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

Anonymous Authors
Anonymous Affiliation

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

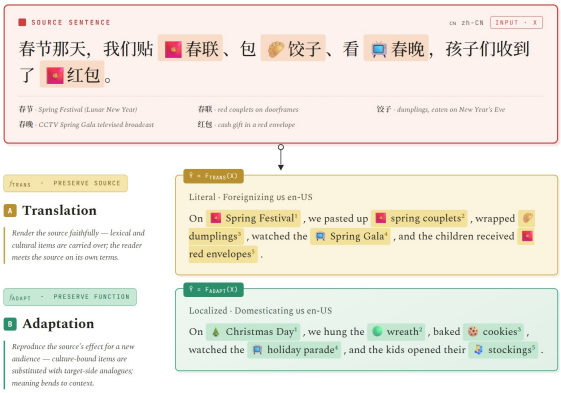

Figure 1: Translation vs. adaptation on a Chinese New Year passage. Translation preserves source-culture items (Spring Festival, couplets, Spring Gala, red envelopes) while adaptation swaps each for target-culture analogue that preserves narrative function (Christmas Day, wreath, holiday parade, stockings).

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 that do not yield to 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), organized around 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 two model families show reduction in identity errors by 65%, dependency errors by 64%, and cultural misalignment errors by 27%. A production A/B test on three Mandarin-to-English-USA serialized shows

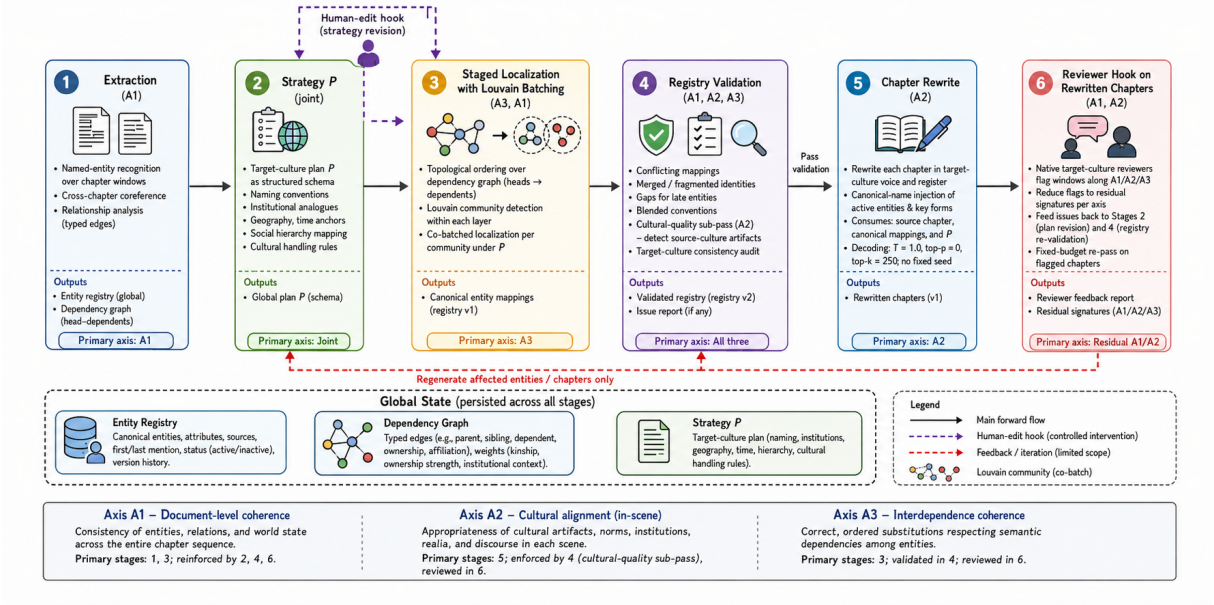

Figure 2: Six-stage pipeline. **S1** Extraction (A1) — entity registry + dependency graph. **S2** Strategy P (joint) — target-culture plan; human-edit hook before downstream stages execute. **S3** Staged Localization (A3, A1) — topological substitution with Louvain dependency-aware batching. **S4** Registry Validation (all axes; A2 via cultural-quality sub-pass). **S5** Chapter Rewrite (A2) — canonical-name injection into the rewriter prompt. **S6** Reviewer Hook on rewritten chapters (residual A1/A2), closing the loop with reader feedback. Persistent global state (registry, plan, dependency graph) is the architectural unit shared across stages.

demonstrates listener retention comparable to that for human-adapted content, reflecting viability of the framework in real-world deployment.

2 Related Work

Translation studies has long recognized that adaptation extends beyond linguistic transfer and includes cultural and functional transformation, as reflected in Skopos theory (Reiss et al., 2014), domestication and foreignization (Yang, 2010), and transcreation (Pedersen, 2014a). Recent work in document-level and literary Machine Translation (MT) has improved consistency beyond sentence level through document context (Jiang et al., 2022; Karpinska and Iyyer, 2023), structured memory, coreference modeling (Vu et al., 2024) and long-context LLM generation. However, they primarily target translation fidelity rather than systemic cross-cultural transformation (Singh et al., 2024), and treat consistency as a contextual retrieval problem rather than a persistent state-management issue.

A growing body of research studies cross-cultural adaptation and localization (Hershcovich et al., 2022; Liu et al., 2025), including culturally grounded translation (Anik et al., 2025), content adaptation (Liu et al., 2025) and cross-cultural evaluation (AlKhamissi et al., 2024). However, existing approaches operate at the passage or document level (Zhang et al., 2025a), and do not explicitly address the long-range coordination needed to maintain consistency across long-context narratives spanning ∼100 chapters (Gao et al., 2025). 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. Rather than treating adaptation as a sequence of independent localization decisions, the proposed framework models it as a stateful process requiring persistent consistency across the narrative arc.

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. Let $S = (c_1, \ldots, c_N)$ denote a source story of N chapters (typically $N \gtrsim 10^2$), written in source language L_S and source culture C_S, and let (L_T, C_T) denote the target language-culture pair. Long-form cross-cultural story adaptation seeks a target story

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

that preserves narrative semantics while rendering the story as native to the target culture. Preservation includes plot structure, character arcs, and causal dependencies; adaptation includes culturally appropriate entities, institutions, settings, and social conventions.

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

A2 – Cultural-system alignment. Cultural artifacts a appearing in target-culture scenes S' must be coherent with the target culture unless explicitly presented as foreign elements. This axis measures whether adaptation extends beyond linguistic translation to cultural localization.

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.

The three axes are non-redundant and jointly necessary: satisfying any subset does not imply satisfaction of the others.

4 System Architecture

To tackle the challenge of long-form adaptation (Section 3), we propose a five-stage pipeline that explicitly manages state across an entire book. Unlike approaches that process text chapter-by-chapter in isolation, our method gradually builds a shared context, including a registry of characters, their relationships, and a global adaptation strategy. This context is then shared across all subsequent steps, ensuring that the final rewrite remains consistent. We evaluate the importance of these individual stages in Section 6, while full prompt templates and hyperparameter details are available in Appendix A and Appendix C.

Stage 1: Entity and Relationship Extraction. Before rewriting any text, the system takes a first pass over the book to understand its characters and how they relate to each other. This involves extracting named entities, resolving coreferences across chapters, and building a dependency graph of relationships. The result is a persistent entity registry (mapping canonical names to all their mentions) and a graph of typed connections between these entities. By establishing this global context upfront, we avoid the inconsistency issues common in document-level translation (Karpinska and Iyyer, 2023; Vu et al., 2024) (addressing A1) and provide the structural foundation needed for later localization (A3).

Stage 2: Defining the Cultural Strategy P. Next, the pipeline establishes a global cultural strategy, denoted as P. This structured plan outlines how to adapt names, institutions, geography, and social hierarchies to the target culture. By finalizing these rules before any prose is actually rewritten, we ensure that the entire narrative is anchored to a coherent cultural setting (A2). This strategy also guides the subsequent stages, as both the localization process and the final rewriting step rely on P to make consistent decisions.

Stage 3: Graph-Based Localization. With the entities extracted and the strategy P defined, we localize the source entities. We process the entities in a topological order based on the dependency graph from Stage 1, ensuring that "head" entities are localized before their dependents. To handle related characters more effectively, we group them using Louvain community detection (Blondel et al., 2008) on the relationship graph, allowing the model to adapt closely tied characters (like a family) in the same batch. We found this community-based approach works better than simple topological sorting or greedy batching (see Appendix A for details). This stage is central to our localization goal (A3) and further reinforces consistency (A1).

Stage 4: Consistency Validation. Before moving on to generate the final text, we run an automated audit on the localized entity registry and the cultural strategy P. This validation step checks for fragmented identities, conflicting name mappings, or awkward blends of source and target conventions (like "Old Mr. Chris-xinsheng"). If the system flags any problematic entries, it specifically regenerates those parts rather than restarting the entire pipeline from scratch. This acts as an additional safeguard for cultural quality (A2).

Stage 5: Chapter Rewriting. In the final stage, the pipeline rewrites each chapter to match the target culture's voice and style. The model is provided with the original chapter, the localized entity mappings from Stage 3, and the cultural strat-

egy P. Crucially, we inject the localized names directly into the prompt to prevent the model from accidentally reverting to the source names or hallucinating new ones, a step we found vital for maintaining surface-level consistency (A1). This is where the actual cultural alignment takes shape in the prose, extending prior work on sentence-level adaptation (Cao et al., 2024a,b; Yao et al., 2024; Singh et al., 2024; Khanuja et al., 2024, 2025) to long-form narratives (A2).

5 Evaluation Framework

We evaluate our approach using two complementary methods. First, an automated evaluation scores adaptations on the three structural goals outlined in Section 3. Second, a live A/B test measures whether listeners stay engaged with our pipeline-adapted shows as much as they do with professionally human-adapted shows. The automated evaluation helps us understand the impact of different pipeline stages, while the A/B test shows how the system compares to a professional adapter in a real-world setting.

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.

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

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

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 also 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 $[0, 1]$. The 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 translated Chinese web novels on a commercial platform. Half the readers received a version drafted by our pipeline (Variant A), while the other half read a version drafted entirely by a human professional (Variant B). Both versions went through the platform's standard human editing process before publication, ensuring a fair comparison of the initial drafting method. We measured listener retention hour by hour. Since we only tested three shows on the Chinese-to-English (US) pair, we report the retention differences descriptively rather than making broad statistical claims. Full details are provided in Section 6.3.

6 Experiments

6.1 Datasets

We evaluate seven serialized stories 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 both German and Hindi. Each story is approximately 100 chapters long, ensuring that early localization decisions have long-term consequences, allowing us to effectively measure long-arc identity and dependency errors.

6.2 Experiment Variants

We compare four experiment variants. Each variant 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 first reads the full story and produces a global target-culture strategy specifying how names, places, institutions, social roles, and recurring cultural references should be adapted. Because this strategy is fixed before rewriting, chapters are adapted independently and in parallel from the chapter text conditioned on the global plan.

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 $1, \ldots, k-1$.

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 Production A/B Test Details

The A/B test was conducted on three Chinese-to-English adapted shows. Platform users were randomly assigned to either the human-drafted or AI-drafted variant upon starting the story, with the version hidden from the user interface. We tracked hour-level listener retention (R_h) from hour 6 to 30, alongside listener-day active users. We report the difference in retention per show with 95% confidence intervals. Details on human edit volumes for the AI variant are available in Appendix B.

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 do not come at the cost of plot fidelity, as m_{A0} scores remain stable (0.980 pooled vs. B1's 0.987).

Per-pair patterns. Breaking down the results by language pair reveals 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 also exceptionally strong for Mandarin→English and English→German. However, English→Hindi adaptation proves more challenging; it is the only pair where the rolling-summary baseline (B2) outperforms the pipeline on identity coherence. Cultural alignment (m_{A2}) shows more variance. While P_{full} decisively wins the English→German setting (0.055), it significantly improves upon B1 and S1 in Mandarin→English (37 errors over 298 chapters, $m_{A2} = 0.124$) but trails B2 by a single

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.

	$m_{A1}\downarrow$	$m_{A2}\downarrow$	$m_{A3}\downarrow$	$m_{A0}\uparrow$
Pooled (all pairs)				
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
Mandarin→English				
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**
English→German				
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**
English→Hindi				
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

error (36 errors over 298 chapters, $m_{A2} = 0.121$). Finally, English→Hindi slightly favors S1 on cultural alignment.

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 of this trial is to establish that P_{full}'s long-form narratives can achieve commercial viability. Across all three shows, the 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 with the human arm. For instance, Variant A significantly outperforms the human arm in AB1 at hr15 (-7.21 pp) and ties in AB3. This indicates that the structural consistency generated by P_{full} effectively translates to real-world listener retention at human-parity levels.

8 Discussion

To contextualize the quantitative improvements reported in Section 7, we perform a qualitative analysis of the outputs. Table 2 illustrates characteristic structural failures of the baselines against P_{full}'s pipeline resolution. This inspection confirms that

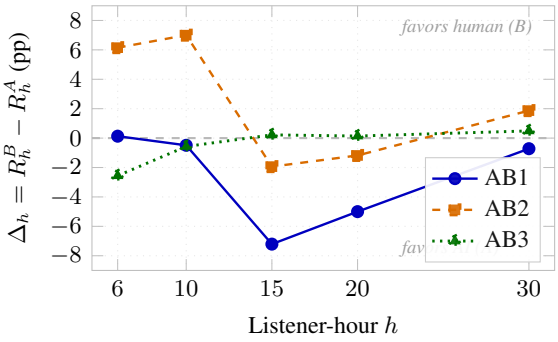

Figure 3: Production A/B retention against professional human adaptation, $(L_S, C_S, L_T, C_T) = (\text{zh}, \text{China}, \text{en}, \text{USA})$. Variant A: drafted by the five-stage pipeline P_{full}; Variant B: drafted end-to-end by a professional human adapter. Both arms undergo the platform's standard human validation step before deployment. $\Delta_h = R_h^B - R_h^A$ in percentage points (positive: human-drafted arm retains more; negative: AI-drafted arm retains more); the dashed line at $\Delta_h = 0$ marks no-effect. Allocated-user counts N_A/N_B and listener-day actives A/B: AB1 = 2,135/2,082 (306/319); AB2 = 769/796 (87/96); AB3 = 2,684/2,700 (570/583). Mixed per-show direction is reported rather than pooled.

the empirical gains on A1, A2, and A3 directly stem from the architectural state introduced in our pipeline.

A1: Resolving Identity Drift via Persistent State. Standard generation paradigms invariably leak identity state over long narrative arcs, regardless of whether they rely on sliding context windows or sequential summaries. The plan-then-adapt baseline (B1) illustrates this clearly: without a globally anchored target-culture surface form, parallel generation splinters a single Chinese character into nine distinct Westernized variants. The rolling summary approach (B2) suffers from a related failure mode driven by lossy compression. Under this bottleneck, a single university in our Mandarin→English evaluation fractured into ten entirely different institutional identities across a 100-chapter arc. P_{full} bypasses these failures by treating identity as an explicit, persistent state rather than implicit context. By extracting the entity registry upfront and injecting those locked forms directly into the local rewriter prompt, the architecture guarantees global consistency.

A2: Enforcing Cultural Alignment. Without a persistent target-culture plan tied to the entity registry, baselines often retain source-culture terminology inside scenes that otherwise require target-

A1: Identity Drift **Source:** (*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:** *Power of Genes*, chapter 90: ; / **Baseline:** Retains cultivation 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:** (*Luo Shiliu*) and (*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 2: Qualitative comparison of structural failures versus P_{full} resolution.

culture or target-genre adaptation. As shown in Table 2, B1 retains the source entity in *Power of Genes*, a term from the Chinese cultivation lexicon conventionally rendered as *True Qi* or internal energy. This creates cultural residue because it interrupts the adaptation's sci-fi biometric register. P_{full} instead maps the concept to *somatic current*, preserving its narrative function while aligning it with the target ontology of bio-energy, genetic nodes, and energy fields. The same pattern appears for /: rather than preserving *Dantian* (energy core), P_{full} renders the concept as *core somatic nexus*, a term that fits the adapted system's scientific vocabulary. Thus, P_{full} mitigates A2 errors by localizing recurring entities at the registry level, ensuring that culturally specific concepts remain consistent with the target genre and are localized together throughout the chapter.

A3: Maintaining Dependency-Chain Integrity. Dependent entities must co-localize with their heads to preserve narrative coherence, particularly when the dependent lexically inherits from the head. Baselines stripped of explicit relational state routinely sever these connections during possessive and institutional chains. In *Folklore Chronicles*, the source text tightly anchors a dependent entity to its head: (*Luo Shiliu's mother*). Without a relational graph to enforce this link, B2 unpredictably 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. This graph-aware batching guarantees that lexical and relational chains remain intact across hundreds of chapters, stably mapping the pair to *Ethan* and *Ethan's mother*.

Remaining Limitations. While the architecture resolves the primary failure modes of the baselines, it remains vulnerable to topological ambiguities in the source text. For example, when multiple unrelated characters share a common source surname, P_{full} conservatively assigns them the same target surname, inadvertently implying kinship. Furthermore, the Louvain graph occasionally partitions dependent entities into different batches than their heads (e.g., clustering "Charles's father" with family terms rather than with "Charles"), leading to localized name divergence within the pipeline. An extended error analysis detailing these failure modes across language pairs is provided in Appendix F.

9 Conclusion

References

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 Toushik 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 Underserved 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 Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024a. Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics (TACL)*.

Yong Cao, Min Lin, Ying Zhou, Yuke Liu, and Daniel Hershcovich. 2024b. Cultural adaptation of menus: A fine-grained approach. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*.

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 ACM CHI Conference on Human Factors in Computing Systems (CHI)*.

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 (EMNLP)*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Radina Dobreva, 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 Hashimoto. 2024. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

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.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, 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 (NAACL)*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam De Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, and 1 others. 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, Daniel Stamenkovic, Ronan Le Bras, Yejin Choi, and Allyson Ettinger. 2024. Evaluating creative short story generation in humans and large language models. *arXiv preprint arXiv:2411.02316*.

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, Clára He, and Graham Neubig. 2025. Towards automatic evaluation for image transcreation. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

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 (EMNLP)*. Best Paper Award.

Chen Cecilia Liu, Anna Korhonen, and Iryna Gurevych. 2025. Cultural learning-based culture adaptation of language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3114–3134.

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 (EMNLP)*.

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 Carmen Mangiron. 2013. *Game Localization: Translating for the Global Digital Entertainment Industry*. John Benjamins.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.

Daniel Pedersen. 2014a. Exploring the concept of transcreation–transcreation as more than translation. *Cultus: The Journal of intercultural mediation and communication*, 7(1):57–71.

Daniel Pedersen. 2014b. Exploring the concept of transcreation: transcreation as "more than translation"? *Cultus*, 7:57–71.

Katharina Reiss, Christiane Nord, and Hans J Vermeer. 2014. *Towards a general theory of translational action: Skopos theory explained*. Routledge.

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, Royal Academy of Sciences, Berlin. English translation in Venuti (ed.), The Translation Studies Reader, Routledge, 2000.

Catarina Silva, Yi-Hui Wu, Vera Cabarrão, and Helena Moniz. 2024. Cultural transcreation with LLMs as a new product. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (EAMT)*.

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 (EMNLP)*.

Gideon Toury. 1995. *Descriptive Translation Studies – and Beyond*. John Benjamins.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*.

Lawrence Venuti. 1995. *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.

Longyue Wang, Zhaopeng Tu, Yan Wang, Derek F. Wong, Lidia S. Chao, and Shuming Shi. 2024. Findings of the WMT 2024 shared task on discourse-level literary translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*.

Minghao Wu, George Foster, Lizhen Qu, and Gholamreza Haffari. 2023. Document flattening: Beyond concatenation-based context-aware translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 448–462.

Wenfen Yang. 2010. Brief study on domestication and foreignization in translation. *Journal of Language Teaching and Research*, 1(1):77–80.

Binwei Yao, Ming Jiang, Tanner Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Jian Zhang, Junyi Guo, Junyi Yuan, Huanda Lu, Yanlin Zhou, Fangyu Wu, Qiufeng Wang, and Dongming Lu. 2025a. Llm-driven completeness and consistency evaluation for cultural heritage data augmentation in cross-modal retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19418–19428.

Ran Zhang, Wei Zhao, and Steffen Eger. 2025b. How good are LLMs for literary translation, really? literary translation evaluation with humans and LLMs. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

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 Architecture details

This appendix collects per-stage operational detail referenced from Section 4. Material is added incrementally; placeholders mark sections to be filled in subsequent revisions.

A.1 Stage 1: Extraction

NER. [Placeholder: model, sliding-window size and stride, output schema.]

Cross-chapter coreference. [Placeholder: canonicalization tuple, embedding model, alias-merge cosine threshold, attribute-conflict handling.]

Relationship extractor. [Placeholder: typed-edge catalogue (parent, sibling, owner, dependent), per-edge confidence threshold, drop-edge policy.]

A.2 Stage 2: Cultural plan P

Schema. [Placeholder: JSON schema with keys for naming conventions, institutional analogues, geographic shifts, time-period anchors, social-hierarchy mapping, cultural handling rules.]

Worked example. [Placeholder: an abridged P for the zh-China $\rightarrow$ en-USA pair.]

A.3 Stage 3: Staged localization with Louvain batching

Louvain vs. simpler alternatives. *Topological-sort-only batching* (one entity per layer, no co-batching) respects head$\rightarrow$dependent ordering but does not co-locate semantically related entities, so cross-entity name affinity (e.g. "Sterling Villa" / "Sterling Estate" / "Sterling monogram") must be re-derived per call, weakening A1 consistency at the rewriter surface. *Fixed-depth BFS* co-locates by neighborhood but treats non-tree relationships (rival families sharing a contested institution; characters with affinities to multiple family units) inconsistently depending on entry point. *Greedy bin-packing by name affinity* co-locates by string similarity but ignores the relationship-graph topology entirely. *Louvain community detection* (Blondel et al., 2008) selects batches by modularity maximization on the weighted relationship graph, jointly respecting (i) the layer ordering induced by edge direction and (ii) the community structure induced by relationship density, with a natural batch granularity that varies across stories without per-story tuning.

Edge weights and Louvain parameters. [Placeholder: per-relationship-type weights (kinship, ownership, naming derivation, shared institution); resolution γ; recursive split policy for oversized communities.]

A.4 Stage 4: Registry validation

Audit categories. [Placeholder: full list of audit categories Stage 4 emits, per-axis grouping, cultural-quality sub-pass scope.]

Repair budget. [Placeholder: per-entity retry budget; escalation to deployment-time human-validation queue on terminal failure.]

A.5 Stage 5: Chapter rewrite

Decoding settings. Temperature $= 1.0$, top-p $= 0$, top-k $= 250$; no fixed seed. Reproducibility implications are discussed in Section **??**.

Canonical-name injection format. [Placeholder: per-entity injection schema, confidence-threshold gating, top-k dependent-surface-form policy.]

B Evaluation details

This appendix collects evaluation-pipeline detail referenced from Section 5 and Section 6. Material is added incrementally; placeholders mark sections to be filled in subsequent revisions.

B.1 Pass 1: Localization-sheet validator

Per-output entity registry. Each record carries a canonical source-language name, the proposed localized name, the mention map across chapters, and a structured localization-reasoning field. Pass 1 consumes the registry and invokes three coupled LLM checks.

Character-localization check. Returns, per character, a list of structured issues; each issue carries an `issue_type`, a free-text description, the list of affected mentions, and a severity tag (`high` / `medium` / `low`). The check covers mention-map inconsistency, one-to-many and many-to-one surname mappings, and conflicts with previously established naming patterns.

Entity-localization check. Returns per-entity issues with the same shape as the character check, plus a `dependency_reference` field recording the character or other entity the flagged item depends on. This dependency field is the load-bearing signal for m_{A3}.

Cultural-verification check. Classifies each localized character and entity as `LOCALIZED`, `ORIGINAL`, or `MIXED`, with confidence, free-text reasoning, and the cultural indicators that drove the decision.

B.2 Pass 2: Chapter-output error detector

Pass 2 scans the adapted chapter text in overlapping 10-chapter batches with 5-chapter overlap (to catch cross-chapter continuity errors) and emits issues across six categories: `Numbers`, `Cultural`, `Facts/Logic`, `Geography`, `Plot/Character`, and `Entity`. Both passes produce structured JSON outputs whose schemas are listed in Appendix C.

B.3 m_{A2} 7-category cultural-artifact taxonomy

The taxonomy below was locked post-pilot and drives the cultural-verification check in Pass 1 as well as the figure-level KEEP/SWAP illustrations in Figure 1.

C1 Food: onigiri, mapo tofu, baozi, dumplings, jianbing, mooncakes. **C2 Religious:** Shinto shrines, Buddhist altars, kowtow, joss paper, ancestor tablets. **C3 Social (relations / hierarchies):** -sama, -kun, dage (older brother), chaebol structure, jia jiao (family discipline). **C4 Architectural:** tatami, shoji screens, siheyuan courtyard, feng shui placement. **C5 Dress:** hanfu, qipao, kimono, geta, kanzashi, Mao suit. **C6 Monetary:** yen, yuan, fen, jiao, ten-thousand naming patterns. **C7 Institutional:** Confucius Institutes, gaokao, household registry (hukou), state-holiday Chinese New Year.

A scene is target-culture iff (i) the setting is geographically in C_T, (ii) characters are framed as C_T residents, and (iii) the narrative pretends the scene is C_T-native. An *enclave* is a source-culture artifact in a target scene without explicit foreign-import framing. Examples: "She grabbed an onigiri from the fridge" in an American suburban kitchen $\rightarrow$ ENCLAVE; "She ordered the unusual Japanese rice ball at the Asian fusion shop" $\rightarrow$ NOT an enclave (framed foreign); "As a third-generation Chinese American, she still kept a small altar" $\rightarrow$ NOT an enclave (heritage framing). Edge cases not counted as enclaves: similes without an actual artifact present; globally recognized brand names; explicit heritage framing; cross-cultural shared references that have entered C_T as commonly-recognized terms (e.g. ramen in modern English-speaking contexts).

The cultural-verification judge prompt asks for, per artifact, an origin classification (SOURCE/-TARGET/NEUTRAL), a scene classification (SOURCE/TARGET/FRAMED-FOREIGN), and an enclave decision; for each enclave the judge outputs the exact phrase, a category C1–C7, and a one-sentence rationale, in a single JSON object. Judges are instructed to be conservative (when in doubt, NOT an enclave). A per-pair surface-marker forbidden-token regex (e.g. zh-en US: "shinto, kowtow, feng shui, qipao, hanfu, yuan, jiao, mapo, baozi, onigiri, kimono, tatami, shoji, siheyuan, gaokao, hukou") supplements the judge count as a model-free sanity check. Calibration target: Cohen's $\kappa \geq 0.6$ per category on a 10-chapter two-annotator overlap; uncategorized clusters with ≥ 10 instances trigger a new category (C8, C9, ...).

B.4 Per-axis scoring

Each axis is normalized to a per-chapter rate (lower is better):

- m_{A1} **(identity coherence):** character-localization issues from Pass 1 + `Plot/Character` errors from Pass 2.
- m_{A2} **(cultural-system alignment):** cultural-verification entries labeled anything other than `LOCALIZED` + `Cultural` errors from Pass 2.
- m_{A3} **(dependency-chain integrity):** entity-localization issues from Pass 1 with a filled `dependency_reference` + `Entity` errors from Pass 2.

The three Pass-2 categories not used above (`Numbers`, `Facts/Logic`, `Geography`) target plot fidelity rather than cultural localization and are reported separately. Headline numbers in Section 7 are mean per-chapter scores across (story, generator) cells with standard error and a paired Wilcoxon signed-rank significance marker against P_{full}.

B.5 Judge panel disagreement protocol

[Placeholder: 1.5σ cross-family-judge disagreement threshold; high-disagreement subset reporting; same-family adjudication policy.]

B.6 Native-annotator calibration design

[Placeholder: per-pair recruitment (3 native annotators), screening criteria (serialized-fiction familiarity), rubric training procedure, inter-annotator agreement target before scoring begins, 50-chapter subset construction (proportional draw across stories), $\kappa_{jh} \geq 0.4$ headline-claim threshold.]

B.7 Extractor sensitivity analysis

[Placeholder: sensitivity of $m_{A1}/m_{A2}/m_{A3}$ to the choice of entity extractor used in Pass 1, comparing the structurally distinct extractor (headline) against the Stage 1 extractor of P_{full} as a robustness check.]

Table 3: Judge calibration. κ_{jh}: judge–human per-axis Cohen's κ on the 50-chapter native-annotator subset (bootstrap 95% CI). κ_{xj}: mean pairwise cross-judge κ. Δ_{sf}: same-family bias gap (issues/chapter). σ_{iss}/μ: across-judge coefficient of variation on issue counts.

Axis	κ_{jh}	κ_{xj}	Δ_{sf}	σ_{iss}/μ
A1	[TBD]	[TBD]	[TBD]	[TBD]
A2	[TBD]	[TBD]	[TBD]	[TBD]
A3	[TBD]	[TBD]	[TBD]	[TBD]

B.8 m_{A0} validation against native annotators

[Placeholder: 30-chapter zh-en-USA subset, two native annotators per chapter, beat extraction with importance tags following the m_{A0} rubric, blinded to LLM-judge output. Reported: judge–human beat-recall correlation, beat-importance agreement, severity-grade agreement on issue lists.]

B.9 Dataset characteristics

[Placeholder: per-pair story selection criteria, chapter-length distribution (median ≈3,200 tokens), genre composition, recurring-entity density beyond chapter 80, source-corpus license/access terms.]

B.10 Competing method configurations

B1 – long-context with planning. B1 comprises two generator calls. The first receives the entire source novel and produces a target-culture plan in the same JSON schema as Stage 2 of P_{full}: cultural mappings, entity-localization principles, narrative-voice strategy, genre framing, and naming rules. The second call receives the source novel together with the plan and produces the full adaptation. No pipeline state is maintained across calls, meaning no cross-chapter entity registry, no relationship graph, no batched localization, no validation pass. This isolates whether a long-context model augmented with an explicit cultural plan can recover the structured global state that P_{full} constructs.

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. Stages 2–5 operate at their P_{full} defaults; in the absence of a relationship graph, Stage 3's batched localization falls back to chapter-co-occurrence batching.

S2 – no strategy. S2 disables Stage 2 entirely. The chapter rewriter receives the Stage 1 entity registry, the target-culture parameters (L_T, C_T), and the source chapter, but no cultural plan, no narrative-voice strategy, no genre framing, and no naming rules. Stages 1 and 3–5 operate at their P_{full} defaults. Stage 3 continues to use the relationship graph for dependency-aware batching but operates without the cultural anchors that ordinarily constrain its naming decisions. The S2 ablation is structured to measure A2 regression; we describe S2's A1/A3 effect, if any, as observational rather than as part of the predicted-failure claim.

B.11 Intra-stage ablations

Within P_{full} we additionally test five intra-stage variants whose hypotheses are detailed below:

- **(a) Stage 1 relationship sub-pass off.** Predicted: large Δ_{A3} because Stage 3 can no longer recover dependency units.
- **(b) Stage 1 cross-chapter coref scope reduced to per-chapter.** Predicted: large Δ_{A1} because identity drifts across chapters when coref scope is local.
- **(e) Stage 3 batching algorithm: Louvain vs. topological-sort-only (the strongest informed alternative) vs. uniform-random at matched batch-size distribution (a weak-baseline floor).** Predicted: largest Δ_{A3} for uniform-random; smaller gap for topological-sort-only relative to Louvain.
- **(g) Stage 4 cultural-quality sub-pass off.** Predicted: largest single-component Δ_{A2}; foreign-enclave artifacts persist when the sub-pass is off.

- **(i) Stage 5 canonical-name injection off.** Predicted: medium-large Δ_{A1} at the surface even when the upstream registry is clean.

B.12 Production A/B trial details

[Placeholder: per-arm allocation procedure (random at first-listen, within-show), variant-identity concealment in the UI, recommender-parity checks across arms, allocated-user counts per arm (range 800–4,000), edit-volume distribution on Variant A (typically under one adapter-hour per chapter; full distribution to characterize).]

C Prompts

We collect the full prompt templates for each LLM call in the pipeline and in the evaluation framework. Curly-brace placeholders `{variable}` indicate runtime substitutions. Prompts not yet finalized are marked as placeholders.

C.1 Pipeline prompts

Stage 1: Entity extraction (NER, coreference, relationship). [Placeholder]

Stage 2: Cultural plan P generation. [Placeholder]

Stage 3: Staged localization. [Placeholder]

Stage 4: Registry validation (cultural-quality sub-pass). [Placeholder]

Stage 5: Chapter rewrite with canonical-name injection. [Placeholder]

C.2 Evaluation prompts

Pass 1: Localization-sheet validator. [Placeholder: combined character-localization + entity-localization + cultural-verification checks; structured-JSON output schema.]

Pass 2: Chapter-output error detector. [Placeholder: six-category issue detector (Numbers, Cultural, Facts/Logic, Geography, Plot/Character, Entity); 10-chapter sliding window with 5-chapter overlap.]

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}
```

D Extended related work

This appendix expands the placeholder 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 account of translator invisibility (Venuti, 1995), and the Skopos theory of translational action of Reiss and Vermeer (1984) all argue that translation is purposive: 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 audiovisual settings (Pedersen, 2014b), 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 whole-chapter context is necessary for stable per-chunk quality. 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, terminology, coreference, and quotation sub-categories on chapter-sized documents; Zhang et al. (2025b) extend literary-evaluation methodology with human-vs-LLM head-to-heads, and **?** inject source-side coreference into a context-aware MT model and report $>$1.0 BLEU gains on WMT En-De/En-Ru and multilingual TED, demonstrating the value of coreference signal for local reference stability even though the evaluation is on news 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; we evaluate on BlonDe-style discourse sub-scores and on WMT 2024 literary as commensurable anchors. Earlier work on *document substructure* for MT (Dobreva 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. (2024a) show this for recipes and Cao et al. (2024b) for menus; 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 framework 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 in-lab reader ratings.

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 (Hada 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., 2024; 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 entity-mapping recall metric across long spans (anchored to BlonDe annotation on a BWB Zh-En subset); m_{A2} is a 7-category foreign-enclave incidence metric; m_{A3} is a dependency-chain coherence rubric; 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 the Bradley–Terry MLE system-ranking protocol of Chiang et al. (2024), and extend the G-Eval-style reference-free scoring of Liu et al. (2023) with axis-decomposed rubrics. Known judge biases motivate our calibration protocol: length effects (Dubois et al., 2024; Saito et al., 2023), same-family self-preference (Panickssery et al., 2024), and judgement biases more broadly (Chen et al., 2024). Off-the-shelf single-judge protocols are uncalibrated for per-axis cross-cultural claims at long-form scale; the symmetric three-judge cross-family panel used in this paper (§5) is the minimum bar we judge sufficient. All judging is length-controlled per Dubois et al. (2024) and ordering-shuffled, with per-axis Cohen's κ and bootstrap CIs reported following the NLG-evaluation guidelines of van der Lee et al. (2021). Throughout the paper, the underlying generators and judges are transformer-based LMs (Vaswani et al., 2017).

E 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 $\rightarrow$ estate $\rightarrow$ 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

Table 4: Coverage of structural axes by the five closest prior systems. • = directly addressed; ○ = partially addressed; — = not in scope. A1 long-span identity coherence; A2 cultural-system alignment within scenes; A3 dependency-chain integrity; LF = long-form output ($\geq$ 100 chapters of consistent semantic state); Adapt = adaptation (vs. literal translation) is the stated objective.

System	A1	A2	A3	LF	Adapt
? (**?**)	○	—	—	○	—
? (**?**)	○	—	—	•	—
? (**?**)	•	—	—	—	—
Khanuja et al. (2024)	—	•	—	—	•
Cao et al. (2024a)	—	•	—	—	•
This work	•	•	•	•	•

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" $\rightarrow$ "Sterling" cannot reach the semantically-dependent estate name. Each of these single-failure cells is observed in actual outputs from the minimal-pipeline floor and from intermediate P_k variants; their joint observability is what the empirical axis-correlation matrix in §7 quantifies.

Why these three axes (and not others). Prior literary-translation evaluation already operationalizes voice/register fidelity (BlonDe (Jiang et al., 2022)), idiomatic naturalness (Karpinska and Iyyer, 2023), and discourse-level pacing (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., 2024a; 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, 1995): 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 7-category taxonomy (§B.3) by inverting the enclave decision; we report results under the domestication regime and discuss the foreignization symmetry in §8.

F 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}.

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

A1: Institutional Identity Drift. In the Mandarin→English adaptation of *Power of Genes*, the protagonist's university (, *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 *Power of Genes*, the source entity belongs to a Chinese cultivation lexicon and is conventionally rendered as *True Qi* or internal energy, while / 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 *Power of Genes*, (*Zhou Sheng*) heads the (*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*.

F.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 *Folklore Chronicles*, seven unrelated characters share the surname (*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 the English→German adaptation of *Royal Accident*, 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-, -X*). Stage 5 sometimes renders these as literal English compounds (e.g., *Bro Troy* for , or *Old Braddock* for). 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., vs.), 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.

F.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.