Context-Adaptive Humor Rewriting: A First-Order Logic Framework Using Large Language Models

Cheng-En Tsai[1], **Fanfan Chen**[2], **Jane Yung-jen Hsu**[3]
[1]Graduate Institute of Networking and Multimedia, National Taiwan University
[2]Department of Creative Technologies and Product Design, National Taipei University of Business
[3]Department of Artificial Intelligence, Chang Gung University
d11944008@csie.ntu.edu.tw, ffchen@ntub.edu.tw, yjhsu@cgu.edu.tw

Abstract

Humor is an important conversational skill that conveys approachability. When addressing different audiences, we often need to adjust and adapt jokes to lower the comprehension barrier while also preventing the original punchline from becoming offensive in a new context. To this end, we propose a humor-rewriting-agent framework. The system converts jokes from an existing humor corpus into a First-Order Logic representation, extracting predicates and constants. The agent then performs context-aware semantic mapping, replacing relevant elements with more suitable equivalents in the target scenario. Finally, the mapped logical structure is realized as fluent, natural text. Through evaluation by participants, this framework provides a controllable, traceable, and interpretable implementation pathway for contextually grounded humor adaptation.

Introduction

Humor is a crucial conversational skill: it signals approachability, reduces social distance, and often functions to build rapport in everyday dialogue. However, humor is also one of the most context-sensitive forms of language. A joke that succeeds with one audience may fail with another due to differences in background knowledge, cultural references, or conversational norms (Xian et al., 2025). This fragility makes humor adaptation, that is, rewriting an existing joke for a new audience and scenario while preserving its comedic mechanism, an important but difficult problem for conversational AI and human–AI interaction (Amin & Burghardt 2020).

The need for controllable humor rewriting is increasingly practical in modern communication environments. In such mixed-audience conditions, a successful joke must lower the comprehension barrier while also reducing the probability of harmful implications. Because offensiveness is highly context-dependent, safety cannot be treated as a single generic filter; rather, it must be modeled as a function of audience and situation, and the rewriting process must remain traceable. Human-centered studies of toxicity detection systems further suggest that automated safety signals can be predictive but should be treated as supportive evidence rather than the sole arbiter of appropriateness (Muralikumar et al. 2023).

Large Language Models (LLMs) have recently demonstrated impressive surface-level rewriting ability, including style transfer and creative paraphrasing. Work on controllable text generation for LLMs emphasizes that real-world generation must satisfy explicit control conditions while retaining fluency and diversity; however, achieving such control often degrades text quality or becomes unstable in deployment (Liang et al. 2024). Relatedly, recent humor-specific research explores the transfer of humorous capability, underscoring that humor is a complex skill that benefits from additional supervisory signals beyond raw imitation (Ravi et al. 2024). Despite these advances, most end-to-end methods remain difficult to interpret: when a rewrite fails (unfunny, confusing, or offensive), it is hard to identify which semantic substitutions caused the problem.

To address these issues, we propose a humor-rewriting-agent framework that improves controllability and interpretability at the same time. The agent first converts jokes from an existing humor corpus (Tseng et al., 2020) into a First-Order Logic (FOL) intermediate representation by extracting predicates and constants (Gordon, 2025). It then performs context-aware semantic mapping to replace contextually risky or inaccessible elements with appropriate equivalents, and finally realizes the mapped logical structure into fluent natural language. This logic-based intermediate form is motivated by recent progress in NL-to-FOL translation with LLMs, which suggests that structured logical representations can improve controllability and support explicit inspection of meaning-preserving transformations (Yang et al., 2024). As a result, the rewriting process is traceable through predicate/constant substitutions in FOL and interpretable through traceable mapping decisions.

We evaluate the framework with literary experts, focusing on whether the rewritten jokes remain funny, become easier for the target audience to understand, and avoid offensiveness in the target scenario.

Related Work

Computational humor has traditionally focused on either generation (producing jokes) or recognition (detecting humorous content), with persistent challenges in modeling humor mechanisms and evaluating outputs reliably. Amin and

Burghardt (2020) provide an overview of joke-generation approaches and emphasize that evaluation often collapses into subjective "humorousness" judgments, making controllable humor a nontrivial target. In parallel, humor recognition has become a mature empirical area, including systematic reviews that summarize datasets, features, and deep-learning methods (Kalloniatis et al. 2024).

A key gap for humor rewriting is that humor must be adapted while also managing social risk: a rewrite must remain funny but also become less offensive or more context-appropriate for a new audience. This dual objective is reflected in shared-task benchmarks that jointly model humor and offense. For example, SemEval-2021 Task 7 (HaHackathon) explicitly combines humor detection and offense rating, providing annotated data and evaluation settings that highlight both subjectivity (variance/controversy) and safety constraints (offense scores) (Meaney et al. 2021). More recently, LLM-era work has explored transferring humorous capability through conditional humor paraphrasing and feedback-driven distillation, suggesting that intermediate signals (e.g., feedback or structured constraints) can help preserve subtle humor while improving controllability (Ravi et al. 2024). These trends motivate humor rewriting as a controlled transformation problem: preserve a comedic mechanism while adapting topical grounding and offensiveness for a target context.

Controllable text generation (CTG) provides a general framework for meeting explicit constraints, such as topic, style, or safety, while maintaining fluency and usefulness. A recent survey frames CTG in the LLM era as a spectrum of content control and attribute control, covering techniques from prompting and decoding-time steering to fine-tuning and RL-based alignment, and noting recurring trade-offs between control strength and text quality (Liang et al. 2024). For our humor rewriting task, topic control is especially important because reference accessibility and taboo sensitivity are audience-dependent.

Beyond what we can talk about, interpretability requires visibility into what was changed. This motivates structured intermediate representations (IRs) that separate semantic structure from surface realization. In particular, natural-language to first-order logic (NL-FOL) translation has advanced significantly with LLMs: LogicLLaMA demonstrates that a fine-tuned model can translate NL statements into FOL rules more reliably than generic LLM baselines (Yang et al. 2024). Training-free approaches also exist; for example, CODE4LOGIC frames NL-FOL translation as progressive code generation and reports improved performance and generalization in an in-context setting (Liu et al. 2025). These developments support a humor rewriting design in which jokes are mapped into an IR (predicates/-constants), transformed via context-aware semantic substitutions, and then realized back into fluent text, yielding a controllable and traceable pathway rather than an opaque end-to-end rewrite.

Method

We propose a humor rewriting agent that adapts text to different audiences and scenarios while preserving the original comedic mechanism and reducing the risk of offensiveness. The core design goal is controllable and traceable humor adaptation: instead of relying on end-to-end rewriting alone, the system uses a First-Order Logic (FOL) intermediate representation to support interpretable semantic mapping.

Algorithm 1: Conversational Core for First-Order Logic (FOL) Extraction

Input: Natural language text T
Output: Formalization JSON $\mathcal{J} = \{\texttt{vocabulary} : V, \texttt{constants} : C, \texttt{formalizations} : F\}$

Initialize conversation history $H \leftarrow [\texttt{SystemPrompt}(\ell)]$
Initialize global state: $V \leftarrow \emptyset$, $C \leftarrow \emptyset$, $F \leftarrow [\,]$
$p_{simp} \leftarrow \texttt{BuildSimplifyPrompt}(T, \ell)$
$r_{simp} \leftarrow \texttt{LLMCall}(H, p_{simp})$
$S \leftarrow \texttt{ValidateAndParseSentences}(r_{simp})$ {simple sentences; coreference resolved}
for $i = 1$ **to** $|S|$ **do**
  $ctx \leftarrow \texttt{FormatContext}(V, C)$ {reuse symbols; reduce re-definitions}
  $p_i \leftarrow \texttt{BuildFormalizePrompt}(S_i, ctx, \ell)$
  $r_i \leftarrow \texttt{LLMCall}(H, p_i)$
  $(\Delta V_i, \Delta C_i, L_i, R_i) \leftarrow \texttt{ValidateAndParseFormalization}(r_i)$
  $V \leftarrow V \cup \Delta V_i$
  $C \leftarrow C \cup \Delta C_i$
  $F \leftarrow F \mathrel{+}= [\texttt{Record}(S_i, L_i, R_i)]$
  $H \leftarrow H \mathrel{+}= [(p_i, r_i)]$ {conversation memory}
end for
return $\{\texttt{vocabulary} : V, \texttt{constants} : C, \texttt{formalizations} : F\}$

Text-to-FOL Conversion

To make rewriting interpretable and traceable, we first map the input text T_s into a First-Order Logic (FOL) representation:

$$LF(T_s) = \{\ell_1, \ldots, \ell_m\}, \tag{1}$$

where each literal ℓ is a predicate applied to constants (entities/objects) and, when necessary, variables.

We implement this conversion via an LLM-assisted, conversational NL→FOL pipeline that follows our core extraction algorithm. Rather than translating a long, stylistically rich text in a single pass, we adopt a progressive procedure that (i) simplifies the text into atomic propositions and (ii) formalizes these propositions iteratively while maintaining a growing shared context. Concretely, the extractor maintains a global *Vocabulary* V and *Constants* C, and uses them as dynamic context to improve entity consistency and avoid redundant predicate definitions.

Operationally, the parser proceeds as follows:

1. **Text simplification.** Decompose T_s into a sequence of simple sentences $S = \langle S_1, \ldots, S_n \rangle$ (removing subordinate clauses and resolving pronouns).
2. **Iterative formalization with memory.** For each sentence S_i, prompt the model with the current context

Original text (exam)	Original FOL skeleton	Adapted FOL skeleton	Final text (tech workplace)
A university holds a final exam.	(holds_exam UNIVERSITY FINAL_EXAM)	(holds_review BIG_TECH QUARTERLY_REVIEW)	A big tech company holds a quarterly review.
Three students arrive late and ask for a makeup, claiming a flat tire.	(arrives_late STUDENT{1,2,3}) (requests_makeup STUDENT{1,2,3} PROFESSOR) (explains STUDENT{1,2,3} ''FLAT_TIRE'')	(arrives_late ENG{1,2,3}) (requests_makeup ENG{1,2,3} DIRECTOR) (explains ENG{1,2,3} ''MODULE_CRASH'')	Three engineers arrive late and ask for a makeup review, claiming a module crash.
A makeup is granted; one week later, a single question checks the excuse detail, their answers diverge, and all fail.	(agrees PROFESSOR ''MAKEUP_NEXT_WEEK'') (given_question PROFESSOR ''WHICH_TIRE?'') (answers STUDENT1 TIRE1), (answers STUDENT2 TIRE2), (answers STUDENT3 TIRE3) (fails STUDENT{1,2,3})	(agrees DIRECTOR ''MAKEUP_NEXT_WEEK'') (given_question DIRECTOR ''WHICH_MODULE?'') (answers ENG1 MODULE_A), (answers ENG2 MODULE_B), (answers ENG3 MODULE_C) (fails ENG{1,2,3})	A makeup review is granted; one week later, a single question asks which module crashed. They answer differently, so all fail.

Table 1: Humor adaptation example with compact FOL notation.

(V, C) and obtain a structured response containing newly introduced items $(\Delta V_i, \Delta C_i)$ and sentence-level literals L_i. Update the global state as $V \leftarrow V \cup \Delta V_i$ and $C \leftarrow C \cup \Delta C_i$, and retain the conversation history to strengthen cross-sentence entity resolution.

3. **Ordered literal assembly.** Concatenate the sentence-level literals in narrative order, $LF(T_s) \leftarrow L_1 \oplus \cdots \oplus L_n$, to preserve the punchline mechanism (e.g., contrast, mistaken identity, literalization) at the level of explicit predicates and constants.

This traceability layer enables explicit inspection: we can directly identify which *constants* encode contextually specific references and which *predicates* encode the joke's logical twist, facilitating systematic debugging and controlled transformations during rewriting.

FOL-guided Abstraction and Scenario Recontextualization for Joke Rewriting

Given an extracted FOL traceability representation $\mathcal{J} = \{\texttt{vocabulary} : V, \texttt{constants} : C, \texttt{formalizations} : F\}$, we design a second LLM-driven pipeline that (i) abstracts V and C into concept-level placeholders, (ii) re-instantiates them into a target scenario while preserving global logical consistency, and (iii) realizes the result as a rewritten joke that matches the source text in length and rhetorical style.

Operationally, the parser proceeds as follows:

1. **Concept-level abstraction.** Lift each constant and predicate in (V, C, F) into higher-level conceptual symbols, producing (V^*, C^*, F^*) and an explicit abstraction map π_{abs}. This forces the representation to encode roles and functional relations (e.g., *authority*, *disguise*, *novice*, *expert tool*) so later substitutions are structurally motivated rather than purely lexical.
2. **Scenario grounding via explicit mapping.** Re-instantiate abstract symbols into a target scenario σ by constructing a grounding map π_σ and generating the grounded structures $(V_\sigma, C_\sigma, F_\sigma)$. The explicit mapping ensures that all mentions of the same abstract entity are instantiated consistently across sentences and literals.
3. **Style-preserving realization.** Realize F_σ back into fluent text under rhetorical constraints Γ extracted from the source (length, beat distribution, dialogue structure, rhetorical devices), so that the rewritten joke remains comparable to the original in scope and style while expressing the new scenario.

Evaluation

To conduct an initial evaluation during the early development phase and to inform subsequent iterative design, we invited three participants from Taiwan to test our humor rewriting agent in a controlled pilot study.

Before the activity began, we provided a short briefing on the operating principles of the system. In particular, we highlighted (1) the possibility of hallucination in LLM-based rewriting and (2) the role of our safety constraints (topic restrictions and toxicity-aware signals) as supportive safeguards rather than absolute guaranties. We also explained how the system's intermediate representations (FOL predicates/constants and mapping logs) enable traceability: par-

	Participant 1	Participant 2	Participant 3
Humorousness	It delivers clean irony. The image is less vivid than the tire version.	It reads as workplace satire. It feels somewhat predictable in review culture.	The twist is understandable. Technical terms reduce immediacy.
Clarity (target audience)	It is mostly clear from context. A brief gloss helps non-engineers.	It is clear and realistic for a tech review scenario.	It is followable overall. Some terms require inference.
Appropriateness / offensiveness	It is appropriate. It targets inconsistency. It does not target any group.	It is appropriate for workplace humor. It avoids identity-based ridicule.	It is not offensive. It reads as a situational joke.
Punchline faithfulness / mechanism preservation	The mechanism is preserved. A shared excuse is tested by one question. Inconsistency is exposed.	The core structure is preserved. The key detail becomes more domain specific.	The logic matches the original. They cannot agree on the crucial detail.
Traceability usefulness (mapping log)	The log helps track meaning-preserving substitutions.	The log helps verification. Target users may not always need it.	The log helps explain the correspondences.

Table 2: Concise qualitative comments from three participants on the rewritten joke across five evaluation dimensions.

ticipants could inspect what was replaced and why, which is especially important when humor crosses situational or cultural boundaries.

After the interaction ended, we conducted semi-structured interviews to gather qualitative feedback on the user experience and to identify areas for improvement. We asked participants to assess each rewritten joke across five dimensions: (1) humorousness, (2) clarity of comprehension for the target audience, (3) appropriateness (i.e., perceived offensiveness) in the target scenario, (4) punchline faithfulness / mechanism preservation, and (5) traceability usefulness (the extent to which the mapping log helped them understand the rewrite). This evaluation design follows prior work that treats humor and offensiveness as scalable human annotations and emphasizes variability and disagreement as informative signals in humor perception.

Future Work

Based on participants' feedback and the current limitations of humor rewriting systems, future development will focus on the following directions:

1. **Support for Multicultural and Cross-Lingual Humor:** Humor is highly culture-dependent and often relies on wordplay and culturally specific references. Future work will extend the audience preference beyond a single contextual setting and incorporate cross-lingual humor transfer and culture-specific reference grounding so that rewrites remain both comprehensible and socially appropriate in diverse contexts.
2. **Multimodal Humor Rewriting (Speech, Memes, Comics):** Real-world humor frequently involves prosody, timing, facial expressions, images, and meme templates. Future versions of the agent will extend the framework to multimodal humor such as memes and comics, where visual context is essential for both interpretation and generation.
3. **Adaptive Learning and Personalization Over Time:** Effective humor adaptation should evolve with repeated interaction. We plan to develop a self-evolving personalization module that updates audience priors and taboo sensitivity based on user feedback, acceptance/rejection signals, and interaction traces, following recent trends in dynamically adapted personalized dialogue agents and personalization research.
4. **Verified Semantic Mapping and Logic Consistency:** While FOL provides traceability, NL$\rightarrow$FOL conversion can still introduce structural errors, especially in small or open-domain settings. We will explore verification-based pipelines (e.g., error taxonomy + correction) and symbolic-prover-assisted checks to ensure that mapped logical structures preserve entailments relevant to the punchline mechanism and do not introduce unintended implications

Conclusion

This study proposes a humor-rewriting-agent framework for contextually grounded joke adaptation that explicitly targets the dual requirement of being funny and appropriate. The framework uses a First-Order Logic (FOL) intermediate representation to make semantic substitutions traceable and inspectable. This design responds to long-standing challenges in computational humor, especially the difficulty of controlling humor mechanisms and evaluating outputs reliably.

We believe this framework has strong long-term potential as a practical pathway toward traceable and interpretable humor adaptation in conversational agents, moderation-assisted rewriting tools, and cross-context communication support. As NL-to-FOL translation techniques continue to improve, logic-grounded rewriting can become more reliable and verifiable, further strengthening transparency in meaning-preserving cultural adaptation.

Acknowledgments

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-].

References

Amin, M.; and Burghardt, M. 2020. A survey on approaches to computational humor generation. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 29–41.

Gordon, A. 2025. Automated Formalization of William Shakespeare's Venus and Adonis Using Large Language Models. In *Twelfth Annual Conference on Advances in Cognitive Systems*.

Kalloniatis, A.; and Adamidis, P. 2024. Computational humor recognition: a systematic literature review. *Artificial Intelligence Review*, 58(2): 43.

Liang, X.; Wang, H.; Wang, Y.; Song, S.; Yang, J.; Niu, S.; Hu, J.; Liu, D.; Yao, S.; Xiong, F.; et al. 2024. Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*.

Liu, J. 2025. Few-Shot Natural Language to First-Order Logic Translation via Code Generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 10939–10960.

Meaney, J.-A.; Wilson, S.; Chiruzzo, L.; Lopez, A.; and Magdy, W. 2021. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 105–119.

Muralikumar, M. D.; Yang, Y. S.; and McDonald, D. W. 2023. A human-centered evaluation of a toxicity detection api: Testing transferability and unpacking latent attributes. *ACM Transactions on Social Computing*, 6(1-2): 1–38.

Ravi, S.; Huber, P.; Shrivastava, A.; Sagar, A.; Aly, A.; Shwartz, V.; and Einolghozati, A. 2024. Small but funny: A feedback-driven approach to humor distillation. *arXiv preprint arXiv:2402.18113*.

Tseng, Y.-H.; Hsu, W.-L.; Wu, W.-S.; Gu, Y.-C.; and Chen, H.-C. 2020. Implementation and Evaluation of a Retrieval-based Chinese Humor Chatbot. *Journal of Library & Information Studies*, 18(2).

Xian, A. F. C.; Ting, N. C.; Cheng, A. K. S.; Tan, W. Y.; Chanthran, M. R.; Soon, L.-K.; and Lee, M. 2025. Laughing Across Borders: A Culturally-Aware Joke Generator for Asian Regions. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 921–925. IEEE.

Yang, Y.; Xiong, S.; Payani, A.; Shareghi, E.; and Fekri, F. 2024. Harnessing the power of large language models for natural language to first-order logic translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6942–6959.