Journal Title
XX(X):2–29
©The Author(s) 0000
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Functional equivalence in cross-context text adaptation: an analyze–generate–verify approach with large language models

Cheng-En Tsai[1], Fanfan Chen[2], and Jane Yung-jen Hsu[3]

Abstract

This study examines whether large language models (LLMs) can convert open-domain texts into representations usable by symbolic research and whether these representations support cross-context adaptation. We define functional equivalence as reconstructing a comparable reader-facing cognitive and emotional transition despite changes in characters, settings, cultural elements, and wording. We propose an analyze–generate–verify architecture: the analyzer extracts an evidence-linked specification of propositions, alternative readings, beliefs, motivations, association paths, and emotional contrasts; the generator realizes these functions in a new context; and the verifier independently checks each candidate for acceptance, repair, or abstention. Experiments on Chinese jokes support verification's diagnostic and selective-release value, but show no significant acceptance gains over matched baselines. Representation errors remain a constraint, and reader-level functional equivalence requires further validation.

Keywords

cross-context text adaptation, functional equivalence, large language models, neurosymbolic artificial intelligence, Chinese humor, effect verification

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

Introduction

The effect of a text depends on its content, on what readers already know, and on the context in which it is read. Adapting a text to a new audience or setting therefore involves more than substituting words: a fable compresses social experience into characters and causal relations, a story spreads an emotion across events, and a joke relies on a late cue that forces the reader to reinterpret what came before. Research on style transfer and cultural transcreation accordingly separates surface change from content preservation, cultural appropriateness, and reader effect [Gero et al.(2019), Mir et al.(2019), Yuan et al.(2026), Wu et al.(2025)].

LLMs produce fluent rewrites, but fluency does not guarantee that the relations behind an effect survive. A model may keep the words and drop a crucial inference, or write a plausible new text whose effect differs from the source; counterfactual story rewriting shows that a local change requires reconstructing the downstream event chain [Qin et al.(2019), Chen et al.(2022), Li et al.(2026)]. This paper asks whether a model can make a source text's effect structure explicit and use it to reconstruct a comparable transition in a new context.

Neurosymbolic motivation. Symbolic representations support comparison, checking, and localized diagnosis, but acquiring them from open-domain text has long required manual knowledge engineering [Kelly et al.(2023)]. Gordon showed that an LLM can decompose a narrative poem into simple sentences and incrementally acquire predicates and constants [Gordon(2025)]. We pursue both directions of the neurosymbolic exchange: the LLM proposes symbols from unannotated text, and the symbolic layer constrains generation and supports executable, localized verification, in the spirit of LLM-Modulo architectures [Kambhampati et al.(2024), Gundawar et al.(2024)]. The symbolic layer contributes only if it improves generation or diagnosis over baselines that receive the same information or the same number of model calls; our experiments are designed around these comparisons.

Why humor. Jokes typically establish an initial reading and then use a late cue to force reinterpretation [Attardo(2002), Suls(1972)]. Consider a joke in which a scientist's wife has twins; when a priest invites both children to be baptized, the

[1]Graduate Institute of Networking and Multimedia, National Taiwan University, Taiwan
[2]Department of Creative Technologies and Product Design, National Taipei University of Business, Taiwan
[3]Department of Artificial Intelligence, Chang Gung University, Taiwan

Corresponding author:
Cheng-En Tsai, National Taiwan University, Taiwan.
Email: d11944008@csie.ntu.edu.tw

scientist replies that one is enough, because the other is the control group. An adaptation to a technology workplace must keep a reason for the ritual, the point at which the experimental frame becomes recoverable, and an ending that stops there, while replacing the scientist, the twins, and the baptism: for instance, a product manager's relatives propose a housewarming ritual, and he agrees that the living room will receive it and the study will serve as the control group. An adaptation fails when a character lacks motivation, events are reordered, an association becomes too remote, or a sentence after the punchline explains the mechanism; it also fails, as an adaptation, when it retells the source in new words. LLMs can recognize such endings without reliably explaining their function [Cheng and Utsuro(2026)], and Chinese humor adds cultural and phonological dependencies [Zhang et al.(2019), Tseng et al.(2020a), He et al.(2025)]. Humor is therefore a sensitive stress test for adaptation; other genres need their own success criteria.

Beyond humor. The same preservation problem arises in fable simplification, plain-language public information, and cultural transcreation [Yuan et al.(2026), Wu et al.(2025), Alva-Manchego et al.(2020), Ayre et al.(2024)]. Each changes surface detail while retaining relations needed by the intended audience. These relations often concern implicit knowledge, attribution, or disclosure order, which lexical similarity cannot establish [Qin et al.(2019), Chen et al.(2022)]. Humor provides a focused test; transfer to other genres remains to be validated.

Research questions and contributions.

- **RQ1 (symbolic structure).** Does generating from a structured, evidence-linked specification preserve effect conditions better than generating from the same information written as prose?
- **RQ2 (verification).** Does criterion-level functional-equivalence verification with local repair improve accepted adaptations beyond generic verification and beyond simply spending the same number of model calls?
- **RQ3 (diagnostic validity).** Does the verifier detect single-factor violations of the specification while accepting meaning-preserving variants, and which violation types does it miss?

Our contributions are (1) an evidence-linked representation of source effect conditions with operational preservation criteria; (2) an analyze–generate–verify procedure that uses this representation to generate, diagnose, repair, or abstain; (3) a controlled evaluation addressing RQ1–RQ3, together with a check of the analyzer against expert reference labels; and (4) an open implementation with full run logs.

Related work

Adaptation as selective preservation

Style transfer, paraphrase, counterfactual rewriting, translation, and transcreation all change some dimensions of a text while preserving others. Prior work evaluates transfer strength, content retention, naturalness, cultural appropriateness, and communicative function as distinct outcomes [Gero et al.(2019), Mir et al.(2019), Yuan et al.(2026), Wu et al.(2025)], and counterfactual story editing shows that local edits propagate through the event chain [Qin et al.(2019), Chen et al.(2022)]. We operationalize preservation at the level of event relations, attributed beliefs, inference conditions, disclosure order, association paths, and effect cues.

Humor mechanisms and cultural knowledge

The General Theory of Verbal Humor treats language, narrative strategy, situation, logical mechanism, and script opposition as interacting resources [Attardo(2002)]. Recent humor-translation and meme-transcreation systems likewise analyze source mechanisms before recomposition [Su et al.(2025), Zheng et al.(2026), Agrawal and Mamidi(2026), Arampatzis and Arampatzis(2026)]. Benchmarks and Chinese corpora show that humor understanding depends on narrative cues, culture, and audience knowledge, and that offensiveness must be judged separately from funniness [Zhang et al.(2019), Tseng et al.(2020a), Meaney et al.(2021), He et al.(2025), Castro et al.(2026), Cheng and Utsuro(2026), Loakman et al.(2025)]. Narrative and reader models distinguish event succession from state transformation and represent agents' beliefs and belief updates [Peng et al.(2022), Todorov(1990), Van Eecke et al.(2023)]; structure-mapping and rhetorical accounts of synecdoche motivate explicit association paths [Gentner(1983), Todorov(1970), Groupe Mu(1981)].

Neurosymbolic formalization, verification, and evaluation

LLM-assisted story planning and LLM-Modulo systems pair neural proposals with symbolic planning or external critics [Kelly et al.(2023), Kambhampati et al.(2024), Gundawar et al.(2024), Vaucher et al.(2026)]; Gordon's incremental formalization addresses the knowledge-acquisition bottleneck directly [Gordon(2025)]. Self-refinement and evidence-grounded revision are relevant generation strategies [Gao et al.(2023), Madaan et al.(2023)], but model-generated feedback does not reliably correct reasoning errors [Huang et al.(2024), Kamoi et al.(2024)], and extra test-time compute alone can improve outputs [Snell et al.(2024)]—which is why our comparison includes a compute-matched baseline. LLM judges show position, verbosity, and self-preference biases, motivating

blinded judges distinct from the generator and human calibration [van der Lee et al.(2019), Zheng et al.(2023), Chen et al.(2024)].

These studies establish methods for source analysis, symbolic planning, and revision, but leave open which parts of an effect specification help when a text moves to a different context. Our evaluation separates three questions: whether a typed specification improves generation over the same information expressed as prose; whether functional criteria and local repair improve outputs over generic checks or additional generation calls; and whether the verifier can identify controlled violations of individual effect conditions. A frozen, evidence-linked source specification connects these comparisons, allowing generation quality and diagnostic validity to be assessed separately.

Task and representation

Functional equivalence

Let T be a text, u a reader, and c a reading context. Reading changes the reader's state:

$$F(T, u, c): \; r_0 = \langle K_0, B_0, E_0 \rangle \longmapsto r_1 = \langle K_1, B_1, E_1 \rangle, \tag{1}$$

where K is accessible knowledge, B beliefs or interpretations, and E emotional response. Source T_s and target T_t are functionally equivalent for specified readers and contexts when they produce comparable belief revision and emotional change. Equivalence is therefore conditional and graded. An adaptation task is $X = (T_s, C_t, U)$, where the target card C_t fixes setting, roles, institutions, available objects, temporal scale, audience, and content restrictions, and U describes the intended readers. The system returns T_t, a role mapping, a verification record, and a terminal status: accept, repair, or abstain. Abstentions are reported separately, following selective-classification practice [El-Yaniv and Wiener(2010)].

Narrative and rhetorical grounding. Two older accounts make this definition operational. In narratological terms [Todorov(1990), Genette(1980)], a joke has a story level (who does what, and why) and a discourse level (the order in which information is disclosed, where the cue falls, and where the telling stops); attribution records focalization, that is, who knows or believes what; and the H1-to-H2 shift is a transformation in Todorov's sense, from misreading to recognition. Cross-context adaptation replaces the story level while keeping the discourse level and the transformation it produces. Rhetorical accounts of synecdoche describe how such a replacement can remain recognizable [Todorov(1970), Groupe Mu(1981)]: a generalizing synecdoche lifts a particular to its genus, and a particularizing one lowers a genus to a new particular. An adaptation is a round trip of this kind. In the joke from the introduction, the twins and the baptism are lifted to "two comparable units, one of which receives a ritual", which is then lowered to a living room and a study. A paraphrase never leaves the particular, and an adaptation that lifts the source too far loses the cue that made the particular work.

Source specification

The analyzer decomposes the source into the smallest independently truth-evaluable propositions, each linked to a source span. Negation, quantity, time, conditions, and speech scope are preserved. Each proposition is attributed as textual, speech, belief, background, or inference, so a character's claim or belief is never promoted to an unqualified fact—the property that lets deliberate misdirection be represented without contradiction. Entities, events, and values receive stable identifiers, and one proposition may map to several typed ground literals.

Ambiguous material keeps parallel analyses: H1 is the natural reading before a later cue and H2 the revised reading after it; when no second reading is supported, the analyzer says so. The specification records the cue supporting the shift, the motivation that licenses a character's key action, and a short association path when a relation (material, functional, phonological, orthographic, referential, or role-based) is needed to recover H2. Each path edge is labeled explicit, lexical, background, inferred, or uncertain. Most path relations—genus and species, part and whole, an object and its attribute, use, or indicator—are synecdochic in the sense of Section 3.1; phonological and orthographic links are not. In the terms of that section, propositions and motivation describe the story level, attribution records focalization, and the cue and the H1–H2 readings describe the discourse-level transformation; disclosure order and the ending are not fields of the executed specification and are checked by the verifier's general checks (Section 5.4).

Effect signature

The source effect signature has seven separate ordinal dimensions (annotation manual v3.2):

$$\mathbf{e}(T) = \langle A, L, S, N, B, M, G \rangle, \qquad e_j \in \{1, \ldots, 5, \text{NA}, \text{UNKNOWN}\}. \tag{2}$$

Table 1 defines them; full anchors are in Supplemental Material, Section S1. A score describes how strongly a dimension is present, not how funny the text is. Scores are never summed or averaged; NA marks an inapplicable dimension and UNKNOWN insufficient evidence.

Operational preservation criteria

After reparsing a target text, the verifier compares source and target and produces a component-wise record. The frozen source specification defines mechanism-critical hard invariants $\mathcal{H}_s = \{h_1, \ldots, h_m\}$ with $h_i(T_t) \in \{1, 0, \perp\}$ (satisfied, violated, insufficient evidence):

$$H(T_t) = \begin{cases} \text{PASS}, & \forall i: \; h_i(T_t) = 1, \\ \text{FAIL}, & \exists i: \; h_i(T_t) = 0, \\ \text{UNKNOWN}, & \text{otherwise.} \end{cases} \tag{3}$$

Table 1. Source effect-signature dimensions. Each score describes how the effect is produced, not how funny the text is.

Dim.	Construct	What is rated
A	Reading difference	How different H1 and H2 are in reference, intention, or causal explanation.
L	Association support	Whether each step of the listed association path is supported.
S	Cue support	How strongly the text's cues support the intended reading.
N	Narrative coherence	Whether events, conditions, or question–answer relations fit together.
B	Knowledge/belief difference	How much a difference in what characters (or reader) know drives the effect.
M	Motivation authorization	How well the text licenses the key character action.
G	Emotional change	The size of the emotional shift or build-up the text cues.

Let μ map source roles to target roles and $\mathcal{R}_{\text{crit}}$ be the relations the mechanism requires. Relational preservation holds when $\text{Preserved}(\mu(r), T_t) = 1$ for every $r \in \mathcal{R}_{\text{crit}}$. For each mechanism-critical dimension $j \in J_{\text{crit}}$, an admissible source–target category relation $\mathcal{A}_j$ is fixed before testing (exact agreement by default), giving $M_j = \mathbb{I}[(e_{s,j}, e_{t,j}) \in \mathcal{A}_j]$. Operational equivalence is the conjunction

$$\text{FE}_{\text{op}}(T_s, T_t \mid X) = 1 \iff H(T_t) = \text{PASS} \wedge P_{\text{graph}} = 1 \wedge \bigwedge_{j \in J_{\text{crit}}} M_j = 1. \quad (4)$$

An observed violation yields REPAIR (if localizable) or FAIL; missing evidence for a required condition yields UNKNOWN and abstention. FE_{op} describes structural conditions for a comparable interpretation; it does not measure experienced humor.

Analyze–generate–verify architecture

Figure 1 summarizes the three modules. They communicate only through declared artifacts: Analyzer A maps (T_s, U) to a frozen specification S_s; Generator G receives S_s, the target card C_t, and realization settings and returns candidates or an abstention; Verifier V receives a candidate, S_s, and C_t—but not G's plans or self-assessments—builds its own target representation S_t, and returns criterion-level judgments with ACCEPT, REPAIR, or ABSTAIN. G cannot modify S_s; repair edits the target, never the source specification.

Analyzer

The analyzer performs eight conceptual operations (Table 2). In the evaluated implementation all eight are requested in a single structured-output call per

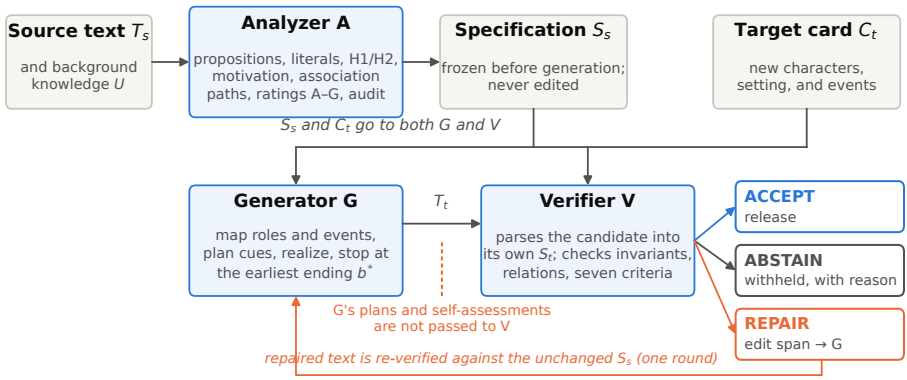


Figure 1. The analyze–generate–verify workflow. Only declared artifacts cross module boundaries: the source specification S_s is frozen before generation, and both the generator and the verifier receive it with the target card C_t. The verifier receives the candidate T_t but not the generator's plans or self-assessments, builds its own representation S_t, and accepts, abstains, or returns a localized repair; a repaired text is verified again from scratch, for at most one round in our experiments.

source, with a frozen prompt, the annotation manual, and a JSON schema; a cross-stage audit and local schema validation follow. Supplemental Material, Section S2, gives the prompts and a stage-by-stage worked example.

Table 2. Analyzer operations and outputs.

#	Operation	Output
1	Segment meaning	Attributed atomic propositions with scope and evidence spans.
2	Entities and events	Stable entity, event, state, and time identifiers; temporal links.
3	Formalize	Typed ground literals; new predicates only when needed.
4	Licensed inferences	Background expectations, knowledge holders, beliefs, supported inferences.
5	Effect mechanism	H1/H2, reinterpretation cue, motivation target, association path.
6	Effect signature	Separate A/L/S/N/B/M/G ratings with evidence.
7	Adaptation constraints	Hard invariants, adaptable surface features, ending constraint.
8	Audit	Coverage, attribution, reference, and evidence checks; local revision flags.

Generator

The generator generalizes the abstraction and grounding maps of our symposium paper [Tsai et al.(2026)]; its prompts are given in Supplemental Material, Section S3. It (1) separates hard invariants, required cues, adaptable features, and prohibited content; (2) proposes a role and event mapping and screens each element against the target card for authority, motivation, temporal scale, availability, and cultural plausibility; (3) plans disclosure order, cue placement, and the intended H1-to-H2 transition; (4) realizes the plan in idiomatic, character-appropriate language; and (5) stops at the earliest boundary b^* at which H2 becomes recoverable, removing any following text that merely explains the mechanism. If no plausible mapping preserves the hard invariants, it abstains and names the unmet condition.

Verifier

The verifier (1) parses the candidate independently into S_t; its prompts are given in Supplemental Material, Section S4; (2) checks attribution, reference, and evidence, marking unsupported fields UNKNOWN; (3) checks each hard invariant and critical relation, recording span, observed relation, and expected source condition; (4) assesses effect criteria separately (interpretation recoverability, motivation, coherence, association support, emotional cues, context plausibility, ending minimality); and (5) returns ACCEPT only if every required criterion passes, REPAIR with a minimal editable span for a localizable violation, and ABSTAIN when essential evidence is missing. A repaired text is re-verified from scratch against the unchanged S_s, up to a fixed repair limit (one round in our experiments).

Experimental setup

Corpus and source selection

The corpus is the Taiwanese Chinese joke collection used in our symposium paper [Tsai et al.(2026)], which is the joke corpus that Tseng et al. released with their retrieval-based humor chatbot [Tseng et al.(2020b)] (`https://github.com/SamTseng/icebreaker`). It contains 4,696 records; 41 automatically flagged near-duplicate groups were manually confirmed and merged into families, and each record keeps its source row and identifier. For the analyzer stage we drew 60 family-distinct records stratified by length (20 short, medium, and long; 51 Chinese, 6 mixed Chinese–English, 3 Taiwanese). Six of these (T001–T006) were used to calibrate prompts and the judge and are excluded from evaluation. From the remaining records we drew 30 sources, 10 per length tercile, with a fixed seed; 8 rely on phonological or orthographic play and are reported separately where relevant. Each source has one target context card, shared by all conditions, that specifies a setting different from the source's (setting and time, roles and relations, available objects and institutions, and assumed readers), a requirement shared by

all cards, and content restrictions. The shared requirement asks the generator to keep the source's mechanism but to replace its characters, setting, events, and associative theme, and, for wordplay, to find a new play of the same kind or abstain. The cards went through three versions (below); all are released with the code and data.

Card and prompt revisions. We revised the cards and prompts twice after inspecting outputs and report all runs (Table 3). The final, far-context run is the primary analysis; earlier near-context runs are secondary.

Near-context cards (v1). An LLM drafted the original cards in an interactive session with access to the source texts. The drafting prompt and model snapshot were not logged. Most cards retained the source setting and allowed, rather than required, a context change. The shared prompt also discouraged adding unsupported facts.

Safety revision (v2). A screen identified eleven cards whose restrictions blocked or altered the source mechanism. Some prohibited content on which the punchline depended; T008 mistakenly prohibited describing the friends' responses, which carry the punchline. We revised these cards to identify inappropriate content and request a substitute theme without prescribing outcomes or describing the mechanism. All seven conditions were rerun on these sources.

Far-context cards (v3). The v2 outputs still largely paraphrased the sources, retaining a median 24%–48% of source character bigrams across conditions. We therefore rewrote all 30 cards to require new settings, characters, and events. For example, field training became an open-plan office, and a director–actress interaction became a homeowner–estate-agent interaction. Claude drafted these cards in an interactive session using the source texts, analyzer outputs, and a written rule released with the cards.

We also updated generation, selection, critique, verification, and repair prompts to recognize new content as preserving a source function rather than requiring literal correspondence (Supplemental Material, Sections S3 and S4). All 210 slots were rerun with the same analyzer, models, budgets, and judge. Because cards and prompts changed together, their effects cannot be separated. Experiment B was unchanged: it perturbs source texts and uses no target cards.

Experiment A: controlled adaptation comparison (RQ1, RQ2)

All conditions receive the same source text and target card and use the same generator model and settings (Table 4). B1 and B2 receive identical information—B2's prose specification is a deterministic rendering of B1's JSON—so their contrast isolates the effect of symbolic structure. B4 and B5 share the same generate–verify–repair loop and differ only in what the verifier checks, isolating functional-equivalence criteria from generic consistency checking. B0-BoN draws four B0 samples and selects one in a fifth call, slightly exceeding B5's maximum of four; it separates the benefit of verification from the benefit of extra test-time compute [Snell et al.(2024)].

Table 3. Runs reported in this paper. Every revision applied one written rule to all cards or all conditions and was followed by a rerun; the far-context run is the primary analysis for RQ1 and RQ2.

Run	Exp.	Size	Target cards	Prompts	Role
Smoke test	A, B	3 sources; 21 slots, 21 perturbations	v1	v1.1, then v1.2	Verifier revision; excluded
Near context, original cards	A	210 slots	v1	v1.2	Secondary
Near context, safety revision	A	77 slots rerun	v2	v1.2	Secondary
Far context	A	210 slots	v3	v2.0	Primary (RQ1, RQ2)
Perturbation test	B	222 perturbations	none	v1.2	Primary (RQ3)

Table 4. Adaptation conditions. All receive the source text and target card. Each pair isolates one factor: B1 against B2 the form of the specification, B4 against B5 what the verifier checks, and B5 against B0-BoN verification against extra calls.

Cond.	Additional input	Post-generation step	Max calls
B0	–	–	1
B0-BoN	–	4 samples, model selects one	5
B1	Structured specification	–	1
B2	Same specification as prose	–	1
B3	Structured specification	Self-critique and one revision	2
B4	Structured specification	Generic consistency check, repair, re-check	4
B5	Structured specification	FE verification, local repair, re-verification	4

Judging. The judge prompt is given in Supplemental Material, Section S5. Each non-abstained final text is assessed by an LLM judge that is a different model from the generator and verifier, runs in a fresh context, receives the source text, target card, and candidate in randomized blind order, and does not receive method labels, generation traces, or the source specification (so it cannot favor spec-conditioned methods by matching the specification's wording). It rates hard-condition status and the seven criteria of Section on 1–5 anchors and returns ACCEPT, REPAIR, or ABSTAIN with cited evidence. The generator is gpt-6-luna and the verifier gpt-6-astra (OpenAI); the judge is claude-opus-5-5 (Anthropic), so the judge comes from a different provider than the models it assesses. All three run with their default sampling: gpt-6-luna and claude-opus-5-5 do not accept a temperature parameter, and the judge's reasoning depth is fixed

at effort `high`. The judge runs without automatic fallback to another model; any refusal is recorded as ABSTAIN, counted as not accepted, and reported separately.

Outcomes and analysis. The primary outcome is judge acceptance over all 30 slots of a condition, so abstentions and failed generations count as not accepted. Candidates that still fail verification after the repair limit (B4, B5) are withheld, as the procedure in Section 4.3 specifies, and count as abstentions; they are not sent to the judge. Secondary outcomes are hard-condition pass rate, each criterion's distribution (reported separately), pipeline status, calls, and tokens. Comparisons are paired by source: we report the mean difference with a source-level bootstrap 95% CI and a two-sided sign-flip permutation test. The pre-specified primary contrasts are B1 vs B2 (RQ1) and B5 vs B0-BoN (RQ2); B5 vs B4 and B5 vs B0 are secondary. Holm correction is applied across all four [Dror et al.(2018)]. A sensitivity analysis repeats these contrasts without the three smoke-test sources (Section 5.4). Results under the near-context cards are reported as a secondary analysis (Section 5.1). With 30 sources, these contrasts can detect only large differences; results are reported as effect estimates with intervals rather than as definitive rankings.

Experiment B: verifier diagnostic validity (RQ3)

For each of the 30 sources we construct eight versions of the source text (Table 5; prompts in Supplemental Material, Section S6). Two are controls that should pass: the unmodified source (C0) and a meaning-preserving paraphrase (C1). Six apply a single targeted violation. A generator model produces each perturbation under an instruction to change only the named factor and to decline when the source has no such factor; declined items are excluded. Labels follow from construction. A domain expert checked a stratified, shuffled sample of 48 items (six per type), without seeing the verifier's decisions, for whether the intended change occurred and whether it was the only change; we report the confirmation rate. A second independent rating was not available in time, so we do not report inter-rater agreement.

The verifier is a different model from the perturbation generator and sees only the frozen specification and the candidate, not the source text, mirroring how it is used on adaptations and preventing a verbatim diff from revealing the edit. We report, per type and overall, the detection rate (UNKNOWN counted as not detected), violation precision/recall/F_1 on decided items, the false-acceptance rate on violations, the false-rejection rate on C0/C1, and the UNKNOWN rate.

Implementation and reproducibility

The pipeline is a resumable Python program using structured JSON-schema outputs. Every call logs the requested and reported model, temperature, token usage, latency, input hash, and retries; every run writes a manifest with prompt and schema hashes. Code, prompts, the source selection, and anonymized run logs

Table 5. Perturbation types and construction labels. Each violation targets one part of the specification, so a type the verifier misses points to the field or check that failed. Layers follow Section 3.1.

Code	Type	Layer	Label	Construction
C0	Original	–	PASS	Unchanged source.
C1	Paraphrase	–	PASS	Wording changes only; events, attribution, order, cues kept.
P1	Early disclosure	Discourse	VIOLATION	Move the reinterpretation cue to the setup.
P2	Motivation removal	Story	VIOLATION	Delete the state or motive licensing the key action.
P3	Order reversal	Discourse	VIOLATION	Swap two events whose order the inference needs.
P4	Attribution swap	Focalization	VIOLATION	Give a key utterance or belief to another character.
P5	Broken association	Association	VIOLATION	Replace a key term on the association path.
P6	Post-punchline explanation	Discourse	VIOLATION	Append one or two sentences explaining the mechanism.

are available from the corresponding author during review and will be deposited in a public repository on publication.

Smoke test and prompt revision. Before the main experiments we ran a smoke test on three of the 30 sources (T008, T009, and T013; 21 adaptation slots and 21 perturbations). It exposed an implementation gap: the executed v3.2 specification has no field for the ending or for disclosure order, so the verifier passed all three post-punchline explanations (P6) and two of the three early disclosures (P1), and noted that the specification did not forbid them. We therefore added two checks that apply regardless of the specification—ending minimality and no early disclosure—to the B5 and perturbation verifier instructions (prompt version 1020-batch-v1.2; Supplemental Material, Sections S4 and S6). On the same 21 perturbations, detection rose from 9/15 to 14/15 violations, with no new false rejections; the one control rejected under both versions was a paraphrase that had in fact changed who performed the key action. Because this revision was informed by three evaluation sources, we report both experiments on all 30 sources and, as a sensitivity analysis, on the 27 sources not used in the smoke test. No smoke-test output enters any reported result.

Results

Analyzer outputs

All 60 analyzer outputs passed local validation against the v3.2 schema (Figure 2). Schema validity establishes that the specifications are well formed and referentially

consistent; it does not establish semantic accuracy, which requires expert reference annotation (Section).

Development check. Two domain experts annotated 17 of the 60 records (T001–T017) under manual v3.2. Before any discussion the experts agreed weakly: Cohen's κ on the categorical fields ranged from 0.07 to 0.40, and the seven scores agreed within one point in 65% of pairs (linear-weighted κ 0.07). Discussion traced most disagreements to one reading of the manual—scoring only the material around the reveal rather than the whole text—and produced a consensus for all 17 records. Table 6 compares the analyzer with this consensus. The specifications used in the experiments were produced through an interactive model interface whose model identifier was not recorded, so we also reran the analyzer with the same manual and prompt through the API (gpt-6-luna). Both runs agree with the consensus about as closely as the two experts agreed with each other before discussion: about two thirds of scores fall within one point, and about a quarter of main-mechanism labels match. Both also score below the consensus on most dimensions, most clearly on emotional contrast (G; by 1.5 and 2.1 points on average) and motivation (M; by 1.1 and 1.2 points). These figures are preliminary. The consensus rests on two experts and 17 records; the same records were used to revise the annotation manual (v3.5) for the remaining 43; and six of them (T001–T006) were used for prompt calibration. Excluding those six, the experiment specifications agree within one point on 63% of scores (weighted κ 0.05). Rerunning the analyzer under the revised manual (v3.5) on the same records raised agreement to 84% within one point (weighted κ 0.42) and narrowed the gap on G to 1.1 points; because the revised manual takes its examples from these records, this is not a held-out estimate.

Held-out check. For the remaining 43 records (T018–T060), none of which was used to revise the manual, reference labels were produced differently. An LLM drafted them under manual v3.4 and reassigned ambiguity type and reading difference (A) under v3.5, and it listed 43 open questions, one per record, with a proposed resolution for each. The two experts then reviewed every record, adopted all 43 proposed resolutions, and thereby changed five records: the ambiguity type of four (two of which the draft had left undecided), the main mechanism of one, the A and L ratings of one, and the recoverability of one. These labels are therefore expert-reviewed rather than independently annotated: we cannot report agreement between the experts on them, and because the experts accepted every draft resolution, the labels may lean toward the drafting model's judgments. Against these labels, the specifications used in the experiments agree within one point on 77% of scores (linear-weighted κ 0.34) and score below the reference on average (-0.60), most on motivation (M, -1.7), emotional contrast (G, -1.2), and knowledge gap (B, -1.1); they match the reference ambiguity type for 49% and the main mechanism for 56% of records (Table 6). Rerunning the analyzer under manual v3.5 raises agreement to 85% within one point (κ 0.45) and narrows the gap on M, but leaves G 1.1 points low and does not improve the categorical

fields; because the drafts and the v3.5 analyzer read the same manual, even this gain may be overstated. As an exploratory check that was not pre-specified, 24 of the 30 experimental sources fall in this set. Under the far-context cards, on the eleven whose specification named a different main mechanism from the reference, the conditions that received the specification (B1–B5) were accepted 22.7 points less often than B0 and B0-BoN, whereas on the other thirteen they were accepted 3.1 points more often (difference 25.8 points, permutation $p = 0.06$); under the near-context cards the difference was 10.2 points ($p = 0.58$). The size of score errors was unrelated to acceptance (Spearman $\rho = 0.21$, $p = 0.33$).

Table 6. Expert checks of the analyzer. Development records (T001–T017): agreement with the two experts' consensus, with the experts' own agreement before discussion for reference; these records were also used to revise the manual. Held-out records (T018–T060): agreement with LLM-drafted reference labels that the two experts reviewed, changing five records. On both sets the specifications used in the experiments score below the reference, most clearly on emotional contrast and motivation. Mean differences are analyzer minus reference.

	Development (17 records)			Held out (43 records)	
	Experiment specs	gpt-6-luna rerun	Expert A vs. B, before discussion	Experiment specs	v3.5 rerun
Scores within one point	70%	66%	65%	77%	85%
Linear-weighted κ, scores	0.20	0.26	0.07	0.34	0.45
Mean difference	−0.55	−0.77	–	−0.60	−0.47
G (emotional contrast)	−1.53	−2.12	–	−1.23	−1.09
M (motivation)	−1.09	−1.18	–	−1.69	−0.45
Ambiguity type, agreement	53%	59%	24%	49%	53%
Main mechanism, agreement	24%	24%	18%	56%	44%

Experiment A

Table 7 reports each condition's outcomes under the far-context cards and Table 8 the paired contrasts. The outputs replaced the source's content: per condition, the median share of the source's character bigrams that reappeared in the output was 2%–7%, against 24%–48% under the near-context cards (Figure 3). Judge acceptance ranged from 7% (B2) to 30% (B0 and B5), and neither primary contrast differed from zero. B1−B2 was +10.0 points (95% CI −3 to 23; Holm $p = 1.00$), the opposite direction from the near-context run; B5−B0-BoN was +6.7 points (−13 to 27). Of the secondary contrasts, B5−B4 was +16.7 points (−3 to 37; Holm $p = 0.91$) and B5−B0 was 0.0. The four conditions that received the specification without functional-equivalence verification (B1–B4) were all

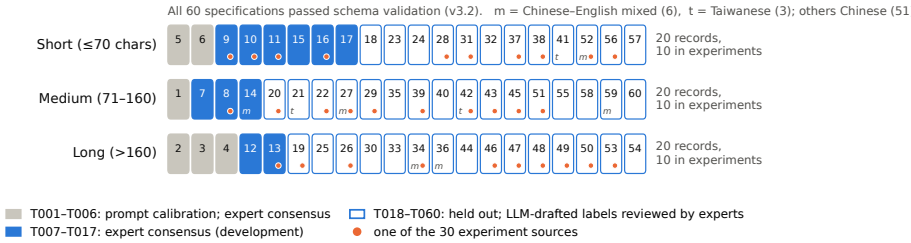


Figure 2. The 60 analyzer records by length tercile (numbers are record IDs). Shading gives each record's role in expert annotation: T001–T006 calibrated the prompts, T001–T017 have the two experts' consensus and served as development data for the manual (Section 6.1), and T018–T060 are held out, with LLM-drafted reference labels that the two experts reviewed. Dots mark the 30 sources used in Experiments A and B, ten per tercile. All 60 specifications are schema-valid; whether their content is correct is not established here (Section).

accepted less often than B0 (7%–23%, against 30%). In an exploratory paired test, which was not pre-specified, their mean acceptance was 15.0 points below B0 (bootstrap 95% CI −30.8 to 0.0; sign-flip permutation $p = 0.09$). Of the four, only B2 differed from B0 at the uncorrected 0.05 level (−23.3 points, −40.0 to −6.7, $p = 0.04$), and this does not survive a Holm correction across the five exploratory tests. Under the near-context cards, the same pooled difference was +0.8 points (−15.0 to 16.7). B5 withheld 10 of 30 candidates, all unresolved repairs (Figure 4), and used 3.6 calls and 9.2k tokens per slot, against 5.0 calls and 8.2k tokens for B0-BoN, so it was matched on calls but used more tokens. On hard-condition pass, which counts withheld slots as failures, B5 passed 57% against 67% for B0-BoN (−10.0 points, −33 to 13). Excluding the three smoke-test sources gives similar contrasts (B1−B2 +14.8 points, bootstrap CI 4 to 30 but permutation $p = 0.13$ and Holm $p = 0.50$; B5−B0-BoN +3.7, −19 to 26). Of the 199 judged outputs, the judge accepted 43, returned REPAIR for 154 and ABSTAIN for 2, and refused none.

Among outputs that were released and judged, B5's were rated 4 or 5 most often on six of the seven criteria, all except emotional cues (criterion shares are in the released summary). The largest gaps were ending minimality, where 70% of B5's judged outputs scored 4 or 5 against 13%–50% for the other conditions, and interpretation recoverability, 70% against 30%–53%. Because B5 withholds the candidates its verifier cannot pass, these shares are conditional on release and are not a causal estimate of verification on output quality.

Near-context cards (secondary analysis). Under the near-context cards (v2), where most outputs retold the source in new words, acceptance was higher (33%–50%) and neither primary contrast differed from zero (B1−B2 −16.7 points, −40 to 7; B5−B0-BoN −6.7, −30 to 17). B5 withheld 5 of 30 candidates, and 88%

Table 7. Experiment A by condition ($n = 30$ sources each; primary analysis, far-context cards v3). Acceptance and hard-condition pass use all slots as denominator; Abstain counts withheld slots, including unresolved repairs (B3: 1 abstention; B5: 10 unresolved repairs). Calls and tokens are means per slot. B5 matches the best baseline while withholding a third of its candidates, and the conditions that use the specification without functional verification (B1–B4) are accepted less often than B0, though not significantly (Section 6.2).

Cond.	Judge accept	Hard pass	Abstain	Mean calls	Mean tokens
B0	9/30 (30%)	19/30 (63%)	0	1.0	1,823
B0-BoN	7/30 (23%)	20/30 (67%)	0	5.0	8,166
B1	5/30 (17%)	23/30 (77%)	0	1.0	3,233
B2	2/30 (7%)	18/30 (60%)	0	1.0	3,324
B3	7/30 (23%)	18/30 (60%)	1	2.0	5,421
B4	4/30 (13%)	16/30 (53%)	0	2.3	5,509
B5	9/30 (30%)	17/30 (57%)	10	3.6	9,161

Table 8. Paired contrasts on judge acceptance, in percentage points ($n = 30$ sources; bootstrap 95% CI; Holm-adjusted p across the four contrasts; far-context cards v3). No contrast differs from zero, and the structure contrast (B1−B2) reverses sign relative to the near-context run.

Contrast	Difference	95% CI	p_{Holm}
B1 − B2 (primary)	10.0	[−3.3, 23.3]	1.00
B5 − B0-BoN (primary)	6.7	[−13.3, 26.7]	1.00
B5 − B4	16.7	[−3.3, 36.7]	0.91
B5 − B0	0.0	[−16.7, 16.7]	1.00

of its released outputs scored 4 or 5 on ending minimality, against 53%–67% for the other conditions. Revising the eleven cards whose restrictions had ruled out or altered the mechanism (v1 to v2) raised acceptance on those sources from 9 to 23 of 77 slots, and no output on them violated a content restriction. Under the original v1 cards, acceptance ranged from 30% to 37% (B1−B2 −6.7 points, −23 to 10; B5−B0-BoN 0.0, −20 to 20). Full tables for both near-context runs are in the released summaries.

Experiment B

Of the 240 planned perturbations, 222 were generated; 18 were declined as not applicable, mostly broken associations (8) and motivation removals (5) in sources

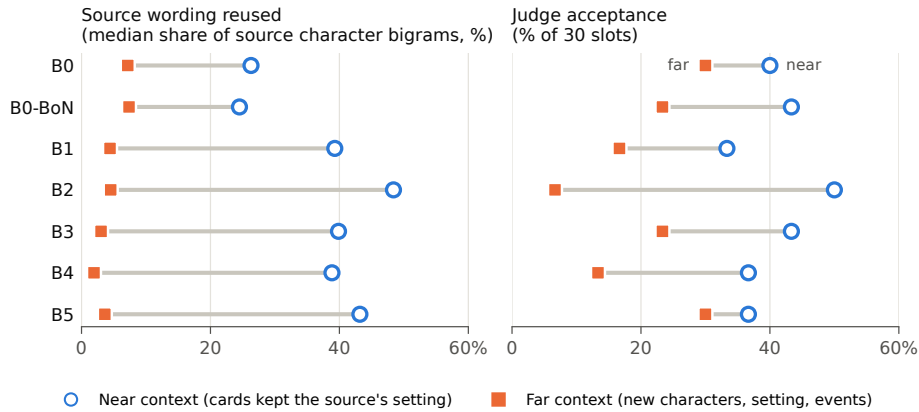


Figure 3. Near- and far-context runs by condition. Left: median share of the source's character bigrams that reappear in the output; right: judge acceptance over 30 slots. Requiring new characters, settings, and events turns near-paraphrases into new texts and lowers acceptance in every condition; B5 and B0 lose the least.

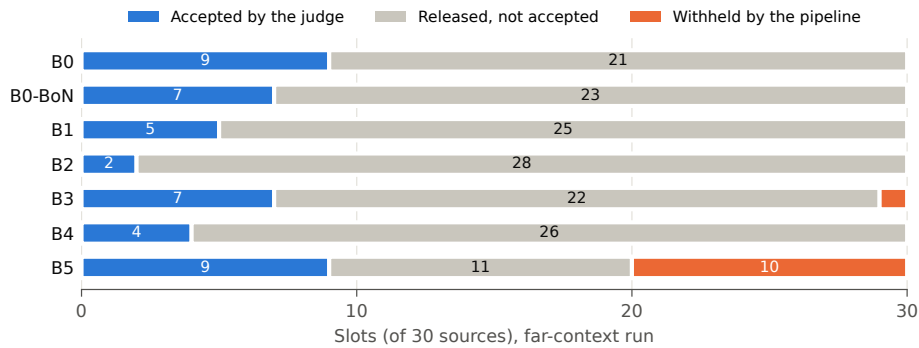


Figure 4. Outcome of every far-context slot by condition. B5 and B0 have the same number of accepted adaptations (9), but B5 withholds 10 candidates its verifier could not pass, so it releases 11 unaccepted outputs against 21 for B0.

that have no association path or no motive to remove. Overall, the verifier detected 154 of 162 violations (95.1%; precision 0.98, F_1 0.97), rejected 3 of 60 controls (5.0%), and never returned UNKNOWN. Table 9 gives verifier outcomes by type. The expert rated 47 of the 48 sampled items (one paraphrase control was left unrated) and confirmed 38 (81%) as valid single-factor changes. All controls (11 of 11) and all early disclosures, order reversals, and added explanations (18 of 18) were confirmed. The exceptions were motivation removals (2 of 6 confirmed; in three the intended change did not take effect, and in one it was not the only

change), broken associations (3 of 6), and attribution swaps (4 of 6). On the 38 confirmed items, the verifier's decision matched the construction label on 37; the exception was the original of T027, discussed below. Of the nine unconfirmed items, all labeled violations by construction, the verifier flagged seven, including all three broken associations. Grouped by the layers of Section 3.1 (Figure 5), discourse-level violations (early disclosure, order reversal, and added explanation) were detected in 83 of 86 cases (96.5%), and all 18 sampled items were confirmed as valid; story-level and focalization violations (motivation removal and attribution swap) were detected in 49 of 54 (90.7%), and 6 of 12 sampled items were confirmed; broken associations were detected in 22 of 22, and 3 of 6 sampled items were confirmed. Excluding the three smoke-test sources, detection is 139 of 146 (95.2%) and false rejection 3 of 54 (5.6%), so the smoke-test revision has little influence on these estimates.

Table 9. Verifier outcomes by perturbation type. Detection counts UNKNOWN as not detected. Detection is complete for early disclosure, broken association, and added explanation and lowest for motivation removal; the expert check found most invalid perturbations among motivation removals and broken associations (Section 6.3).

Type	n	Detection / false rej.	False accept	UNKNOWN
C0 Original	30	1/30 (3.3%)	–	0
C1 Paraphrase	30	2/30 (6.7%)	–	0
P1 Early disclosure	29	29/29 (100%)	0/29	0
P2 Motivation removal	25	22/25 (88.0%)	3/25 (12.0%)	0
P3 Order reversal	27	24/27 (88.9%)	3/27 (11.1%)	0
P4 Attribution swap	29	27/29 (93.1%)	2/29 (6.9%)	0
P5 Broken association	22	22/22 (100%)	0/22	0
P6 Added explanation	30	30/30 (100%)	0/30	0
All	222	154/162 (95.1%); false rej. 3/60 (5.0%)	8/162 (4.9%)	0

Worked example

Source T029 (paraphrased) is the joke from the introduction: a scientist's wife has twins, and when a priest invites both to be baptized, the scientist answers that one is enough, because the other is the control group. The analyzer separates the narrated event (`HasTwins(ScientistFamily)`, `Requests(Baptism(Children))`) from the scientist's reply (`Says(Scientist, BaptizeOne, OtherControl)`),

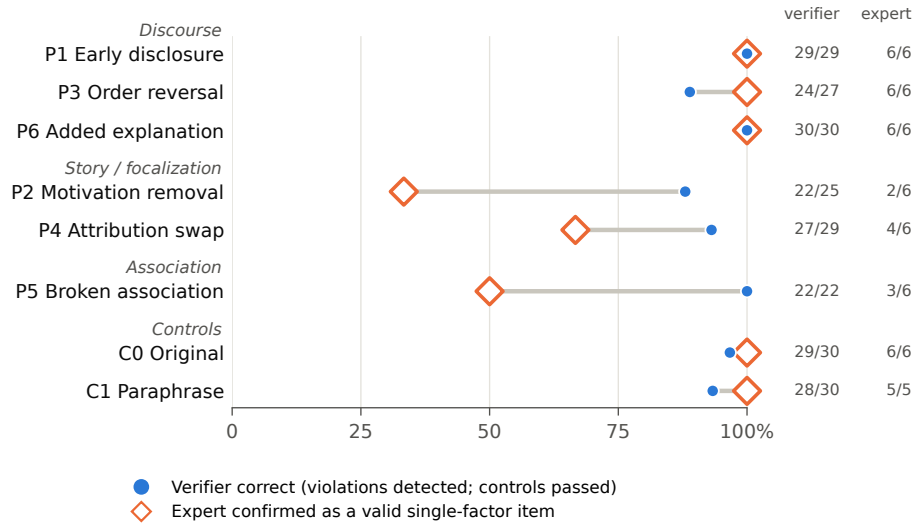


Figure 5. Verifier correctness and expert confirmation by perturbation type, grouped by narrative layer. Counts give verifier-correct items over all items and expert-confirmed items over rated items. The two agree on the discourse layer; on the story, focalization, and association layers the verifier flags violations that the expert often did not confirm as valid, so detection there partly reflects invalid perturbations.

records H1 (an ordinary invitation to baptize both children) and H2 (the scientist recasts the ritual as an experiment with a control group), and links the twins to treatment and control groups as an explicit association. Under the near-context card, B5 retold the source almost word for word. The far-context card moves the joke to a product manager whose elders propose a customary ritual, and B5's first draft read:

> 家裡剛買了新房，長輩提議照習俗辦入厝儀式。產品經理說：「當然辦！客廳照習俗走，書房先當對照組；住一個月我拉張數據報表，看看哪邊先傳出好消息，再貼到家族群組。」

(*The family has just bought a flat, and the elders propose the customary housewarming ritual. The product manager: "Of course! The living room gets the ritual, and the study is the control group; after a month I'll pull a report, see which side brings good news first, and post it to the family chat."*) The verifier returned VIOLATION on ending minimality: once "the study is the control group" completes the reinterpretation, the rest restates it. Repair deleted the final clause, the verifier passed the shortened text, and the blinded judge accepted it with an ending-minimality score of 5. The judge also named the adaptation's limit: its associative core, an experimental control, is the same as the source's, so the joke is renewed in characters, setting, and events but not in its central idea. In the

terms of Section 3.1, the round trip rose only as far as "an experimental control" before descending again.

Discussion

What the symbolic layer contributes. Neither structured specifications (RQ1) nor criterion-level verification (RQ2) significantly increased acceptance in the matched comparisons. With 30 sources, the intervals allow substantial differences in either direction; absence of a detectable effect is not equivalence. Specification quality may be one constraint. The experiment specifications underrated motivation and emotional contrast against the 43 expert-reviewed reference records. Their mechanism errors were also associated with a larger acceptance deficit for specification-conditioned methods, but only in an exploratory comparison ($p = 0.06$; Section 6.1).

Verification's clearest contribution was selective release. B5 matched B0's observed acceptance count while withholding a third of its candidates; its released outputs received the highest ratings on six of seven criteria. This is selectivity, not evidence of better generation. B1–B4 all fell below B0 under far-context cards, but the pooled exploratory difference was not significant ($p = 0.09$) and did not appear under near-context cards. Source specifications may anchor generation to source details: near-context outputs retained 39%–48% of source bigrams with specifications, against 24%–26% without them. Whether targeted repair can overcome such anchoring remains open.

The target card shapes the task. Near-context instructions allowed paraphrases; revised cards and prompts required new content and lowered acceptance. Because both changed together, their separate effects cannot be identified. A useful card must require a context change without prescribing the plot or prohibiting the mechanism. Restrictions should identify inappropriate content and request a substitute theme. The account in Section 3.1 offers an interpretation: adaptation requires abstracting from source particulars and grounding the shared relation anew. T029 shows that this can renew characters and events without renewing the central associative idea.

Diagnosis. Construction labels and expert checks must be read together. Discourse-level edits were detected in 83 of 86 cases, and all 18 sampled edits were confirmed. Motivation removals, attribution swaps, and broken associations were less reliably constructed: only 9 of 18 sampled items were confirmed. Some deleted motives remained inferable, and replacing an association term could leave another interpretive route intact. The verifier flagged all three sampled associations that the expert considered unbroken. Thus high detection on construction labels does not alone establish diagnostic validity.

All three false rejections arose from the general ending check added after the smoke test. Sources T027 and T034 already contain explanation after the reveal, so a source-independent minimality rule can reject a meaning-preserving text.

Future checks should distinguish preservation of the source ending from a separate preference for compression.

Representation and realization. First-order literals capture entities, events, negation, and some preconditions [Gordon(2025)]; attribution, association paths, and ordinal dimensions represent additional effect conditions [Attardo(2002), Van Eecke et al.(2023)]. These outputs support auditing, not claims about the model's internal processing [Bender and Koller(2020), Turpin et al.(2023)].

A detailed specification also risks leaking analytic language into the adaptation. The earliest-sufficient-boundary rule distinguishes a cue enabling resolution from an explanation of that resolution [Suls(1972)]; P6 tests its enforcement, while the false rejections show why it needs source-sensitive application.

Beyond humor. Reusing propositions, attribution, and verifier interfaces for fable simplification, plain-language public information, or educational adaptation is a hypothesis; each task needs its own effect signature and validation with intended readers [Alva-Manchego et al.(2020), Ayre et al.(2024)].

Opportunities and Challenges of LLM-Based Formalization

Gordon's formalization of *Venus and Adonis* illustrates the acquisition opportunity and its limits [Gordon(2025)]. The model introduced 939 predicates and 1,292 constants across 1,817 simple sentences, but 71.7% of predicates and 65.6% of constants appeared only once. The study also reports near-duplicate predicates, pronoun-driven referent drift, lost negation, and a constant assigned to two referents.

These problems motivate stable identifiers, evidence spans, polarity and attribution checks, and multiple literals per proposition. They do not establish that our implementation resolves those problems. The present experiments test generation and diagnosis; predicate reuse and cross-text vocabulary stability require separate evaluation. Our symposium pilot likewise found mapping logs useful for inspection but technical language disruptive to immediacy [Tsai et al.(2026)].

Defining predicates is only part of domain axiomatization [Gordon(2025)]. Schema validity establishes well-formed output, not correct meaning; analyzer agreement and verifier validity must therefore be evaluated separately.

Limitations

Reference labels. The 17 development records informed the manual. For the 43 held-out records, the experts reviewed the AI-drafted annotations and revised five records before finalizing the reference labels. This expert review differs from independent double annotation, and shared manual assumptions may favor revised-manual runs. Independent double annotation remains necessary.

Judging and perturbations. Experiment A uses one uncalibrated LLM judge, so provider separation and blinding do not eliminate style bias [Zheng et al.(2023), Chen et al.(2024)]. Experiment B uses construction labels; 9 of 47 expert-rated items were not confirmed as valid single-factor changes.

Post-output revisions and target cards. Verifier instructions changed after a smoke test on three evaluation sources. The 27-source sensitivity analysis checks robustness but does not create a fully held-out test. Cards and prompts were then revised twice after inspecting Experiment A outputs. Earlier runs are reported, but no card set was preregistered, and card effects cannot be separated from prompt effects. Shared cards control inputs across methods but do not establish neutral task framing. In v2, the explicit mechanism-preservation requirement appeared only on eleven revised cards, limiting cross-source comparison of hard-condition pass rates.

Scope. Thirty sources yield wide intervals, not fine-grained rankings. The narrative-layer and synecdochic interpretations were not independently annotated. Structural judgments do not measure experienced humor, and other genres require task-specific validation. The corpus's age and cultural coverage limit generality. Because its repository states no license, we do not redistribute it; released records identify source rows, and paper examples are paraphrased or newly written.

Ethics

The corpus includes jokes about death, sex, identity, and insults. Analysis of such mechanisms is kept separate from generation for release, and target cards name content that is inappropriate in the target setting and ask for a substitute theme that keeps the mechanism; the system abstains only when it cannot produce an adaptation that satisfies the card.

Conclusion

We defined operational conditions for cross-context functional equivalence and implemented an analyze–generate–verify pipeline with repair and abstention. Requiring new content reduced paraphrasing, but neither structured specifications nor verification significantly improved judge acceptance over matched baselines. Verification's supported role is narrower: it selected which outputs to release and detected most construction-labeled violations. Moderate analyzer agreement and imperfect perturbations limit stronger claims. Priorities are independent source annotation, expert-calibrated judging, source-sensitive verification, and reader studies testing whether the structural criteria predict interpretation and emotion.

Acknowledgements

This research was supported by the National Science and Technology Council, Taiwan (grant numbers NSTC 114-2634-F-002-002-, NSTC 112-2410-H-141-008-MY3, and NSTC 114-2420-H-002-010-).

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Supplemental material

The supplemental file contains effect-signature anchors (S1), analyzer prompts with a worked example (S2), generation prompts (S3), verification and repair prompts (S4), the blind judge prompt (S5), and perturbation prompts (S6). Code, prompts, target cards, the annotation manual, the source selection, and run logs are available from the corresponding author during review and will be deposited in a public repository on publication; the corpus itself is not redistributed.

References

[Agrawal and Mamidi(2026)] Agrawal, J., & Mamidi, R. (2026). j10official at SemEval-2026 Task 1: Neurosymbolic humor generation via GTVH-guided LLM decomposition. In *Proceedings of the 20th International Workshop on Semantic Evaluation* (pp. 3015–3026). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.semeval-1.379

[Alva-Manchego et al.(2020)] Alva-Manchego, F., Martin, L., Bordes, A., Scarton, C., Sagot, B., & Specia, L. (2020). ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4668–4679). https://doi.org/10.18653/v1/2020.acl-main.424

[Arampatzis and Arampatzis(2026)] Arampatzis, G., & Arampatzis, A. (2026). DUTH at SemEval-2026 Task 1: Prompt-based zero-shot large language models for constrained multilingual humor generation. In *Proceedings of the 20th International Workshop on Semantic Evaluation* (pp. 579–584). https://doi.org/10.18653/v1/2026.semeval-1.83

[Attardo(2002)] Attardo, S. (2002). Translation and humour: An approach based on the General Theory of Verbal Humour (GTVH). *The Translator*. https://doi.org/10.1080/13556509.2002.10799131

[Ayre et al.(2024)] Ayre, J., Muscat, D. M., Mac, O., Beller, E., McCaffery, K. J., & Dodd, R. H. (2024). Online plain language tool and health information

quality: A randomized clinical trial. *JAMA Network Open, 7*(10), e2437955. https://doi.org/10.1001/jamanetworkopen.2024.37955

[Bender and Koller(2020)] Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. https://aclanthology.org/2020.acl-main.463/

[Castro et al.(2026)] Castro, S., Chiruzzo, L., Góngora, S., Deng, N., Rahili, S., Sastre, I., Rosá, A., Amoroso, V., Rey, G., Moncecchi, G., Meaney, J. A., Prada, J. J., & Mihalcea, R. (2026). SemEval-2026 Task 1: MWAHAHA, models write automatic humor and humans annotate. In *Proceedings of the 20th International Workshop on Semantic Evaluation* (pp. 3797–3822). https://doi.org/10.18653/v1/2026.semeval-1.454

[Chen et al.(2022)] Chen, J., Gan, C., Cheng, S., Zhou, H., Xiao, Y., & Li, L. (2022). Unsupervised editing for counterfactual stories. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(10), 10473–10481. https://doi.org/10.1609/aaai.v36i10.21290

[Chen et al.(2024)] Chen, G. H., Chen, S., Liu, Z., Jiang, F., & Wang, B. (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* (pp. 8301–8327). https://doi.org/10.18653/v1/2024.emnlp-main.474

[Cheng and Utsuro(2026)] Cheng, J., & Utsuro, T. (2026). Can LLMs understand punchlines? LLMs' narrative understanding evaluation with short-shorts. In *Proceedings of the 15th Language Resources and Evaluation Conference* (pp. 2024–2034). https://doi.org/10.63317/4n2p36736i24

[Dror et al.(2018)] Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 1383–1392). https://doi.org/10.18653/v1/P18-1128

[El-Yaniv and Wiener(2010)] El-Yaniv, R., & Wiener, Y. (2010). On the foundations of noise-free selective classification. *Journal of Machine Learning Research, 11*. https://www.jmlr.org/papers/v11/el-yaniv10a.html

[Gao et al.(2023)] Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A. T., Fan, Y., Zhao, V., Lao, N., Lee, H., Juan, D.-C., & Guu, K. (2023). RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 16477–16508). https://doi.org/10.18653/v1/2023.acl-long.910

[Genette(1980)] Genette, G. (1980). *Narrative discourse: An essay in method* (J. E. Lewin, Trans.). Cornell University Press.

[Gentner(1983)] Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*(2), 155–170. https://doi.org/10.1207/s15516709cog0702_3

[Gero et al.(2019)] Gero, K., Kedzie, C., Reeve, J., & Chilton, L. (2019). Low level linguistic controls for style transfer and content preservation. In *Proceedings of the 12th International Conference on Natural Language Generation* (pp. 208–218). https://doi.org/10.18653/v1/W19-8628

[Gordon(2025)] Gordon, A. S. (2025). Automated formalization of William Shakespeare's *Venus and Adonis* using large language models. *Advances in Cognitive Systems, 12*, 259–274. https://openreview.net/pdf?id=xczKwRv1t0

[Groupe Mu(1981)] Groupe Mu. (1981). *A general rhetoric*. Johns Hopkins University Press.

[Gundawar et al.(2024)] Gundawar, A., Valmeekam, K., Verma, M., & Kambhampati, S. (2024). Robust planning with compound LLM architectures: An LLM-Modulo approach. *arXiv:2411.14484*. https://arxiv.org/abs/2411.14484

[He et al.(2025)] He, R., He, Y., Bai, L., Liu, J., Sun, Z., Tang, Z., Wang, H., Xia, H., Mihalcea, R., & Deng, N. (2025). Chumor 2.0: Towards better benchmarking Chinese humor understanding from (Ruo Zhi Ba). In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 21799–21818). https://doi.org/10.18653/v1/2025.findings-acl.1122

[Huang et al.(2024)] Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2024). Large language models cannot self-correct reasoning yet. In *Proceedings of the Twelfth International Conference on Learning Representations*. https://proceedings.iclr.cc/paper_files/paper/2024/hash/8b4add8b0aa8749d80a34ca5d941c355-Abstract-Conference.html

[Kambhampati et al.(2024)] Kambhampati, S., Valmeekam, K., Guan, L., Verma, M., Stechly, K., Bhambri, S., Saldyt, L., & Murthy, A. (2024). LLMs can't plan, but can help planning in LLM-Modulo frameworks. *arXiv:2402.01817*. https://arxiv.org/abs/2402.01817

[Kamoi et al.(2024)] Kamoi, R., Zhang, Y., Zhang, N., Han, J., & Zhang, R. (2024). When can LLMs actually correct their own mistakes? A critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics, 12*, 1417–1440. https://doi.org/10.1162/tacl_a_00713

[Kelly et al.(2023)] Kelly, J., Calderwood, A., Wardrip-Fruin, N., & Mateas, M. (2023). There and back again: Extracting formal domains for controllable neurosymbolic story authoring. In *Proceedings of the 19th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (pp. 64–74). https://doi.org/10.1609/aiide.v19i1.27502

[Li et al.(2026)] Li, J., Guo, X., Wu, Y., Lee, R. K.-W., Li, H., & Xie, Y. (2026). Lost in stories: Consistency bugs in long story generation by LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026* (pp. 8400–8428). https://doi.org/10.18653/v1/2026.findings-acl.410

[Loakman et al.(2025)] Loakman, T., Thorne, W., & Lin, C. (2025). Who's laughing now? An overview of computational humour generation and explanation. In *Proceedings of the 18th International Conference on Natural Language Generation* (pp. 780–794). https://aclanthology.org/2025.inlg-main.45/

[Madaan et al.(2023)] Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (2023). Self-Refine: Iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. https://dl.acm.org/doi/10.5555/3666122.3668141

[Meaney et al.(2021)] Meaney, J. A., Wilson, S. R., Chiruzzo, L., Lopez, A., & Magdy, W. (2021). SemEval-2021 Task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation* (pp. 105–119). https://doi.org/10.18653/v1/2021.semeval-1.9

[Mir et al.(2019)] Mir, R., Felbo, B., Obradovich, N., & Rahwan, I. (2019). Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 495–504). https://doi.org/10.18653/v1/N19-1049

[Peng et al.(2022)] Peng, X., Xie, K., Alabdulkarim, A., Kayam, H., Dani, S., & Riedl, M. (2022). Guiding neural story generation with reader models. In *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 7087–7111). https://doi.org/10.18653/v1/2022.findings-emnlp.526

[Qin et al.(2019)] Qin, L., Bosselut, A., Holtzman, A., Bhagavatula, C., Clark, E., & Choi, Y. (2019). Counterfactual story reasoning and generation. In *Proceedings of EMNLP-IJCNLP 2019* (pp. 5043–5053). https://doi.org/10.18653/v1/D19-1509

[Snell et al.(2024)] Snell, C., Lee, J., Xu, K., & Kumar, A. (2024). Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv:2408.03314*. https://arxiv.org/abs/2408.03314

[Su et al.(2025)] Su, Y., Zhu, Y., Chen, Y., Benavides-Prado, D., & Witbrock, M. (2025). Psychology-driven enhancement of humour translation. *arXiv:2507.09259*. https://arxiv.org/abs/2507.09259

[Suls(1972)] Suls, J. M. (1972). A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. In J. H. Goldstein & P. E. McGhee (Eds.), *The psychology of humor* (pp. 81–100). Academic Press. https://doi.org/10.1016/B978-0-12-288950-9.50010-9

[Todorov(1970)] Todorov, T. (1970). Synecdoques. *Communications, 16*, 26–35. https://doi.org/10.3406/comm.1970.1227

[Todorov(1990)] Todorov, T. (1990). The two principles of narrative. In C. Porter (Trans.), *Genres in discourse*. Cambridge University Press.

[Tsai et al.(2026)] Tsai, C.-E., Chen, F., & Hsu, J. Y.-J. (2026). Context-adaptive humor rewriting: A first-order logic framework using large language models. *Proceedings of the AAAI Spring Symposium Series, 8*(1), 592–596. https://doi.org/10.1609/aaaiss.v8i1.42592

[Turpin et al.(2023)] Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. https://dl.acm.org/doi/10.5555/3666122.3669397

[Tseng et al.(2020a)] Tseng, Y.-H., Wu, W.-S., Chang, C.-Y., Chen, H.-C., & Hsu, W.-L. (2020a). Development and validation of a corpus for machine humor comprehension. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 1346–1352). https://aclanthology.org/2020.lrec-1.168/

[Tseng et al.(2020b)] Tseng, Y.-H., Hsu, W.-L., Wu, W.-S., Gu, Y.-C., & Chen, H.-C. (2020b). Implementation and evaluation of a retrieval-based Chinese humor chatbot. *Journal of Library and Information Studies, 18*(2), 73–101. https://doi.org/10.6182/jlis.202012_18(2).073

[van der Lee et al.(2019)] van der Lee, C., Gatt, A., van Miltenburg, E., Wubben, S., & Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation* (pp. 355–368). https://doi.org/10.18653/v1/W19-8643

[Van Eecke et al.(2023)] Van Eecke, P., Verheyen, L., Willaert, T., & Beuls, K. (2023). The Candide model: How narratives emerge where observations meet beliefs. In *Proceedings of the 5th Workshop on Narrative Understanding* (pp. 48–57). https://doi.org/10.18653/v1/2023.wnu-1.7

[Vaucher et al.(2026)] Vaucher, M., Silveira, S., Góngora, S., & Chiruzzo, L. (2026). IVIE: A neuro-symbolic approach to incremental and validated generation of interactive fiction worlds. *arXiv:2606.13348*. https://arxiv.org/abs/2606.13348

[Wu et al.(2025)] Wu, H., Silva, B., Cabarrão, V., & Moniz, H. (2025). Cultural transcreation in Asian languages with prompt-based LLMs. In *Proceedings of Machine Translation Summit XX* (pp. 42–51). https://aclanthology.org/2025.mtsummit-2.5/

[Yuan et al.(2026)] Yuan, Z., Ye, Y., Feng, X., Li, B., Hong, Q., Lu, Y., Tu, D., & Qin, B. (2026). Culture-aware machine translation in large language models: Benchmarking and investigation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 29636–29661). https://doi.org/10.18653/v1/2026.acl-long.1367

[Zhang et al.(2019)] Zhang, D., Zhang, H., Liu, X., Lin, H., & Xia, F. (2019). Telling the whole story: A manually annotated Chinese dataset for the analysis of humor in jokes. In *Proceedings of EMNLP-IJCNLP 2019* (pp. 6402–6407). https://doi.org/10.18653/v1/D19-1673

[Zheng et al.(2023)] 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., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv:2306.05685*. https://arxiv.org/abs/2306.05685

[Zheng et al.(2026)] Zheng, J., Liu, Y., Peng, Z., Hu, T., Zhao, Y., Zheng, X., & He, X. (2026). TransMeme: A multi-agent framework for cross-cultural meme transcreation. *arXiv:2608.27127*. https://arxiv.org/abs/2608.27127