PAUSE: Editable Strategy Artifacts for Long-Form Cultural Story Adaptation

Taaha Kazi [1] **Vasu Sharma** [1] **Mohammad Saifullah** [1] **Abdur Rahman** [1]

Abstract

Generative AI systems increasingly mediate cultural adaptation, but their cultural decisions are often hidden inside prompts, transient model plans, or final prose. We study PAUSE (**P**ause-**A**nd-**U**pdate **S**trategy **E**diting), an intervention that exposes an editable adaptation strategy as a human control surface for cultural decisions in long-form story adaptation. The strategy is a structured artifact that can be inspected, edited, and then projected through downstream character, entity, and chapter-localization stages. In two Chinese-source serialized novels, we test whether human edits to this strategy propagate into chapter-level prose. Across 9 edited-vs-control chapter comparisons, judges select the edited-strategy output in all 9; a marker audit shows target markers in $8/9$ edited outputs and $0/9$ controls, with forbidden markers absent from edited outputs and present in all controls. We frame these results as a smoke-scale edit-adherence study, not a claim that the outputs are culturally authoritative or literary-quality improvements. PAUSE offers one practical way to make AI-mediated cultural adaptation more inspectable and contestable before decisions propagate through long-form generation.

1. Introduction

Generative AI systems increasingly mediate cultural artifacts, but the cultural decisions they make often remain hidden inside prompts, transient model plans, or final prose. In long-form narrative adaptation, these decisions become especially consequential because names, settings, institutions, social hierarchies, and genre register must remain stable across many chapters (Wang et al., 2023a;b; 2024; Jin et al., 2024). When the task is also cross-cultural, the problem is not only fluency: translation may require transcreation, entity localization, and context-sensitive cultural substitution (Liu et al., 2025; Conia et al., 2024).

Current LLM workflows give human authors and editors two familiar intervention points. They can steer through a prompt before generation begins, hoping that the instruction survives the model's internal planning and many downstream calls. Or they can post-edit the final prose, after a global adaptation plan has already shaped names, relationships, institutions, and style. Both points are useful, but neither gives the editor a direct handle on the cultural decision layer itself. For AI used as a cultural technology, this opacity matters: cultural adaptation can erase source specificity, flatten cultures into stereotypes, and automate parts of creative labor. The goal should not be to make such transformations automatically correct, but to make them inspectable and contestable before they propagate.

This paper studies PAUSE, a pause-and-update strategy editing intervention built around an editable, interpretive intermediate strategy. In our long-form adaptation pipeline, an LLM emits a structured *adaptation strategy* that captures the intended target setting, naming conventions, institutional analogues, register choices, and entity-localization principles. We expose this strategy as a JSON artifact, allow a human editor to revise it, and resume the pipeline from the edited copy without fine-tuning the model. The strategy is consumed by metadata-localization stages and projected into chapter writing through localized character records, entity records, mention mappings, and, in the patched experimental run, strategy-conditioned adaptation prompts; the final polish pass is deliberately strategy-blind. The empirical question is therefore not whether the system improves translation in general, but whether edits to this visible strategy artifact reach the prose that readers see.

We evaluate three edited-vs-control runs. Across the resulting 9 chapter-level pairwise comparisons from two Chinese-source web novels, judges select the edited-strategy output in all 9. A deterministic marker audit points in the same direction: target markers appear in $8/9$ edited outputs and $0/9$ controls, while forbidden markers appear in $0/9$ edited outputs and $9/9$ controls. We treat this as a smoke-scale edit-adherence result. It shows that schema-aligned strategy edits can propagate into chapter prose; it does not establish cultural adequacy,

[1]Pocket FM. Correspondence to: Taaha Kazi <taaha.kazi@pocketfm.com>.

Accepted at the 1st Workshop on Culture x AI: Evaluating AI as a Cultural Technology, ICML 2026.

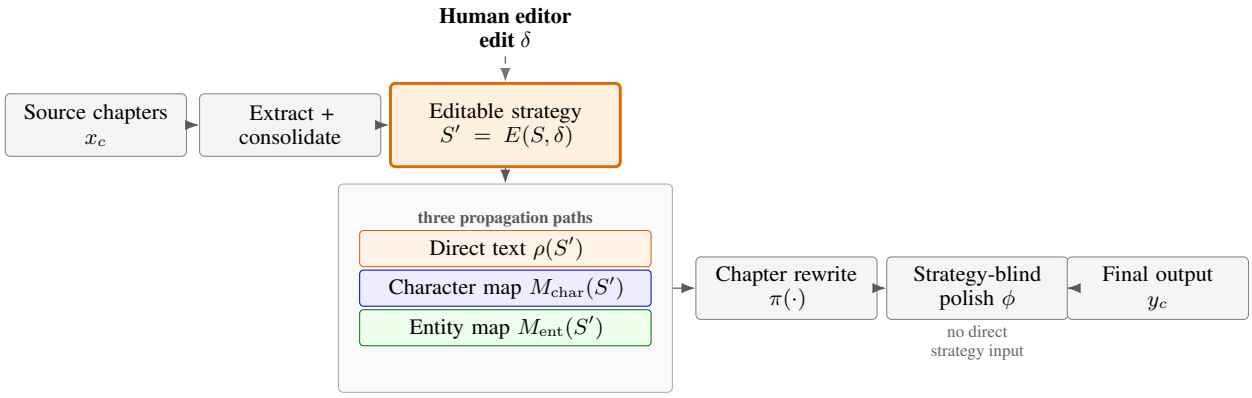

Figure 1. PAUSE mechanism. The pipeline captures an adaptation strategy S from source chapters; a human edit δ produces $S' = E(S, \delta)$. The edited strategy reaches chapter rewrite through direct prompt text, character mappings, and entity mappings, while the strategy-blind polish pass receives only the rewritten draft.

literary quality, or full-novel coherence.

Contributions. (i) We introduce PAUSE, identifying the LLM-generated adaptation strategy as an editable control surface for cultural decisions in long-form narrative adaptation. (ii) We implement its pause/edit/resume mechanism around this artifact without model fine-tuning. (iii) We characterize three propagation paths from strategy edits into prose: direct prose conditioning, indirect character mapping, and indirect entity mapping. (iv) We report a smoke-scale propagation study over 9 chapter comparisons, supported by pairwise judgments and deterministic marker checks. (v) We discuss the limits of this evidence and the prerequisites for making cultural adaptation workflows more inspectable and contestable. Figure 1 gives the mechanism overview before we separate the surrounding pipeline substrate from the edit-propagation experiment.

2. Related Work

Long-form literary translation and cultural adaptation. Document- and chapter-level MT motivates our focus on decisions that must remain consistent beyond a single sentence or scene (Maruf et al., 2021). Recent WMT literary-translation shared tasks and the Disco-Bench benchmark make discourse-level Chinese–English literary translation a concrete evaluation setting (Wang et al., 2023a;b; 2024); Jin et al. (2024) further study chapter-to-chapter context. A parallel line argues that translation across cultures cannot be reduced to literal transfer: culturally aware NLP and cross-cultural MT require attention to entities, social context, and target-culture knowledge (Liu et al., 2025; Conia et al., 2024). Our work uses this setting, but asks a different question: whether the cultural decision layer can be exposed as an editable artifact before it affects many chapters of prose.

Human control through intermediate writing artifacts. Human-AI writing systems have often exposed control at the text-editing or suggestion level, such as collaborative editors and interaction datasets for creative writing (Coenen et al., 2022; Lee et al., 2022). Long-form generation systems instead emphasize plans, outlines, memories, and revision loops as intermediate objects that guide downstream drafting (Yao et al., 2019; Yang et al., 2022; Schick et al., 2022). We follow this intermediate-artifact pattern, but at a different granularity: the edited object is not a local passage or plot outline, but a cross-chapter adaptation strategy whose values are compiled into names, entities, and chapter-localization prompts.

Evaluation of open-ended generation. Creative and cultural adaptation tasks rarely have one reference answer, so pairwise and rubric-based LLM judgments are common but must be narrow and bias-aware (Zheng et al., 2023; Liu et al., 2023). We therefore frame the judge task as edit adherence rather than global quality, use judges from a different model family than the generator, and add a deterministic marker audit. The marker audit follows the broader evaluation pattern of decomposing long-form outputs into targeted, checkable units, as in atomic factuality evaluation (Min et al., 2023). Together these choices position our evidence as a small propagation test, not a benchmark for cultural quality or literary translation.

3. Pipeline Substrate

We describe only the parts of the long-form adaptation pipeline needed to make the strategy-edit experiment legible. The pipeline separates a *metadata layer*, which extracts and localizes characters, entities, relationships, and mention mappings, from a *text layer*, which uses those localized records to rewrite chapters. In the inspected implementation these are separate entrypoints, but conceptually they are

one adaptation workflow. The model and judge details are reported with the evaluation protocol.

Metadata layer For a source novel and target language/culture, the metadata layer runs: (i) per-chapter extraction of characters, entities, and relationships into a knowledge graph (KG); (ii) a strategy step that condenses the KG, family structure, entity graph, and target setting into a structured adaptation plan; (iii) character localization; (iv) entity localization; and (v) validation and cultural-status checks over the localized records. The strategy is a JSON document containing declarative rules for names, families, places, institutions, social address, genre terms, and cultural references. Its schema is source-conditional: editors must revise the keys the strategy LLM actually emitted, rather than assume a fixed canonical schema.

Localized records and mention projection. Character localization produces canonical source-to-target name records, including first names, surnames, honorifics, and mention maps. Entity localization first handles locations, then localizes other entities by graph segment so related institutions, places, objects, and cultural items can be adapted together; graph segmentation uses Louvain-style community detection (Blondel et al., 2008). After localization, bridge scripts merge source and localized records and project them into `mentions_by_chapter.json`, the machine-readable handoff from metadata to chapter writing.

Text layer Chapter writing is a separate prototype entrypoint. It reads the source DOCX and `mentions_by_chapter.json`, optionally augments mentions from merged character/entity JSON, and writes each chapter in two passes. Step 1 performs adaptation and mention replacement; in the patched experimental run, relevant strategy text can also be injected into this localize prompt. Step 2 rewrites the draft for fluency and style. The polish pass is *strategy-blind*: it does not see the strategy JSON directly, so the evaluation tests whether strategy-sensitive decisions introduced before polishing survive into the final chapter text.

Under PAUSE, we can summarize this claim boundary compactly. For source chapter x_c, captured strategy S, human edit δ, chapter rewrite π, and strategy-blind polish ϕ:

$$\begin{aligned} S' &= E(S, \delta), \\ \mathcal{P}(S') &= \big(\rho(S'), M_{\text{char}}(S'), M_{\text{ent}}(S')\big), \\ \tilde{y}_c &= \pi\big(x_c, \mathcal{P}(S')\big), \\ y_c &= \phi(\tilde{y}_c). \end{aligned} \tag{1}$$

Equation 1 makes the claim boundary explicit: E is the edit operation, $\tilde{y}_c$ is the pre-polish chapter draft, and y_c is the final chapter output. $\mathcal{P}(S')$ is only a compact notation for the three strategy-derived paths in Figure 1: $\rho(S')$ is direct strategy text injected into the localize prompt, while M_{char} and M_{ent} are the indirect character and entity mappings. A strategy edit does not control every sentence; it can affect prose only through these paths.

4. PAUSE: Editable Strategy Steering

PAUSE is deliberately small: expose the strategy artifact, edit it, and resume from the edited copy. This places human input between prompt-time steering and final-prose post-editing. The editor is not asked to rewrite the whole output, and the model is not fine-tuned; the editor changes the visible cultural decision layer that downstream metadata-localization stages use.

Pause/edit/resume hook We implemented a local prototype hook around the generated strategy file. In the first invocation, the pipeline runs through extraction and the strategy LLM call, writes the strategy JSON to disk, and halts. The editor revises that file. A second invocation loads the edited JSON, continues metadata localization from that edited strategy, and projects the resulting localized records into chapter writing. The interface is therefore just the emitted schema: any text editor can author an edit, but the edit must target fields that actually exist in the captured strategy.

What is editable An edit is any change to the JSON document that the strategy step emits, applied between the pause and the resume. In the family-musical-drama case study (*Love and Strings*), the editor reframes the localization from a New York classical-piano family to a Nashville country-guitar family. **One simple field edit** changes the target-setting prose:

> **Before.** "...invoking the architecture, weather, and atmosphere of New York City and the broader New England area (e.g., brownstones, autumn foliage, the pace of Manhattan life)."
>
> **After.** "...invoking the architecture, weather, and atmosphere of Middle Tennessee: rolling hills, limestone bedrock, hot summers, the neon glow of Lower Broadway, and the quiet elegance of Franklin's antebellum homes."

Other changes in the same edit affect surname conventions, institutions, and honorifics. Appendix A shows a longer before/after excerpt. Whether any such edit actually reaches final prose is empirical; Section 5 evaluates that propagation.

5. Evaluation

We evaluate PAUSE through *strategy-edit propagation*: whether a human edit to the captured strategy changes the

Table 1. Edited-vs-control propagation runs. G = *Power of genes*; S = *Love and Strings*.

Case	Edit	Ch.	Wins
G-2	Denver, CO → Oakdale, IL	2	2/2
G-99	Pittsburgh, PA → Oakdale, IL	5	5/5
S-20	NYC/Ashford → Nashville/Branson	2	2/2
Total		9	9/9

final chapter prose in the intended direction. This is a smoke-scale study, not a full benchmark. The controlled evidence covers two Chinese-source serialized novels, *Power of genes* and *Love and Strings*, each with 99 native chapters, localized to American English.

Protocol Each comparison contains one source chapter, one control localization resumed from the unedited captured strategy, one edited localization resumed from the same captured strategy after the human edit, and an edit description specifying target and forbidden phenomena. The two resumes share the same source, captured baseline strategy, and downstream configuration; the only intended input difference is the strategy edit. A blinded pairwise judge sees the edit description and the two outputs, then selects the output that better instantiates the edit while avoiding forbidden remnants of the unedited strategy. This focused pairwise setup follows common LLM-judge practice for open-ended outputs while keeping the criterion to edit adherence rather than global quality (Liu et al., 2023; Zheng et al., 2023). Because our judges are LLM-only, the pairwise results measure edit adherence, not cultural authority or representativeness of target-culture values.

Pairwise result Table 1 summarizes the three runs. The first *Power of genes* run edited a Denver/Colorado setting into a fictional Midwestern town, Oakdale, Illinois. The second re-captured the strategy from the full 99-chapter source and edited a Pittsburgh/Pennsylvania setting to Oakdale. The *Love and Strings* run changed the target setting and music culture from a New York classical-piano family to a Nashville country-guitar family. Across all 9 judged chapters, judges selected the edited-strategy output in all 9. The result shows that schema-aligned edits can reach chapter prose; it does not show that the system solves cultural adaptation or literary quality.

Marker audit To reduce dependence on judge preference, we also count edit-specific surface markers in the saved outputs. Per-chapter target and forbidden token counts are computed by case-insensitive word-boundary regex match against a fixed token list derived from each edit description, applied to the localized chapter outputs. For example, the Oakdale edits target tokens such as *Oakdale* and *Illinois*, while forbidding baseline tokens such as *Denver* or *Pittsburgh*. Table 2 gives the aggregate.

Table 2. Marker audit over the 9 judged localized chapter outputs.

Side	Target present	Forbidden present
Edited	8/9	0/9
Control	0/9	9/9

Table 3. Secondary rubric audit (n=18 order-shuffled judgments).

Axis	Edited	Control	Δ
Edit adherence	4.67	1.00	+3.67
Source faithfulness	4.22	4.22	+0.00
Target-culture naturalness	4.50	4.06	+0.44
Prose quality	4.44	4.17	+0.28

The one edited output without a target marker is an interior testing-room scene that names neither city nor institution, so we treat it as a no-op opportunity rather than a contradicted edit. The forbidden-marker separation is complete: no edited output retains the forbidden baseline markers, while every control output contains at least one.

Secondary rubric audit As a secondary guardrail, we rejudged all 9 chapter pairs with Opus 4.7 using shuffled A/B ordering and a four-axis 1–5 rubric. This yields 18 forced-choice judgments; the edited side is selected in $18/18$, and all 9 chapters have edited-side wins under both orderings. These scores are transcribed into the aggregation script; raw judge-response files are not part of the current reproducibility package. Table 3 shows that the main difference is edit adherence: source faithfulness is tied, while target-culture naturalness and prose quality are similar in this small LLM-judged sample. We treat this as a secondary audit, not as evidence that the outputs are culturally correct.

Run settings The reported strategy, localization, and adaptation calls used Gemini 2.5 Pro as the main generation model; two pairwise runs were judged by GPT-5 with high reasoning effort and the five-chapter run by Opus 4.7. The strategy step and chapter rewrite use live model aliases without a fixed seed. Chapter rewrite is therefore nondeterministic in absolute output, but each edited/control comparison is resumed from the same captured baseline strategy and processed with the same downstream configuration. This controls for the captured strategy and run setup, with residual stochasticity remaining a limitation.

6. Discussion and Limitations

Discussion PAUSE does not merely make cultural adaptation "automatically correct"; it makes it transparent, controllable, and verifiable. The system exposes a consequential decision layer that is visible, editable, and testable before it propagates through any long-form generation workflow.

This is the core claim of this paper: the most useful human interface is not another prompt box or final-prose post-edit, but the model's own intermediate cultural plan presented as an inspectable control surface. Cultural adaptation carries risks: erasing source specificity, flattening cultures into stereotypes, and automating creative labor. The editable strategy directly addresses these risks by making transformations inspectable and contestable before they reach the prose stage.

Limitations and future directions Our evaluation covers two source novels and Chinese-to-American-English localization. The natural next step is to widen the study to more stories and more language-culture pairs (the mechanism is already in use internally across more than a dozen pairs, e.g., English→Hindi and Korean→Spanish) and to confirm that the edit-adherence signal generalizes.

Conclusion A pause-and-resume hook on the strategy step lets a human revise the model's intermediate cultural plan before it propagates into chapter prose. We verify this propagation under three orthogonal signals—a blinded pairwise judge, a model-free surface-marker audit, and a multi-axis re-judge with ordering shuffle. The result is a precise and practical control surface for the global creative choices of long-form localization.

References

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

Coenen, A., Ippolito, D., Hedayatnia, B., Bakst, A., Bosma, M., Yuan, A., Borgeaud, S., Pushkarna, M., and Cai, C. Wordcraft: A human-AI collaborative editor for story writing. In *Proceedings of the 27th International Conference on Intelligent User Interfaces Companion*, pp. 42–44, 2022.

Conia, S., Lee, D., Li, M., Minhas, U. F., Potdar, S., and Li, Y. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of EMNLP*, pp. 16343–16360, 2024.

Jin, L., An, L., and Ma, X. Towards chapter-to-chapter context-aware literary translation via large language models. *arXiv preprint arXiv:2407.08978*, 2024.

Lee, M., Liang, P., and Yang, Q. CoAuthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022.

Liu, C. C., Gurevych, I., and Korhonen, A. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics (TACL)*, 13:652–689, 2025.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of EMNLP*, 2023.

Maruf, S., Saleh, F., and Haffari, G. A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys*, 54(2), 2021.

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P. W., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of EMNLP*, 2023.

Schick, T., Dwivedi-Yu, J., Jiang, Z., Petroni, F., Lewis, P., Izacard, G., You, Q., Nalmpantis, C., Grave, E., and Riedel, S. PEER: A collaborative language model. *arXiv preprint arXiv:2208.11663*, 2022.

Wang, L., Du, Z., Liu, D., Cai, D., Yu, D., Jiang, H., Wang, Y., Cui, L., Shi, S., and Tu, Z. Disco-Bench: A discourse-aware evaluation benchmark for language modelling. *arXiv preprint arXiv:2307.08074*, 2023a.

Wang, L., Tu, Z., Gu, Y., Liu, S., Yu, D., Ma, Q., Lyu, C., Zhou, L., Liu, C.-H., Ma, Y., et al. Findings of the WMT 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of LLMs. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, pp. 55–67, 2023b.

Wang, L., Liu, S., Liu, C., Lyu, C., Tu, Z., et al. Findings of the WMT 2024 shared task on discourse-level literary translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*, 2024.

Yang, K., Tian, Y., Peng, N., and Klein, D. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of EMNLP*, 2022.

Yao, L., Peng, N., Weischedel, R., Knight, K., Zhao, D., and Yan, R. Plan-and-write: Towards better automatic storytelling. In *Proceedings of AAAI*, 2019.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

A. A worked example of an edit

To make the Section 4 description concrete, we show a before/after excerpt from the family-musical-drama case study (*Love and Strings*). The author reframes a New York classical-piano "Ashford" family as a Nashville country-guitar "Branson" family. The two strategy sections shown are consumed by downstream localization, allowing the edit to reach chapter writing without rewriting prompts.

Naming convention edit.

> **Before.** "All character names belonging to the primary narrative (the American setting) must be fully Americanized with common Anglo-Saxon, Latinate, or otherwise recognizable American names. The specific pronunciation quirk of the source surname (Tán/Qín) will be discarded in favor of a clear, consistent American surname. . . . A new, consistent naming scheme will be established (e.g., The '**Ashford**' family)."
>
> *Foreign-character rule:* "Should any character be explicitly introduced as foreign (e.g., a guest conductor from Germany, an exchange student from Japan), their name will retain its original phonetic structure to serve its narrative function as an outsider."
>
> **After.** ". . . A new, consistent naming scheme will be established (e.g., The '**Branson**' family)."
>
> *Foreign-character rule:* "Should any character be explicitly introduced as foreign (e.g., **a record producer from Sweden, a touring fiddler from Ireland**), their name will retain its original phonetic structure . . . "

Honorifics-and-titles edit. The original strategy specifies a complete title-mapping rubric tied to the classical-music milieu; the edit rewrites the entries that carry milieu-specific connotations and leaves the rest unchanged.

> **Before.** "*Patriarchal Titles:* 'Old Master' becomes '**The Maestro**' in public/professional contexts . . . 'Big Master/Current Head' becomes '**Mr. Ashford**'. *Female Family Heads:* 'Senior Aunt' becomes 'Ms. Catherine' or '**Ms. Ashford**'. The matriarch's professional title ('Professor') is used where appropriate ('**Dr. Ashford**')."
>
> **After.** "*Patriarchal Titles:* 'Old Master' becomes '**The Legend**' in public/professional contexts **(echoing the reverence given to iconic guitarists and bandleaders in Nashville)** . . . 'Big Master/Current Head' becomes '**Mr. Branson**'. *Female Family Heads:* 'Senior Aunt' becomes 'Ms. Catherine' or '**Ms. Branson**'. The matriarch's professional title ('Professor') is used where appropriate ('**Dr. Branson**')."

What this illustrates. Both edits are intentional interpretive choices stored as prose values under existing strategy keys. No new top-level keys are introduced; no rule grammar is invoked. The edited decisions are projected through character localization, entity localization, mention mappings, and the chapter rewrite; the final polish pass does not see the strategy directly. The corresponding chapter-level outputs that result from this edit are reported in Section 5.