the two source surnames differ, the conflict is resolvable at the registry level: the repair regenerates only the affected mapping, 牛婉兮 → *Annie Caldwell*, while 李庆丰 → *James Beaumont* and every other registry entry are left untouched and Stage 1 extraction is not rerun. Every subsequent Stage 5 rewrite call queries the corrected registry, so the false shared surname never reaches the adapted text.

B Prompts

We collect the prompt templates used by the evaluation framework. Curly-brace placeholders {variable} indicate runtime substitutions.

B.1 Evaluation prompts

This subsection gives the canonical evaluation-prompt templates used for the reported MA1, MA2, and MA3 scores in Section 5. The prompts are shown with runtime placeholders in braces. Language-pair-specific examples and target-culture labels are substituted at run time, but the task definitions, decision rules, and output fields are fixed across variants.

MA1/MA3 mention mapper. The mapper is the bridge from the human-vetted recurring-item list to the metric-bearing audit artifact. It reads the source chapter, the adapted chapter, and candidate mentions selected from the vetted list. It then records how each observed source-side item is realized in the adapted text.

```
[SYSTEM]
You are a precise bilingual localization auditor. Return
only valid JSON.
Do not invent names or mappings that are not supported
by the adapted text.

[USER]
You are given:
1. The original source chapter in {source_language}.
2. The adapted chapter in {target_language} for {
target_culture}.
3. Candidate recurring-item mentions selected from the
vetted item list.

Task:
Identify how each candidate source mention appears in the
adapted chapter.

Rules:
- Use only evidence from the adapted chapter for the
adapted span.
- Preserve the provided character ID or entity ID exactly.
- Keep source mentions exactly as provided in the
candidate list.
- If the candidate is omitted from the adapted chapter,
mark it "unmapped".
- If the adapted chapter refers to the concept without a
stable name, mark it
"conceptual".
- If several adapted mentions are plausible and the
evidence is not decisive,
mark the row "ambiguous".
- If a concrete adapted mention is present, mark the row "
mapped".
- Return JSON only, with no markdown.

Return:
{
  "mappings": [
    {
      "type": "character|entity",
      "char_id": "source character id, or empty string",
      "entity_id": "source entity id, or empty string",
      "source_canonical_name": "source item name",
      "source_mention": "exact source mention",
      "observed_adapted_mention": "exact adapted
mention or empty string",
      "status": "mapped|unmapped|ambiguous|conceptual",
      "confidence": "high|medium|low",
      "source_evidence": "short source evidence",
```

```
      "adapted_evidence": "short adapted evidence",
      "notes": "brief reason"
    }
  ]
}

Candidate mentions:
{candidate_mentions}

Source chapter:
{source_chapter}

Adapted chapter:
{adapted_chapter}
```

Rows from this mapper are grouped by vetted item ID before scoring. The grouped view defines the validated item set I_v and provides the evidence used by the MA1 and MA3 validators.

MA1 identity-consistency validator. This validator reads grouped character rows from the mention mapper. It asks whether the same recurring source character keeps one coherent adapted identity across the evaluated chapters.

```
[SYSTEM]
You are an expert localization validator for long-form
story adaptation.
Return only valid JSON.

[USER]
Validate the adapted character mappings for {
target_language} readers in
{target_culture}.

Input:
- Characters to validate: {characters}
- Related entity context: {entity_context}

Scope:
Validate the adapted character-name fields and mention-
map values. Treat
descriptions and source-canonical names as context only.
They may contain
source-language names and should not by themselves
create an issue.

Decision rule:
For each character, decide whether the adapted mentions
consistently realize
the same source character. Allow normal title changes,
inflection, casing, and
minor spelling variation. Flag an issue when the same
source character is split
across incompatible adapted identities, when the mention
map preserves an
unadapted source form where a stable adapted form is
required, or when a
character mention conflicts with the adapted entity or
relationship context.

Metric-bearing issue labels:
- "mention_map_inconsistency": the adapted mentions for
the same character do
not form a stable identity.
- "cross_character_inconsistency": related character names
or aliases conflict
with each other.
- "entity_inconsistency": a character mention conflicts
with the adapted entity
context.

Return only characters with issues:
{
  "character_issues": [
    {
      "char_id": "character id",
      "localization_issues": [
        {
          "issue_type": "one allowed label",
```

```

      "description": "short explanation",
      "affected_mentions": ["mention"],
      "entity_reference": "entity id if relevant",
      "severity": "high|medium|low"
    }
  ]
}

```

The MA1 count $C_v^{(1)}$ treats a grouped item as consistent when this pass finds no metric-bearing mention-map inconsistency for that item.

MA3 dependency-coherence validator. This validator reads grouped non-character entity rows and character context. It asks whether adapted dependent names still cohere with the adapted characters, institutions, places, objects, and concepts they rely on.

```

[SYSTEM]
You are an expert localization validator for recurring story
entities.
Return only valid JSON.

[USER]
Validate the adapted entity mappings for {
target_language} readers in
{target_culture}.

Input:
- Entities to validate: {entities}
- Character context: {character_context}

Scope:
Validate adapted entity names and mention-map values.
Treat descriptions,
source names, and source-side aliases as context only.

Decision rule:
For each entity, decide whether its adapted mentions
remain coherent with
the adapted story context. Flag an issue when an entity
mention refers to a
character by an incompatible adapted name, when related
entities no longer
share the name component or institutional pattern that
their relationship
requires, or when a dependent place, title, organization,
object, or concept
drifts away from the adapted head item it depends on.

Metric-bearing issue labels:
- "mention_map_inconsistency": the adapted mentions for
the same entity are
internally unstable.
- "character_context_mismatch": an entity mention
conflicts with the adapted
character name it depends on.
- "cross_entity_inconsistency": related entities are
adapted with incompatible
dependent names or naming patterns.

Return only entities with issues:
{
  "entity_issues": [
    {
      "entity_id": "entity id",
      "localization_issues": [
        {
          "issue_type": "one allowed label",
          "description": "short explanation",
          "affected_mentions": ["mention"],
          "severity": "high|medium|low"
        }
      ]
    }
  ]
}

```

The MA3 count $C_v^{(3)}$ treats a grouped item as dependency-consistent when this pass finds no metric-bearing related-item issue for that item.

MA2 cultural-residue detector. The MA2 detector evaluates the adapted chapters as a target-culture reading experience. It receives prior adapted-story context packets, a short window of previous adapted chapters, and the current chapter chunk to analyze. The context packets help the judge decide whether a cultural detail has already been explained by the story; the quoted evidence for an issue must come from the current chunk.

```

[SYSTEM]
You are an expert cultural localization QA editor for
fiction localized from
a source culture into an established target locale. Return
only valid JSON.

[USER]
Analyze the adapted chapters for source-culture residue.

You receive three input sections:
1. SYNTHESIZED OVERALL CONTEXT
  Prior adapted-story context for the target chapters. Use
it to understand
  what the story has already established. Do not quote it
as evidence.
2. PREVIOUS CHAPTERS
  Adapted prose before the current chunk. Use it only as
context.
3. CHAPTERS TO ANALYZE
  The current adapted chapter chunk. Report only issues
whose exact quote is
  a continuous substring of this section.

Source culture: {source_culture}
Target culture: {target_culture}
Target chapters: {target_chapters}

Definition:
Source-culture residue is a detail whose narrative logic still
depends on the
source culture, source-language convention, or original
setting after the
adaptation has moved the relevant character, location,
institution,
relationship, or social context into the target locale.

Localization test:
Flag a finding when the quoted detail is treated as a
default target-locale
norm, institution, ritual, economic assumption,
relationship rule, ordinary
behavior, naming logic, or expected procedure, and when
the surrounding story
does not give a concrete explanation that makes the detail
natural in the
adapted setting.

Allowed explanations include character background, family
heritage, travel,
religion, imported goods, restaurant setting, fandom,
personal habit, genre
worldbuilding, historical setting, or explicit multicultural
context.

Hall level:
- "Technical / Visible": explicit culture-bound terms,
names, objects,
  places, institutions, currency, festivals, or titles.
- "Formal / Semi-visible": learned routines, etiquette,
rituals,
  institutional procedures, family/workplace/school
conduct, or role behavior.
- "Informal / Invisible": deeper assumptions about
authority, obligation,
  shame, privacy, gender, kinship, status, politeness, taboo

```

, or worldview.

Priority:

- High: breaks the target setting or causes major institutional, plot, or character-logic damage.
- Medium: likely confuses target-locale readers or weakens plausibility.
- Low: minor or uncertain but text-supported cultural residue.

Return:

```
{
  "hall_iceberg_summary": {
    "technical_visible": {"score": 1-5, "issue_count": 0, "assessment": "..."},
    "formal_semi_visible": {"score": 1-5, "issue_count": 0, "assessment": "..."},
    "informal_invisible": {"score": 1-5, "issue_count": 0, "assessment": "..."}
  },
  "inconsistencies": [
    {
      "chapter_number": 1,
      "priority": "High|Medium|Low",
      "error_type": ["Source-Culture Residue"],
      "hall_level": "Technical / Visible|Formal / Semi-visible|Informal / Invisible",
      "quote_with_error": "exact continuous quote from CHAPTERS TO ANALYZE",
      "explanation": "target expectation, retained source logic, and why context does not justify it"
    }
  ]
}
```

SYNTHESIZED OVERALL CONTEXT:

{prior_context_packet}

PREVIOUS CHAPTERS:

{previous_chapters}

CHAPTERS TO ANALYZE:

{chapters_to_analyze}

MA2 acceptance rule. The detector is run in a high-recall mode, but a finding contributes to $F_v^{(2)}$ only when it satisfies the localization test above, has a chapter number in the evaluated chunk, and quotes an exact span from the analyzed adapted chapters. Malformed findings, duplicate reports of the same quoted issue, and findings justified only by prior-context material are excluded from the count used in m_{A2} .

m_{A0} : chapter beat-analysis prompt. The m_{A0} evaluator (§5.1) issues a single LLM call per (source, adapted) chapter pair and returns a structured beat-level analysis with five sub-scores. The full prompt template is shown below.

[SYSTEM]

You are a rigorous evaluator for cultural adaptation quality. Your task is to evaluate MA0: whether the adapted chapter preserves the source chapter's story beats. Focus on narrative/content fidelity, not whether localization choices are culturally perfect. The source chapter may be in Chinese/Mandarin and the adapted chapter may be in American English.

[USER]

Evaluate one source/adapted chapter pair for MA0 story-beat preservation.

MA0 definition:

- Measures whether the adapted text preserves the source

story's plot events, character actions, relationships, causal links, stakes, reveals, and important entities.

- Do not penalize culturally appropriate substitutions by themselves if they preserve the same narrative function.
- Penalize missing source beats, hallucinated new beats, wrong causality, wrong relationships, wrong character/entity mapping, and major sequence changes that alter the story.

Scoring guidance:

- beat_recall: 1.0 means all important source beats are preserved; 0.0 means almost none are preserved.
- beat_precision: 1.0 means the adapted chapter adds no meaningful hallucinated/contradictory beats; 0.0 means many adapted beats are unsupported.
- sequence_fidelity: 1.0 means the order/causal flow is preserved; 0.0 means the adapted flow changes the story.
- character_entity_fidelity: 1.0 means characters/entities map consistently in narrative function; 0.0 means major people/entities are confused.
- overall_ma0_score: holistic 0.0-1.0 score weighted toward beat_recall and story-changing errors.

Return only valid JSON with this exact top-level shape:

```
{
  "chapter_index": <integer>,
  "source_heading": "<string>",
  "adapted_heading": "<string>",
  "source_beats": [{ "beat_id": "S1", "beat": "<atomic source story beat>", "importance": "critical|major|minor" }],
  "adapted_beats": [{ "beat_id": "A1", "beat": "<atomic adapted story beat>", "source_alignment": "S1|S2|none|unclear" }],
  "beat_alignment": [{ "source_beat_id": "S1", "status": "preserved|partial|missing|contradicted", "adapted_beat_ids": ["A1"], "evidence": "<short evidence>", "notes": "<status reasoning>" }],
  "missing_or_weakened_beats": [{ "source_beat_id": "S1", "severity": "high|medium|low", "explanation": "<what is missing/weakened>" }],
  "added_or_changed_beats": [{ "adapted_beat_id": "A1", "severity": "high|medium|low", "explanation": "<unsupported or story-changing addition>" }],
  "character_entity_issues": [{ "severity": "high|medium|low", "explanation": "<character/entity mapping or relationship issue>" }],
  "scores": { "beat_recall": <float>, "beat_precision": <float>, "sequence_fidelity": <float>, "character_entity_fidelity": <float>, "overall_ma0_score": <float> },
  "verdict": "pass|borderline|fail",
  "reasoning": "<concise evaluator reasoning>"
}
```

Source language: {source_language}

Target/adapted language: {target_language}

Source culture: {source_culture}

Target culture: {target_culture}

Chapter index: {chapter_index}

Source heading: {source_heading}

Adapted heading: {adapted_heading}

SOURCE CHAPTER:

{source_chapter_text}

ADAPTED CHAPTER:

{adapted_chapter_text}

C Extended related work

This appendix expands the positioning in §2.

Translation-studies foundations. The conceptual scaffold for separating *translation* from *adaptation* has a long history in translation studies. Schleiermacher's foreignization-domestication dichotomy (Schleiermacher, 1813), Toury's descriptive translation studies (Toury, 1995), Venuti's ac-

count of translator invisibility (Venuti, 2017), and the Skopos theory of translational action of Reiss and Vermeer (1984) together establish that translation strategy is a choice conditioned on purpose and target culture: when the skopos prioritizes target-culture experience, the regime shifts from source-faithful equivalence to target-culture rewriting. Industry practice has formalized this shift as *transcreation* in game localization (O’Hagan and Mangiron, 2013; Mangiron and O’Hagan, 2006) and in marketing and advertising (Pedersen, 2014), and more recent NLP work has begun to import transcreation as an operational construct (Silva et al., 2024). Our work is, to our knowledge, the first NLP formalization of long-form cross-cultural adaptation with named structural axes and an architectural decomposition aligned to those axes; we draw on these foundations conceptually rather than as quantitative comparison points.

Document-level and literary machine translation. A line of work has tightened the sentence-level gap in literary settings. Karpinska and Iyyer (2023) show that document-level context lets LLMs outperform sentence-by-sentence LLM translation on literary paragraphs across 18 language pairs, while leaving a tail of critical errors that still requires human intervention; Jin et al. (2024) introduce a chapter-level Chinese–English benchmark and show that chapter-to-chapter translation remains challenging for current LLMs. Thai et al. (2022) contribute the Par3 paragraph-aligned multilingual literary corpus, and Wang et al. (2024) run the WMT 2024 Discourse-Level Literary Translation shared task. Discourse-level evaluation has matured around BlonDe (Jiang et al., 2022), which scores entity, tense, pronoun, and discourse-marker consistency on document-sized inputs; Zhang et al. (2025) extend literary-evaluation methodology with human-vs-LLM head-to-heads. Separately, Vu et al. (2024) inject source-side coreference into a context-aware MT model and report BLEU gains of up to ~ 2 points on IWSLT En-De, OpenSubtitles En-Ru, and multilingual TED, demonstrating the value of coreference signal for local reference stability even though the evaluation is on subtitles and talks rather than on book-length fiction. These systems narrow the sentence-level gap and improve local discourse cohesion, but they operate without an explicit cross-chapter entity registry, without a global target-culture plan,

and without dependency-aware batching: identity drifts across hundreds of chapters (A1) and source-culture artifacts persist inside target-culture scenes (A2). Our Stage 1 cross-chapter coreference plus entity registry, Stage 2 cultural plan P , and Stage 3 Louvain-based batching (Blondel et al., 2008) are the persistent global state these systems lack. Earlier work on *document substructure* for MT (Dobrev et al., 2020) encoded Wikipedia section topics as side constraints or cache state, anticipating the idea that documents carry exploitable internal structure; our cross-chapter entity registry and plan P generalize this from intra-document sections to inter-chapter narrative state. Subsequent work scales the *quantity* of context—e.g., DocFlat’s flat-batch attention and neural context gate push beyond pseudo-document boundaries on En-De (Wu et al., 2023)—but treats context as a flat token stream rather than as a structured cross-chapter state, which is the gap our Stage 1 registry and Stage 2 plan P target.

Atomic-scale cross-cultural adaptation in NLP.

A parallel line of NLP work establishes that adaptation outperforms literal translation at *atomic* scale: Cao et al. (2024) show this for recipes; Yao et al. (2024) benchmark MT cultural awareness on culture-specific items; Singh et al. (2024) treat cultural adaptation as an intralingual rewriting problem; and Khanuja et al. (2024) extend adaptation to images for cross-cultural relevance, with Khanuja et al. (2025) proposing automatic evaluation for image transcreation. These artifacts share a defining property: they have no recurring identity across chapters, no dependent sibling entities, and no narrative continuity, so the techniques do not lift directly to long-form output. We extend the adaptation paradigm from atomic artifacts to multi-chapter narrative by introducing a structured dependency graph (Stage 1) and Louvain-based dependency-aware batching (Stage 3) that operate across recurring entities, plus a global cultural plan P (Stage 2) that constrains the space of acceptable mappings so atomic-scale “adaptation > literal” transfers coherently to long-form output.

Contrast with image transcreation. The closest prior work in spirit is the image-transcreation line opened by Khanuja et al. (2024) and continued by Khanuja et al. (2025), which established cultural *adaptation* (rather than literal translation) as the right regime for cross-cultural image rendering and provided the first automatic evaluation frame-

work for that task. Our work shares the “adaptation > literal” thesis but operates in a regime where three new structural difficulties appear simultaneously and the [Khanuja et al. \(2024\)](#) apparatus does not transfer: (i) *recurring identity across the long arc* (an image has no chapter-50 follow-up; a character does), which forces the A1 machinery and persistent global state; (ii) *dependency-chain structure* among co-referring entities (an image has no head/dependent semantics; a family does), which forces the A3 machinery and the Louvain dependency-aware batching; and (iii) *cultural-system alignment within scenes* (an image is one scene; a novel is a hundred), which forces the per-scene A2 evaluation framework rather than per-artifact judgment. We import the methodological commitment of [Khanuja et al. \(2024\)](#) to axis-decomposed evaluation and to native target-culture validation, and extend it to the long-form regime along all three new axes; in our setting the native target-culture validation comes from listener engagement on a deployed production A/B rather than from crowdsourced ratings by native annotators.

Long-form generation and discourse-level evaluation. Long-form generation evaluation has converged on discourse and engagement measurement. *BlonDe* ([Jiang et al., 2022](#)) is a discourse-overlap metric for document-level MT; broad multilingual benchmarks such as MEGAVERSE ([Ahuja et al., 2024](#)) stress-test LLMs across 22 datasets and 83 languages but treat tasks at sentence or short-passage granularity rather than at chapter scale; and creative-narrative evaluation work ([Ismayilzada et al., 2025](#); [Chakrabarty et al., 2024](#)) probes generation quality on creative short stories. These instruments measure intra-language coherence, engagement, and creativity, but none isolate cross-cultural alignment as an evaluation axis or pair operational metrics with an external listener-engagement instrument at production scale. Our evaluation framework (§5) addresses this gap: m_{A1} is an identity error rate over validated recurring entities across the full arc; m_{A2} counts confirmed cultural-residue errors per chapter under the Hall-triad taxonomy (Appendix A.3); m_{A3} is a dependency-coherence error rate; and the architecture-level result is externally validated by a production A/B against professional human adaptation on hour-level listener retention.

LLM-as-judge methodology and calibration.

We follow the LLM-as-judge methodology established by [Zheng et al. \(2023\)](#) and extend the G-Eval-style reference-free scoring of [Liu et al. \(2023\)](#) with axis-decomposed rubrics. Known judge biases motivate our fixed cross-family judge assignment (§6.3) and validation against human annotation (Appendix A.5): length effects ([Dubois et al., 2024](#); [Saito et al., 2023](#)), self-preference of evaluators for their own generations ([Panickssery et al., 2024](#)), and judgement biases more broadly ([Chen et al., 2024](#)). Off-the-shelf judge protocols are uncalibrated for per-axis cross-cultural claims at long-form scale; the pipeline used in this paper, validated against native-speaker annotation (Appendix A.5), is the minimum bar we judge sufficient.

D Task scope, non-redundancy, and regime

This appendix expands the scope and methodology paragraphs of §3.

Non-redundancy: structural realizability.

The three axes are non-redundant in the strong sense that each is independently realizable as the sole failure mode of an otherwise coherent output. *A1 alone*: a chapter rendered with consistent cultural artifacts (no enclaves) and coherent dependent references (head → estate → honorific all in C_T) can still fail A1 at chapter 87 when the localized name of a character introduced in chapter 3 drifts to a different target-side spelling. *A2 alone*: every recurring entity is renamed consistently and every dependent reference is coherent under the new head, yet the kitchen scene still contains *baozi* where it should contain the target-culture equivalent. *A3 alone*: identity is stable across the arc and cultural artifacts inside scenes are target-native, yet the *Qin Family Villa* retains the source-culture form because lexical find-and-replace on the head “Qin” → “Sterling” cannot reach the semantically-dependent estate name. Each of these single-failure cells is observed in baseline outputs (Appendix E).

Why these three axes (and not others). Prior literary-translation evaluation already covers discourse consistency (*BlonDe* ([Jiang et al., 2022](#))), span-level error typology and translator preference ([Karpinska and Iyyer, 2023](#)), and shared-task human assessment ([Wang et al., 2024](#)); we

do not introduce new axes for these because existing instruments already measure them and our pipeline does not change their behavior. The three axes A1/A2/A3 are precisely those that emerge at the intersection of *long-form* (where chapter- k identity decisions are visible only at $k + 50$) and *cross-cultural* (where the cultural substrate of the target differs from the source): A1 is the long-span generalization of the document-level coreference problem; A2 is the long-form extension of atomic-scale cultural adaptation (Khanuja et al., 2024; Cao et al., 2024; Yao et al., 2024); and A3 captures dependency structures specific to entities-with-culturally-bound-extensions, a class neither short-context nor atomic-scale work has had to address. We make no claim of exhaustiveness over all long-form-cross-cultural concerns—behavioral markers (bowing, register-marked address forms) sit at the A2/A3 boundary and are discussed as a known scoping question in §8—only that A1/A2/A3 are non-redundant and jointly load-bearing for the failure mode this paper is about.

Scope: the domestication regime in detail.

A2 as defined in §3 encodes a specific stance within the foreignization–domestication tradition of translation studies (Schleiermacher, 1813; Venuti, 2017): source-culture artifacts inside target-culture scenes are treated as failures *unless* explicitly framed as foreign imports. This is a *domestication-oriented* setting, appropriate to the commercial localization regime where the skopos prioritizes target-culture-native reading experience: video games, audiovisual content, and serialized fiction platforms whose readers expect a target-culture surface. We make this stance explicit rather than implicit because the alternative regime—foreignization, where preserving source-culture texture is a feature rather than a bug—is equally legitimate and motivates a different metric. Three classes of reference are correctly admitted by the current A2 definition: (i) *heritage scenes*, where a character framed as heritage-keeping retains source-culture artifacts (the third-generation Chinese-American keeping an altar); (ii) *tourist scenes*, where the narrative explicitly visits the source culture and source-culture markers are part of the setting; (iii) *cross-culturally shared references*, items that have crossed into the target culture as commonly-recognized terms (ramen in modern English-speaking contexts). The residue A2 measures is unflagged source-culture artifacts

inside scenes that have already been moved to the target culture. A symmetric foreignization-regime metric (rewarding source-culture texture preservation) is straightforward to construct from the same Hall-triad taxonomy (§A.3) by inverting the enclave decision; we report results under the domestication regime only (§3).

E Qualitative Error Analysis

This section provides extended qualitative examples demonstrating both the structural failures of the baseline methods and the remaining limitations of P_{full} .

E.1 Pipeline Successes: Resolving A1, A2, and A3

A1: Identity Drift

Source: 华夏基因进化大学 (*Huaxia Genetic Evolution University*)

Baseline: Fractures into 10 distinct institutions across the arc (e.g., *NMGEU*, *United States Genetic Evolution College*).

P_{full} : Locked globally as *the national academy for genetic advancement*.

A2: Cultural Residue

Source: Mandarin sci-fi adaptation, ch. 90: 真气; 基础丹田/丹田

Baseline: Retains source culture terms (*True Qi*, *Dantian*), violating the sci-fi register.

P_{full} : Localizes them as *somatic current* and *core somatic nexus*.

A3: Dependency-Chain Integrity

Source: 罗十六 (*Luo Shiliu*) 和 罗十六的母亲 (*Luo Shiliu's mother*)

Baseline: Severs the possessive chain, drifting the dependent between disconnected proper nouns (*Margaret Hayes*) and detached descriptors (*Mother-and-Child Haunt*).

P_{full} : Preserves the explicit lexical dependency chain, co-localizing them stably as *Ethan* and *Ethan's mother*.

Table 7: Qualitative comparison of structural failures versus P_{full} resolution.

Table 7 summarizes one instance per axis; the paragraphs below give the full context for each and a second A3 example.

A1: Institutional Identity Drift. In our Mandarin→English evaluation, the protagonist’s university (华夏基因进化大学, *Huáxià Genetic Evolution University*) is mentioned across 60+ chapters. The B2 (rolling summary) baseline suffers from severe compression loss, referring to the institution by over ten distinct names across the arc (e.g., *NMGEU*, *GAT*, *Golden City Genetic Evolution University*, *United States Genetic Evolution*

College). P_{full} stores a single canonical form (*the national academy for genetic advancement*) in its persistent registry, remaining completely stable.

A2: Cultural Residue. Without a persistent target-culture plan tied to the entity registry, B1 retains recurring source-culture terminology inside target-genre scenes. In chapter 90 of the same source, the entity 真气 belongs to a Chinese cultivation lexicon and is conventionally rendered as *True Qi* or internal energy, while 基础丹田/丹田 is commonly preserved as *Dantian* (energy core). These terms create cultural residue in the adapted sci-fi biometric register. P_{full} instead localizes them together as *somatic current* and *core somatic nexus*, preserving their narrative function while aligning them with the target ontology of bio-energy, genetic nodes, and energy fields.

A3: Dependency-Chain Integrity. Dependent entities must co-localize with their heads. In the same source, 周胜 (*Zhou Sheng*) heads the 金城特情署 (*Jincheng Special Ops Department*). In B1 and S1, these entities drift independently across chapters: the head becomes *Vance Zhou* or *Joe*, while the agency becomes *Alliance Security* or *CIA Special Operations Division*. P_{full} batches them together during Stage 3, producing the stable pair *Victor Sterling* and the *Genetic Security Bureau*.

E.2 Pipeline Failures and Limitations

F1: Shared-Surname Blanket Replacement (Task-Level Failure). When multiple unrelated source characters share a common surname, all variants struggle. In another Mandarin source, seven unrelated characters share the surname 刘 (*Liu*). P_{full} conservatively assigns them all the localized surname *Boone* (e.g., *Curtis Boone*, *Carpenter Boone*, *Mrs. Boone*), creating ambiguity. Stage 4 flags these collisions (e.g., “Character shares localized last name ‘boone’ with unrelated characters”), but the architecture cannot safely split them without explicit relational evidence.

F2: Cluster-Boundary Divergence (A3 Desync in P_{full}). While Louvain batching generally succeeds, dependents that cluster topically rather than relationally can diverge. In an English→German adaptation, the head *Charles* is localized to *Carl*. However, *Charles’s father* and *Charles’s doctor* clustered separately and were localized to *Christians Vater* and *Christians Arzt*.

F3: Honorific Residue (A2 Leakage). P_{full} ’s registry stores source-language honorific structures (*X-哥*, *老-X*). Stage 5 sometimes renders these as literal English compounds (e.g., *Bro Troy* for 小腿哥, or *Old Braddock* for 老程). While the entity name is localized, the relational pragmatic is not, which reads awkwardly to an American audience.

F4: Script-Split Artifacts. When the source text mixes traditional and simplified Chinese across chapters (e.g., 刘浪 vs. 劉浪), Stage 1 extracts them as separate entities if no normalization is applied, leading to diverging target names (*Leo Trent* vs. *Leo Thorne*). This is an easily correctable pipeline artifact but remains a failure mode in the current snapshot.

E.3 Cross-Cultural Generalization: Variance by Language Pair

The empirical results (Table 1) reveal that while P_{full} is highly effective overall, its performance profile varies significantly depending on the source and target culture pair.

Consistent Success on Dependency Integrity (A3). Across all three language pairs (Mandarin→English, English→German, and English→Hindi), P_{full} consistently achieves the lowest error rate on the dependency axis (m_{A3}). This confirms a broad architectural conclusion: the structural mechanism of Louvain-batched graph localization (Stage 3) generalizes across languages. Because it relies on the topological structure of the entity relationship graph rather than language-specific pragmatics, it successfully forces heads and dependents to co-localize regardless of the target culture.

Variance in Identity (A1) and Cultural Alignment (A2). In contrast, P_{full} ’s performance on A1 and A2 is sensitive to the language pair:

- **High-Resource, High-Proximity (English→German):** P_{full} dominates, achieving near-zero error rates across A1 (0.014) and A2 (0.055). The underlying LLM possesses strong priors for German naming conventions and cultural norms, allowing the Stage-2 cultural strategy and Stage-1 registry to operate near perfectly.
- **Moderate-Proximity (Mandarin→English):** P_{full} wins decisively

on A1, but narrowly trails the B2 baseline on A2 (0.124 vs. 0.121). As detailed earlier in this appendix, this slight A2 regression is primarily due to the pipeline awkwardly retaining literal translations of Chinese honorifics (e.g., *Bro Troy* or *Old Braddock*), a culturally specific artifact that simpler baselines incidentally drop.

- **High-Divergence (English→Hindi):** P_{full} loses to B2 on A1 (0.147 vs. 0.111) and to S1 on A2 (0.050 vs. 0.041). This suggests that when adapting into culturally divergent target systems like Hindi, the rigid global constraints of Stage 1 (fixed registry) and Stage 2 (global plan) can become brittle. In such cases, the flexible, sequential re-evaluation of the B2 rolling-summary baseline actually produces more coherent identity mappings than a strict top-down registry.

In summary, while the pipeline’s graph-based dependency resolution (A3) is universally robust, its strict global anchoring (A1/A2) thrives primarily when the underlying model has strong target-culture priors. For highly divergent pairs like English→Hindi, enforcing a rigid global strategy occasionally underperforms local, context-driven adaptation.