K-Humor: Dynamic Multi-Agent LLM Framework for Generation of Culturally Grounded Korean Jokes

Minsu Park[1] **Heejeong Jeon**[1,2] **DongJae Kim**[1] **Minwoo Oh**[1] **Eunil Park**[1,3,*]
[1]Sungkyunkwan University [2]ALLBIGDAT [3]Jaume I University
{mspark501, iamheej, bronze.ash, mwoh0208}@g.skku.edu, eunilpark@skku.edu

Abstract

Humor generation is a challenging task in natural language generation that requires fluency, surprise, and cultural grounding. Although large language models (LLMs) show promise in English, little work addresses low-resource and culturally specific humor, such as in Korean. We propose **K-Humor**, a *dynamic multi-agent framework* for generating Korean jokes. Unlike static one-shot models, K-Humor integrates specialized agents for cultural awareness, audience simulation, editing, and punchline creation, which are coordinated through a dynamic iterative refinement loop. A human-data-grounded mechanism further guides the punchlines using authentic Korean jokes. Evaluation with automatic metrics, agent-internal diagnostics, LLM-as-a-judge, and human ratings shows that K-Humor consistently outperforms a single-prompt baseline and ablations in criteria such as fluency and cultural appropriateness. Although human-written jokes remain stronger overall, particularly on core creative dimensions, K-Humor demonstrates that dynamic, culturally grounded multi-agent systems can improve humor generation over standard LLM baselines in low-resource settings. Code is available at https://github.com/dxlabskku/k-humor.

1 Introduction

Computational humor generation has long been considered a challenging benchmark for natural language generation (NLG) because it requires modeling linguistic incongruity, cognitive surprise, and cultural context. Early systems produced rule-based puns (Binsted, 1996), and later, neural approaches introduced data-driven methods (Yu et al., 2018; He et al., 2019). However, such methods are limited by their contextual grounding.

With recent advances in LLMs, humor generation has expanded using methods such as multistep reasoning (Tikhonov and Shtykovskiy, 2024) and curriculum learning (Chen et al., 2024b). Nonetheless, most systems remain single-pass generators and are largely trained in English, overlooking the cultural grounding required for humor (Myung et al., 2024; Kirk et al., 2024). These issues are particularly acute for Korean: multilingual LMs show degraded performance on Korean tasks, motivating the development of monolingual Korean models (Ko et al., 2023), and baseline single-pass LLM systems struggle with pragmatics, such as sarcasm in Korean corpora (Kim et al., 2024). Combined with the unique morphological features of Korean (e.g., case particles, honorifics, and flexible word endings) (Oh et al., 2021; Kwon and Sturt, 2024), these challenges make culturally resonant humor generation in Korean underexplored.

Meanwhile, LLMs are increasingly organized into **multi-agent systems** that collaborate through debate (Du et al., 2024) or dynamic communication (Liu et al., 2024). Such systems allow adaptive workflows that go beyond the limits of a single LLM, enabling iterative refinement, critique,

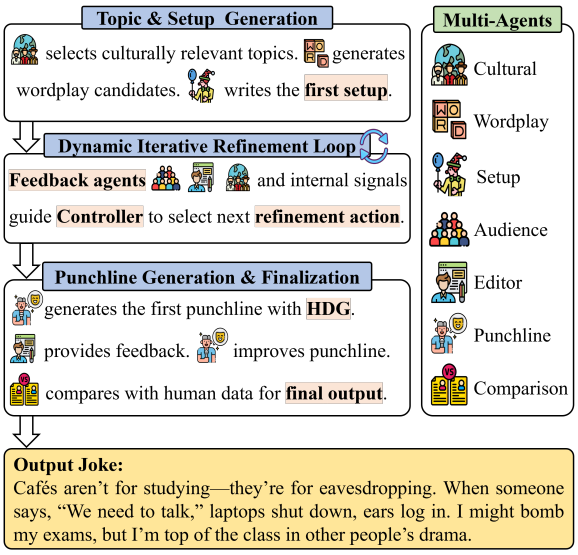


Figure 1: Overall framework of K-Humor.

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 41292–41315
October 24-29, 2026 ©2026 Association for Computational Linguistics

and labor division. Such competencies extend to creative NLG. Chen et al. (2024a) explored multi-agent use with role-playing and feedback-based writing, but still their potential for low-resource, culturally aware humor generation remains unexplored. We argue that multi-agent coordination is a natural fit for humor, as comedy writing is an iterative, collaborative process shaped by audience feedback and cultural norms. However, existing humor generation methods rarely model this iterative and collaborative nature of real-world comedy writing.

To address this gap, we propose **K-Humor**, a dynamic multi-agent pipeline for Korean humor generation. The system decomposes humor writing into specialized agents for cultural grounding, wordplay, setup construction, punchline generation, and multi-perspective feedback (audience, editor, and culture). At its core is a dynamic iterative loop, where candidate setups are iteratively refined. Unlike prior pipelines, where the next step is implicitly determined by the language model, our system introduces an explicit decision mechanism that selects the next refinement action based on intermediate feedback and quality signals. To address data scarcity, we introduce a **human-data-grounded (HDG)** mechanism. HDG retrieves the top analogous human-written Korean jokes as few-shot exemplars, guiding punchlines toward a culturally authentic style and structure.

Experiments combining LLM, human, internal, and automatic evaluations show that K-Humor outperforms the single-pass baseline and ablations, while human-authored jokes remain stronger on core creative dimensions.

Our contributions are:

- We introduce a **dynamic multi-agent system** for Korean humor generation that incorporates an explicit refinement loop with adaptive action selection.
- We propose a novel **HDG** retrieval method that leverages a curated Korean joke corpus to guide stylistic and cultural adaptation in low-resource settings.
- We provide a **comprehensive evaluation** using automatic, LLM-based, and human assessments, showing gains over baselines and ablations.

2 Related Work

2.1 Computational Humor Generation

Computational humor has evolved from symbolic rule systems (e.g., *JAPE* (Binsted, 1996)) to neural models and, recently, LLM-based frameworks. The core challenge remains capturing the interplay of linguistic structure, surprise, and cultural context.

To overcome the rigidity of templates, the field has shifted toward data-driven methods. Data-driven pun generation seeks to learn humor from general corpora with minimal annotation, such as the retrieve-and-edit technique for local–global surprisal (He et al., 2019). Variants (seq2seq for homographs; GANs for diversity) improved fluency but remained largely lexical and context-shallow (Yu et al., 2018; Luo et al., 2019).

LLMs enable the modeling of the cognitive and social facets of humor. Wang et al. (2025a) proposed a reinforcement-learning (RL) framework that leverages external information for humor reasoning, and Chen et al. (2024b) employed a multi-stage curriculum learning framework to progressively teach a model the structure and preferences of puns. Mittal et al. (2022) exploited contextual ambiguity via one-shot prompting, and Kim and Chilton (2025) developed the *HumorSkills* pipeline to generate meme captions tailored to a specific audience (Gen Z). The creation of new, culturally specific resources, such as the *TalkFunny* dataset with its "chain-of-humor" explanations for Chinese dialogue (Chen et al., 2024c), further underscores this trend. Despite these advances, most systems are single-pass and fail to capture iterative and feedback-driven comedy writing.

2.2 LLMs for Low-Resource Languages and Cultures

Multilingual LLMs (CONNEAU and Lample, 2019; Xue et al., 2021) have improved linguistic coverage by training on diverse corpora with shared word pieces; however, they often reflect a Western cultural bias (Johnson et al., 2022; Liu, 2023; Myung et al., 2024). Although benchmarks have broadened language inclusion (Bandarkar et al., 2024; Team et al., 2022; Muhammad et al., 2023), few target cultural diversity issues.

Recent work continues to expose this cultural gap in dealing with low-resource settings. *CVQA* demonstrates multimodal tasks' failures on non-Western cultural events and objects (Romero et al., 2024). *PRISM* dataset reveals that human feedback

on LLM outputs is highly subjective and varies significantly across different cultural and demographic groups (Kirk et al., 2024).

To achieve better cultural coverage, Tao et al. (2024) employed cultural prompting as a control strategy to increase the cultural alignment. However, the *CultureCare* benchmark demonstrates that simple role-play prompting is insufficient for generating culturally appropriate responses (Liu et al., 2026).

Overall, inadequate cultural grounding impairs LLMs in nuanced generation tasks, particularly those that are culturally sensitive.

2.3 Multi-Agent and Dynamic LLM Systems

Recent research increasingly framed LLMs as autonomous or semi-autonomous agents that collaborate to perform complex tasks. Early work (Du et al., 2024) demonstrated that a fixed number of debating agents can enhance factuality and reasoning. Recently, dynamic frameworks such as DyLAN focus on adaptive communication structures among agents, allowing flexible coordination (Liu et al., 2024).

Large-scale orchestration frameworks, such as TaskMatrix.AI, show how LLMs can be integrated with external tools and APIs to achieve compositional task solving (Liang et al., 2024). Wang et al. (2025b) demonstrates fully autonomous multi-agent systems with a single prompt, and in the creative domain, HoLLMwood illustrates the potential of multi-agent systems for feedback-based writing and character playing (Chen et al., 2024a).

Given this trend, we posit that multi-agent systems offer an effective solution for culturally grounded humor generation in Korean, thereby addressing the limitations of single-pass LLM approaches.

3 System Architecture and Methodology

3.1 Overall System Architecture

Our system is an iterative multi-agent pipeline designed to generate culturally specific humor in low-resource settings, as illustrated in Figure 1. Inspired by professional comedy writing workflows, such as television writers' rooms, in which creative teams pitch, critique, and iteratively rewrite scripts (Phalen and Osellame, 2012; Taylor and O'Meara, 2025), our architecture enables agents to play analogous roles in the refinement of jokes. This design departs from conventional

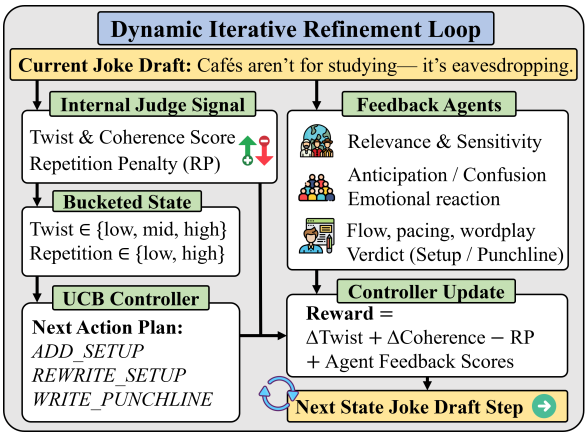


Figure 2: Dynamic iterative refinement loop overview.

single-prompt generation by implementing a dynamic multistage process in which agents collaborate and refine their outputs based on cultural, structural, and audience-driven feedback. The pipeline operates in three phases:

Phase 1: Topic, Wordplay, and Setup Generation The process begins with the **cultural agent**, which proposes a culturally relevant topic, ensuring that the generated humor resonates with the target audience. Building on this, the **wordplay agent** introduces creative linguistic devices, such as puns, double meanings, or phonetically playful expressions that can serve as seeds for humor. These wordplay suggestions enrich the semantic space and provide a foundation for creating original content. The **setup agent** then produces the initial premise or "setup" of the joke, optionally incorporating wordplay if appropriate. In contrast to prior one-shot generation approaches, this phase establishes both cultural grounding and comedic potential from the outset.

Phase 2: Dynamic Iterative Refinement Loop This phase, illustrated in Figure 2, forms the dynamic core of the proposed system. The evolving joke draft is evaluated by three feedback agents: the **Audience Agent** (simulating audience reactions), the **Editor Agent** (providing structural critique and quality control), and the **Cultural Agent** (ensuring cultural appropriateness). Their feedback, together with intermediate quality signals such as twist, coherence, and repetition, is used by an adaptive controller that determines the next refinement action. Based on the current joke state, the controller selects actions such as extending the setup, revising the existing setup, or transitioning to punchline generation. This process is repeated iteratively, al-

lowing the system to progressively refine the joke until the controller determines that the punchline should be generated.

Phase 3: Punchline Generation and Final Selection Once the setup is finalized, the **Punchline Agent** generates the conclusion of the joke. The *Editor Agent* then provides a final review of the complete joke, resulting in a polished output. At this stage, the system also incorporates the **Human-Data-Grounded (HDG)** mechanism, which provides stylistic and cultural guidance by leveraging human-generated jokes. Finally, the **Comparison Agent** evaluates the generated joke against a randomly selected human-written reference and scores it based on dimensions such as humor, clarity, surprise, naturalness, originality, and consistency. This final step acts as an internal quality filter, helping the system to retain only the strongest outputs. The complete agent prompts are provided in Appendix B.

3.2 Dynamic Refinement Loop

The core of our system is a controller-guided refinement loop that incrementally develops a joke before producing the final punchline, inspired by recent iterative and agent-based refinement loop works (Madaan et al., 2023; Chen et al., 2024a; Lin et al., 2025). Unlike conventional single-pass generation, we model joke construction as a sequence of refinement steps where the system repeatedly observes the current draft, receives feedback, and selects the next editing action.

At each iteration, the evolving draft is evaluated by several feedback agents. The **Audience Agent** estimates likely audience reactions, the **Editor Agent** evaluates structural quality and punchline readiness, and the **Cultural Agent** checks the contextual and cultural appropriateness. In parallel, an internal judge scores the draft with two signals: *twist*, capturing surprise potential, and *coherence*, capturing structural consistency. We also compute a *repetition* signal that penalizes drafts that become semantically similar to previously generated jokes.

A controller then selects the next refinement action from a small action space centered on setup expansion, setup revision, and punchline generation (`ADD_SETUP`, `REWRITE_SETUP`, and `WRITE_PUNCHLINE`), with termination handled separately through a `STOP` decision. Instead of training a full reinforcement-learning policy, we adopt a lightweight contextual bandit with an upper confidence bound (UCB) strategy (Auer et al., 2002). This choice is motivated by the short decision horizon and the availability of immediate reward signals, which makes bandit optimization more stable and sample-efficient than full RL (Shi et al., 2023; Salgia, 2023).

To obtain reliable state-action statistics under a limited generation budget, we discretize the draft state into coarse buckets defined by twist and repetition levels. Although these signals could in principle be treated as continuous controller inputs, a continuous state space would substantially reduce the number of repeated observations for similar states, making the UCB estimates sparse and less stable. Moreover, the controller operates over a small set of coarse editing decisions, such as continuing setup refinement or transitioning to punchline generation; discrete states therefore provide a natural and interpretable abstraction for these decision regimes. Specifically, twist is divided into three ranges (low, medium, high) and repetition into two levels (low, high), producing six contextual buckets. For each bucket, the controller maintains action statistics and selects actions using a UCB policy. Coherence is logged and used in reward computation but is not included in the bucket definition because its values tend to cluster within a narrow, high range.

After each action, the controller receives an immediate reward computed as $r = \alpha\Delta\textit{twist} + \beta\Delta\textit{coherence} - \gamma\,\textit{rep} + w_a s_{\text{aud}} + w_e s_{\text{edit}} + w_c s_{\text{cult}}$, where $\Delta\textit{twist}$ and $\Delta\textit{coherence}$ denote improvements in the internal judge scores, *rep* is a semantic repetition penalty computed from the embedding similarity with previous outputs, and $s_{\text{aud}}, s_{\text{edit}}, s_{\text{cult}}$ are scalar feedback scores derived from the Audience, Editor, and Cultural agents. The controller updates the action statistics of the corresponding bucket using this reward and decides whether to continue expanding the setup or to generate the final punchline. Unlike prior iterative prompting approaches that rely on fixed heuristics, our controller explicitly optimizes the action selection using reward-driven bandit learning. Further details are available in Appendix Section E.

3.3 Human-Data-Grounded (HDG) Generation

To address the low-resource nature of humor in Korean, we introduce an **HDG** mechanism (Figure 3). Unlike retrieval-augmented generation (RAG), which typically retrieves factual knowl-

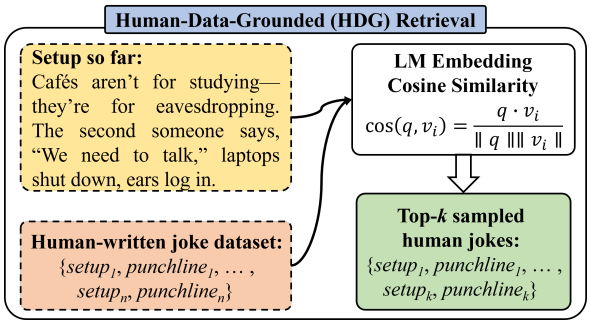


Figure 3: Illustration of Human-Data-Grounded (HDG) Retrieval.

edge, HDG dynamically retrieves human-written jokes to provide culturally and stylistically relevant examples for punchline generation.

The HDG pipeline proceeds as follows:

- **Dataset:** We curate 180 Korean jokes from the Namuwiki[1] page on the *Choi Bul-am Series*, each split into 'setup'/'punchline'.
- **Embedding Representation:** We encode the *setups* using a RoBERTA based Korean sentence embedding model (KoSimCSE[2]), which captures semantic similarity between jokes beyond surface lexical overlap.
- **Semantic Similarity Retrieval:** Given the generated setup, we compute cosine similarity between its embedding and the embeddings of all corpus setups, and rank candidates according to their similarity.
- **Few-Shot Example Construction:** We retrieve the top $\mathbf{k{=}4}$ most similar human-written jokes and format them as few-shot exemplars for the Punchline Agent, to provide human guidance.

By grounding punchline generation in human-written Korean examples, HDG provides additional stylistic guidance under the limited-data setting. This ensures that the generated jokes not only follow effective structural patterns but also reflect the cultural nuances necessary for natural humor.

4 Experiments

4.1 Dataset

We evaluate our system on the collected Korean joke dataset, the *Choi Bul-am Series*, a well-known collection of short comedic stories featuring a recurring character. The dataset is analogous to "Francesco Totti jokes" in Italian culture and is regarded as a canonical set of humorous narratives. Each joke was segmented into a 'setup' and 'punchline', with the final line treated as the punchline and the preceding narrative as the 'setup'. A sample is provided in Appendix Table 17. For fair experiments, the jokes were randomly sampled for all evaluations.

4.2 Model Setup

All generation experiments use the OpenAI API model `gpt-4.1`, while the lightweight internal judge used by the controller is `gpt-4.1-mini`. Both model identifiers refer to the API aliases available at the time of our experiments (accessed on March 10, 2026). Unless otherwise noted, we set the temperature and top-p to 1.0. The Wordplay Agent uses a slightly higher temperature (1.2) to encourage creative associations, while setup expansion uses a lower sampling diversity ($p = 0.5$) to maintain narrative coherence.

The dynamic refinement loop is implemented as described in Section 3.2. The controller operates under a fixed budget of eight refinement steps and uses a contextual bandit with a UCB exploration coefficient of $c = 1.0$. The action space consists of three operations: `ADD_SETUP`, `REWRITE_SETUP`, and `WRITE_PUNCHLINE`. Details on reward computation are provided in Appendix Section E.

For HDG retrieval, the system retrieves the top-k nearest human joke examples ($k = 4$) using cosine similarity over KoSimCSE setup embeddings. These retrieved examples provide few-shot guidance for punchline generation. We also compare alternative retrieval representations (TF-IDF and KS-BERT) in Appendix Section F.

4.3 Baselines and Ablation Studies

We first establish a **single-prompt baseline**, in which `gpt-4.1` is directly asked to "generate a funny joke in Korean" with a setup and punchline. This model uses no iterations, retrieval, or cultural checks, representing the most straightforward use of an LLM for humor generation. The prompt is provided in Appendix Table 16.

We then conduct **ablation studies** by systematically removing the components of our full pipeline. Removing the iteration disables the refinement loop, leaving only a single 'setup'-'punchline' pass with HDG and cultural guidance. Removing the

[1] https://namu.wiki
[2] https://huggingface.co/BM-K/KoSimCSE-roberta

System	Funny ↑	Surprise ↑	Clarity ↑	Cultural ↑	Fluency ↑	Original ↑	Coherence ↑	Overall ↑
Baseline	2.86	2.69	4.37	4.87	4.54	2.61	3.93	3.04
No-Culture	3.06	2.62	4.67	3.33	4.78	2.76	3.90	3.11
No-HDG	3.56	2.95	4.76	4.93	4.84	3.04	4.12	3.77
No-Iter	3.03	2.78	4.07	4.48	4.26	2.8	3.75	3.09
K-Humor	3.58	2.92	**4.92**	**4.97**	**4.95**	3.12	4.05	3.82
Human	**4.13**	**4.21**	4.64	4.51	4.73	**4.01**	**4.88**	**4.31**

Table 1: Likert-scale LLM-as-a-Judge results (1-5). Arrows indicate preferred direction.

Comparison	Non-Ties	Win Rate	95% CI
vs. Baseline	48	0.71	[0.57, 0.82]
vs. Human	50	0.62	[0.48, 0.74]
vs. No-Culture	48	0.58	[0.44, 0.71]
vs. No-HDG	40	0.60	[0.45, 0.74]
vs. No-Iter	39	0.67	[0.51, 0.79]

Table 2: LLM-as-a-Judge pairwise evaluations. Values are the proportion of full pipeline wins (ties excluded). The win rate is calculated as the proportion of comparisons in which the Full pipeline was preferred.

HDG prevents the retrieval of exemplar jokes. Removing the Cultural Agent disables both culturally guided topic proposal and cultural feedback during refinement, while preserving the remaining iterative and HDG components.

Controller baselines. To assess whether adaptive action selection is necessary, we further compare the proposed Bucket-UCB controller with two non-adaptive heuristic policies. The **Fixed 2-Step** policy performs two setup-refinement steps before transitioning to the punchline generation. The **Coherence Threshold** policy continues setup refinement until a predefined coherence criterion is satisfied before generating a punchline. These baselines retain the remaining components of the K-Humor pipeline and differ only in how the refinement trajectory is controlled.

4.4 Evaluation Methodology

LLM-as-a-Judge. To complement the internal and automatic metrics, we employ the OpenAI API model `gpt-5` (accessed March 10, 2026) as the LLM evaluator, following recent conventions (Kocmi and Federmann, 2023; Zheng et al., 2023; Liusie et al., 2024). The model is used in two modes: (i) *pairwise comparison*, where two candidate jokes are presented side-by-side and the LLM selects the better one (or declares a tie); and (ii) *Likert-style scoring*, where a single joke is rated on a 1-5 scale across seven criteria and an overall assessment. To mitigate positional bias in pairwise evaluation, each joke pair is evaluated twice with the A/B order reversed. We evaluate all **60 jokes per system**, using two pairwise judgments per comparison and three Likert ratings per joke. This practice results in several hundred evaluations, providing both relative preferences and absolute-quality estimates. Such LLM-based judgments provide a cost-efficient and reproducible supplement to human evaluations. The complete pairwise and Likert evaluation prompts are provided in Appendix D.

Human Evaluation. Finally, 30 native Korean speakers evaluated the jokes using two tasks: (i) pairwise preference between K-Humor and either baseline or human-written jokes, with the A/B order randomized, and (ii) independent 1-5 Likert ratings on funniness, surprise, clarity, cultural appropriateness, fluency, originality, setup-punchline coherence, and overall quality. The original cultural appropriateness criterion jointly captured cultural relevance, audience suitability, and the absence of inappropriate content. Because these aspects were not annotated separately, we retain the original score and do not interpret it as an isolated measure of cultural grounding or safety. Each participant evaluated 10 K-Humor-baseline pairs, 10 K-Humor-human pairs, and nine Likert items (three per source). We report means, 95% confidence intervals, and Krippendorff's α (Krippendorff, 2011). Agreement is low overall ($\alpha = 0.124$), so fine-grained Likert differences should be interpreted with caution. This dual design balances relative system comparisons with absolute quality measurements, consistent with recent practices in NLG evaluation (Gatt and Krahmer, 2018; Howcroft et al., 2020; Mohankumar and Khapra, 2022; Ruan et al., 2024; van der Lee et al., 2021). The survey is provided in Appendix A.

Human-LLM agreement. To examine the agreement between human and LLM-based evaluations, we align nine jokes evaluated by both native Korean annotators and GPT-5, comprising three K-

System	Funny ↑	Surprise ↑	Clarity ↑	Cultural ↑	Fluency ↑	Original ↑	Coherence ↑	Overall ↑
Baseline	2.57	2.59	2.92	3.17	3.22	2.90	2.81	2.62
K-Humor	2.76	2.74	3.34	**3.73**	3.63	3.07	3.01	3.03
Human	**3.30**	**3.57**	**3.83**	3.43	**4.01**	**3.45**	**3.45**	**3.46**

Table 3: Human Likert ratings (1-5). Means are averaged across all items and raters for each source. Overall inter-rater agreement is low (Krippendorff's $\alpha = 0.124$).

Comparison	Non-Ties	Win Rate	95% CI
vs. Baseline	251	**0.65**	[0.58, 0.70]
vs. Human	255	0.44	[0.38, 0.50]

Table 4: Human evaluation on pairwise preference.

Humor outputs, three baseline outputs, and three human-written jokes. For each joke and criterion, we average the available human ratings and pair the resulting score with the corresponding GPT-5 Likert scale score. We then compute Pearson's r and Spearman's ρ across the nine aligned jokes for each evaluation criterion.

Automatic Metrics. We also report several automatic statistics to characterize the surface properties of the generated jokes. Lexical diversity is measured using Distinct-n (Dist-2 and Dist-3), defined as the ratio of unique n-grams to the total n-grams at the corpus level. We additionally report self-BLEU (S-BLEU), which computes the BLEU scores of each joke against the remaining outputs from the same system; lower values indicate less lexical overlap among the generated jokes. Finally, setup-punchline coherence is estimated using the embedding cosine similarity computed with a Korean SentenceTransformer model (Gao et al., 2021; Reimers and Gurevych, 2019; Chuang et al., 2022).

These statistics provide coarse indicators of diversity, repetition, and semantic relatedness between the setups and punchlines. Because humor quality depends on subtle semantic shifts and cultural context, we use these metrics as complementary diagnostics rather than primary evaluation criteria.

5 Results

5.1 LLM-as-a-Judge Evaluation

Pair-wise Evaluation. To evaluate the quality of generated jokes, we first employ an **LLM-as-a-Judge** framework using GPT-5 to conduct pairwise comparisons of jokes. Each pair is evaluated in both A/B orders and aggregated into a pairwise outcome. We report win/loss proportions together with 95% Wilson confidence intervals over the resulting non-tied comparisons (Wilson, 1927).

Table 2 presents results across conditions. Compared to the baseline single-agent model, our full system won a clear majority of comparisons with a **71% win rate**. Compared to human-written jokes, the full pipeline achieved a **62% win rate**, although the confidence interval overlaps with chance. Against the No-Iter ablation, the Full pipeline maintained a strong advantage with a 67% win rate. Performance against No-HDG also favored the full system at 60%, while improvements over No-Culture were more modest at 58%.

Overall, K-Humor outperforms the baseline and ablations, while remaining competitive with human-written jokes.

Likert-scale Evaluation. We further evaluate the jokes using a Likert-scale rubric with seven criteria and an overall assessment. Scores range from 1 (very poor) to 5 (excellent). Table 1 reports average scores of all systems.

Human-written jokes unsurprisingly achieve the highest scores across most criteria, particularly in terms of funniness, surprise, and originality. Among the model-based systems, K-Humor achieves the highest overall score and the highest scores in clarity, cultural appropriateness, and fluency. As in the human evaluation, the cultural appropriateness score reflects the original combined criterion and therefore should not be interpreted as an independent measure of cultural grounding or safety. Compared with the baseline system, K-Humor substantially improves the overall quality (3.82 vs. 3.04 baseline) while also achieving higher scores in clarity, originality, and surprise.

The ablations show that the full system achieves the highest overall score, although individual ablations can outperform it on specific dimensions. Although human humor remains stronger overall, K-Humor achieves the strongest overall performance among the evaluated model-based systems.

5.2 Human Evaluation.

Tables 3 and 4 present the results of human evaluation. We recruited 30 native Korean speakers to evaluate the jokes using two complementary tasks: pairwise preference and Likert-scale rating.

Pairwise Evaluation. When directly compared with the baseline system, K-Humor is preferred in 65% (95% CI: [0.58, 0.70]), indicating that human evaluators generally find jokes generated by the proposed system more appealing overall. This result suggests that the multi-agent refinement process and cultural grounding mechanisms improve the perceived quality of the generated humor relative to a single-pass baseline.

When compared with human-written jokes, however, K-Humor achieves a win rate of 44%, meaning that human-authored jokes remain preferred overall. This gap is expected, given the complexity of humor generation, and highlights the remaining challenges in producing humor that matches human creativity and comedic timing.

Likert-scale Evaluation. The Likert ratings in Table 3 provide a more detailed view of the system behavior. Compared with the baseline system, K-Humor receives higher scores across all dimensions. The largest improvement is observed in the original cultural appropriateness criterion (3.73 vs. 3.17). Because this criterion jointly captures cultural relevance and audience suitability, including non-offensiveness, the improvement should not be attributed to cultural grounding alone. Rather, it indicates that K-Humor is rated more favorably on the combined cultural-appropriateness criterion used in our original evaluation. Detailed statistical significance tests for all the pairwise and Likert-scale evaluations are reported in Appendix Section G.

Despite these gains over the baseline, human-written jokes still achieve the highest scores on most criteria. This pattern suggests that while the proposed system improves structural quality and cultural alignment, generating genuinely surprising or highly humorous punchlines remains a challenging problem for current LLM-based approaches.

Overall, the human evaluation indicates that the proposed multi-agent framework improves over a standard LLM baseline, while a clear gap remains relative to human-written humor.

Category	Metric	Value
Generation process	Samples	60
	Steps per joke	5.68 ± 0.81
	Joke length (chars)	341.3 ± 74.56
	Step range	3-8
Action selection	ADD_SETUP	172 (50.4%)
	REWRITE_SETUP	50 (14.6%)
	WRITE_PUNCHLINE	59 (17.3%)
	STOP	60 (17.6%)
Reward and penalties	Mean reward	1.10 ± 1.06
	Mean repetition penalty	0.69 ± 0.14
	WRITE vs. SETUP reward	1.64 vs. 0.96
Humor signal change	Twist (pre $\rightarrow$ post)	$2.91 \rightarrow 3.11$
	Coherence (pre $\rightarrow$ post)	$4.62 \rightarrow 4.74$
	Mean Δ twist	+0.20
	Mean Δ coherence	+0.12
Punchline dynamics	Editor punchline verdict	163
	Audience punchline verdict	150
	Lag to WRITE_PUNCHLINE	1.76 steps

Table 5: Agent-internal statistics of the K-Humor controller. The table summarizes generation length, controller action distribution, reward signals, and changes in humor-related scores during iterative refinement.

Criterion	Pearson r	Spearman ρ
Funniness	**0.83**	**0.92**
Surprise	**0.86**	**0.88**
Clarity	-0.29	-0.48
Cultural Appropriateness	-0.22	-0.21
Fluency	-0.12	-0.17
Originality	**0.76**	**0.76**
Coherence	0.47	0.39
Overall	**0.76**	**0.86**

Table 6: Correlation between mean human ratings and GPT-5 Likert scores over nine aligned jokes.

5.3 Human-LLM Agreement

Human-LLM agreement. Table 6 reports item-level correlations between mean human ratings and GPT-5 scores over the nine aligned jokes. Strong positive correlations are observed for funniness ($r = 0.83$, $\rho = 0.92$), surprise ($r = 0.86$, $\rho = 0.88$), originality ($r = 0.76$, $\rho = 0.76$), and overall quality ($r = 0.76$, $\rho = 0.86$). In contrast, agreement is weak or negative for clarity, fluency, and the original cultural-appropriateness criterion, while coherence shows only moderate positive correlation. These results suggest that GPT-5 captures some relative variation in core humor-related dimensions, but does not consistently reproduce native-speaker judgments across all evaluation criteria. Given the small aligned subset ($N = 9$), we treat this analysis as indicative rather than as a definitive validation of LLM-based evaluation.

Policy	Funny	Surprise	Coherence	Overall
Fixed 2-Step	3.06	2.78	3.85	3.34
Coherence Threshold	3.06	2.58	3.77	3.26
Bucket-UCB	**3.59**	**2.92**	**3.98**	**3.84**

Table 7: Comparison of the proposed Bucket-UCB controller with non-adaptive heuristic policies using LLM-as-a-Judge Likert scores.

System	Funny	Surprise	Original	Coherence
w/o Audience	3.06	2.53	2.71	3.77
w/o Editor	2.93	2.44	2.76	3.49
w/o Comparison	3.05	2.50	2.73	3.71
Full K-Humor	**3.59**	**2.92**	**3.07**	**3.98**

Table 8: Fine-grained ablation of individual agents using LLM-as-a-Judge Likert scores.

5.4 Controller and Component Ablations

The controller and fine-grained agent ablations were evaluated in separate controlled runs, so the Full K-Humor reference scores in these tables may differ slightly from those in the main evaluation.

Controller comparison. Table 7 shows that Bucket-UCB outperforms both heuristic policies across all reported dimensions, with the largest gains in overall quality and funniness. These results support adaptive controller-based refinement over the evaluated fixed schedules.

Fine-grained agent ablation. We further examine the contribution of individual agents by removing the Audience, Editor, or Comparison Agent from the full pipeline while retaining all other components. Each variant is evaluated using the same LLM-as-a-Judge Likert protocol.

As shown in Table 8, removing any of the three agents reduces performance. The Audience Agent most visibly affects funniness and surprise, the Editor Agent affects coherence, and the Comparison Agent also contributes to originality. Overall, the agents provide complementary rather than redundant signals.

5.5 Agent-Internal Analysis

To understand the controller-guided pipeline, we analyze internal logs recorded during generation. Table 5 summarizes statistics.

Controller behavior and refinement effects. The controller uses 5.68 steps per joke on average, with setup expansion and revision accounting for 65.1% of actions. It typically continues refinement after agents indicate punchline readiness, suggesting that the controller does not immediately commit to a punchline. Across iterations, twist increases from 2.91 to 3.11 and coherence from 4.62 to 4.74, while punchline-writing actions receive higher rewards than setup actions (1.64 vs. 0.96). We treat these corpus-level statistics as descriptive diagnostics rather than formal significance tests.

System	Diversity		S-BLEU ↓	Cohe. ↑
	Dist-2 ↑	Dist-3 ↑		
Baseline	0.92	0.97	0.04	0.33
K-Humor	0.89	**0.98**	0.06	**0.61**
Human	**0.94**	**0.98**	**0.01**	0.44

Table 9: Automatic metrics results covering diversity, redundancy, and coherence.

5.6 Automatic Metrics

We report automatic statistics for diversity (Distinct-n), Self-BLEU, and setup–punchline coherence. Table 9 compares K-Humor, the baseline, and human-written jokes.

Human-written jokes achieve the highest Distinct-2 (0.94), while K-Humor and Human obtain the same Distinct-3 score (0.98). Among the model-based systems, the Baseline has slightly higher Distinct-2 than K-Humor (0.92 vs. 0.89). K-Humor has slightly lower Distinct-2 (0.89) but comparable Distinct-3 (0.98). Human jokes obtain the lowest Self-BLEU (0.01), whereas K-Humor achieves the strongest setup-punchline coherence (0.61).

These results illustrate the limits of lexical automatic metrics for humor evaluation: diversity favors loosely structured outputs, whereas K-Humor yields stronger semantic alignment.

6 Conclusion

We introduced **K-Humor**, a dynamic multi-agent system for Korean humor generation that combines iterative refinement, cultural feedback, and human-data-grounded retrieval. Across LLM-based, human, and internal evaluations, K-Humor improves over a single-prompt baseline, particularly in terms of clarity and fluency. These results support culturally grounded multi-agent generation for low-resource creative tasks. Beyond humor generation, our framework highlights the potential of dynamic multi-agent systems for culturally sensitive and creative language tasks. Future work may explore scaling the approach to broader domains and more diverse cultural settings.

Limitations

Despite these contributions, K-Humor has several limitations. First, humor is subjective and context-dependent. Our evaluation relied on native Korean speakers; however, it did not fully capture variations across demographics, such as age, region, or subcultural backgrounds.

Second, the HDG mechanism is constrained by a relatively small corpus of 180 human-written jokes from the *Choi Bul-am Series*, which represents a specific and relatively narrow Korean humor tradition. Its recurring characters, temporal context, and stylistic conventions may bias retrieved exemplars and consequently influence the style of generated jokes. The HDG mechanism itself is corpus-agnostic, however, and can incorporate broader and more contemporary Korean humor collections as they become available.

Also, the controller in our system is implemented as a lightweight contextual bandit rather than a fully trained reinforcement learning policy. While this design improves stability and sample efficiency, it may not capture long-term dependencies across refinement steps or guarantee globally optimal action sequences. Although we provide ablation studies and agent-internal analysis, fully disentangling the contribution of each agent remains challenging. The multi-agent system introduces complex interactions between components, making it difficult to attribute performance gains to specific modules or coordination strategies. More controlled experiments, such as isolating agent interactions or learning explicit coordination policies, would improve interpretability, and future work could investigate more expressive decision-making frameworks, such as reinforcement learning with longer horizons or planning-based approaches.

Another limitation concerns our evaluation of cultural appropriateness. The human evaluation used a single criterion that jointly considered cultural relevance, audience suitability, and the absence of inappropriate content, while the LLM evaluator likewise returned only a single cultural-appropriateness score. Because neither evaluation separately measured cultural grounding and safety, the resulting scores cannot disentangle these dimensions. Future evaluations should assess them separately with native-speaker judgments.

Finally, the dynamic multi-agent architecture, although effective, introduces additional computational costs. Iterative refinement requires multiple agent interactions per joke, making generation less efficient than static methods. Although our results show that this cost yields substantial quality gains, future research should investigate adaptive cost-aware orchestration to improve practicality.

Acknowledgments

This research was supported by the "Advanced GPU Utilization Support Program" funded by the Government of the Republic of Korea (Ministry of Science and ICT). This research was also supported by the MSIT (Ministry of Science and ICT), Korea, under the Graduate School of Virtual Convergence support program (RS-2023-00254129), the ICAN (ICT Challenge and Advanced Network of HRD) support program (RS-2023-00259497), the ITRC (Information Technology Research Center) program (RS-2024-00436936), and AI Star Fellowship Support Program (Ulsan National Institute of Science and Technology, Inter-University; RS-2026-25616928), supervised by the IITP (Institute of Information & communications Technology Planning & Evaluation).

References

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Kim Binsted. 1996. *Machine humour: An implemented model of puns*. Ph.D. thesis, University of Edinburgh.

Jing Chen, Xinyu Zhu, Cheng Yang, Chufan Shi, Yadong Xi, Yuxiang Zhang, Junjie Wang, Jiashu Pu, Tian Feng, Yujiu Yang, and Rongsheng Zhang. 2024a. HoLLMwood: Unleashing the creativity of large language models in screenwriting via role playing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8075–8121, Miami, Florida, USA. Association for Computational Linguistics.

Yang Chen, Chong Yang, Tu Hu, Xinhao Chen, Man Lan, Li Cai, Xinlin Zhuang, Xuan Lin, Xin Lu, and Aimin Zhou. 2024b. Are U a joke master? pun generation via multi-stage curriculum learning towards a

humor LLM. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 878–890, Bangkok, Thailand. Association for Computational Linguistics.

Yuyan Chen, Yichen Yuan, Panjun Liu, Dayiheng Liu, Qinghao Guan, Mengfei Guo, Haiming Peng, Bang Liu, Zhixu Li, and Yanghua Xiao. 2024c. Talk funny! a large-scale humor response dataset with chain-of-humor interpretation. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218, Seattle, United States. Association for Computational Linguistics.

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Bradley Efron and Robert Tibshirani. 1995. An introduction to the bootstrap.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1):65–170.

He He, Nanyun Peng, and Percy Liang. 2019. Pun generation with surprise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1734–1744, Minneapolis, Minnesota. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3. *Preprint*, arXiv:2203.07785.

Sean Kim and Lydia B. Chilton. 2025. Ai humor generation: Cognitive, social and creative skills for effective humor. *Preprint*, arXiv:2502.07981.

Yumin Kim, Heejae Suh, Mingi Kim, Dongyeon Won, and Hwanhee Lee. 2024. KoCoSa: Korean context-aware sarcasm detection dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9890–9904, Torino, Italia. ELRA and ICCL.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott Hale. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 105236–105344. Curran Associates, Inc.

Hyunwoong Ko, Kichang Yang, Minho Ryu, Taekyoon Choi, Seungmu Yang, Jiwung Hyun, Sungho Park, and Kyubyong Park. 2023. A technical report for polyglot-ko: Open-source large-scale korean language models. *Preprint*, arXiv:2306.02254.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Nayoung Kwon and Patrick Sturt. 2024. When social hierarchy matters grammatically: Investigation of the processing of honorifics in korean. *Cognition*, 251:105912.

Yaobo Liang, Chenfei Wu, Ting Song, Wenshan Wu, Yan Xia, Yu Liu, Yang Ou, Shuai Lu, Lei Ji, Shaoguang Mao, Yun Wang, Linjun Shou, Ming Gong, and Nan Duan. 2024. Taskmatrix.ai: Completing tasks by connecting foundation models with millions of apis. *Intelligent Computing*, 3:0063.

Yi-Cheng Lin, Kang-Chieh Chen, Zhe-Yan Li, Tzu-Heng Wu, Tzu-Hsuan Wu, Kuan-Yu Chen, Hung-yi Lee, and Yun-Nung Chen. 2025. Creativity in LLM-based multi-agent systems: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27584–27607, Suzhou, China. Association for Computational Linguistics.

Chen Cecilia Liu, Hiba Arnaout, Nils Kovačić, Dana Atzil-Slonim, and Iryna Gurevych. 2026. Tailored emotional LLM-supporter: Enhancing cultural sensitivity. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 535–574, Rabat, Morocco. Association for Computational Linguistics.

Zhaoming Liu. 2023. Cultural bias in large language models: A comprehensive analysis and mitigation strategies. *Journal of Transcultural Communication*, 3(2):224–244.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2024. A dynamic LLM-powered agent network for task-oriented agent collaboration. In *First Conference on Language Modeling*.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.

Fuli Luo, Shunyao Li, Pengcheng Yang, Lei Li, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. Pun-GAN: Generative adversarial network for pun generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3388–3393, Hong Kong, China. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Anirudh Mittal, Yufei Tian, and Nanyun Peng. 2022. AmbiPun: Generating humorous puns with ambiguous context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1053–1062, Seattle, United States. Association for Computational Linguistics.

Akash Kumar Mohankumar and Mitesh Khapra. 2022. Active evaluation: Efficient NLG evaluation with few pairwise comparisons. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8761–8781, Dublin, Ireland. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Alipio Jorge, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, and 8 others. 2023. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.

Soyoung Oh, Jisu Kim, Seungpeel Lee, and Eunil Park. 2021. Jujeop: Korean puns for k-pop stars on social media. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 170–177, Online. Association for Computational Linguistics.

Patricia Phalen and Julia Osellame. 2012. Writing hollywood: Rooms with a point of view. *Journal of Broadcasting & Electronic Media*, 56(1):3–20.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John

Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 57 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 11479–11505. Curran Associates, Inc.

Jie Ruan, Xiao Pu, Mingqi Gao, Xiaojun Wan, and Yuesheng Zhu. 2024. Better than random: reliable nlg human evaluation with constrained active sampling. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Sudeep Salgia. 2023. Provably and practically efficient neural contextual bandits. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29800–29844. PMLR.

Qicai Shi, Feng Xiao, Douglas Pickard, Inga Chen, and Liang Chen. 2023. Deep neural network with linucb: A contextual bandit approach for personalized recommendation. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 778–782, New York, NY, USA. Association for Computing Machinery.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Stayci Taylor and Radha O'Meara. 2025. Power dynamics, inclusion, trust and leadership in tv script development. *Studies in Australasian Cinema*, 18(1):19–33.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Alexey Tikhonov and Pavel Shtykovskiy. 2024. Humor mechanics: Advancing humor generation with multistep reasoning. In *Proceedings of the 15th International Conference on Computational Creativity, ICCC 2024, Jönköping, Sweden, June 17-21, 2024*, pages 31–41. Association for Computational Creativity (ACC).

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

Han Wang, Yilin Zhao, Dian Li, Xiaohan Wang, sinbadliu, Xuguang Lan, and Hui Wang. 2025a. Innovative thinking, infinite humor: Humor research of large language models through structured thought leaps. In *International Conference on Learning Representations*, volume 2025, pages 42061–42085.

Qian Wang, Tianyu Wang, Zhenheng Tang, Qinbin Li, Nuo Chen, Jingsheng Liang, and Bingsheng He. 2025b. All it takes is one prompt: An autonomous LLM-MA system. In *ICLR 2025 Workshop on Foundation Models in the Wild*.

Edwin B. Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. 2018. A neural approach to pun generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1660, Melbourne, Australia. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Human Evaluation Survey

We present screenshots of the English prompts in the survey form used for human evaluation. The survey was deployed using Google Forms. One out of every six participants was randomly selected to receive a Starbucks coupon. Figure 4 shows the general interface, while Figures 5 and 6 illustrate the pairwise comparison and Likert rating questions, respectively.

B Multi-Agent Prompts (Full Pipeline)

This section provides the complete set of prompts used in the **K-Humor** multi-agent framework. Each agent including the Cultural Agent, Wordplay Agent, Setup Agent, Punchline Agent, Audience Agent, Editor, and Comparison Agent was initialized with distinct role-specific instructions. The prompts are presented in Korean with English glosses for clarity. Table 14 and its continuation provide the complete prompt set.

C Baseline (Single-Prompt) Prompts

For comparison, we also implemented a **single-prompt baseline**, in which the model directly generated a setup–punchline pair without iterative refinement, cultural grounding, or human-data guidance. The system prompt frames the model as a Korean comedy writer, and the user prompt specifies the audience and context. This approach allows us to isolate the contribution of dynamic multi-agent collaboration. Table 16 lists the exact baseline system and user prompts together with brief glosses.

D LLM-as-a-Judge Prompts

For reproducibility, we provide the exact prompts used for the LLM-as-a-Judge evaluation. We use `gpt-5` as the evaluator. The pairwise evaluator compares two jokes and returns an overall preference, while the Likert evaluator assigns scores to seven criteria and an overall quality score.

D.1 Pairwise Evaluation Prompt

For pairwise evaluation, the order of Joke A and Joke B is reversed and evaluated twice to reduce positional bias.

System prompt.

```
You are an expert Korean comedy critic and
↪  rater.
Compare two Korean jokes.
Do not consider the length of the jokes. Do
↪  not necessarily give longer jokes the
↪  better score.
Return STRICT JSON with keys:
 - overall_preference: one of {"A","B","Tie"}
 - rationale: short explanation (<=50 Korean
 ↪  words)
```

User prompt.

```
Evaluate the two jokes using the rubric. Be
↪  concise.

Joke A:
{joke_a}

Joke B:
{joke_b}

Output JSON ONLY.
```

D.2 Likert Evaluation Prompt

For Likert evaluation, each joke is independently evaluated on seven criteria, funniness, surprise, clarity, cultural appropriateness, fluency, originality, and setup-punchline coherence together with an overall score.

System prompt.

```
You are an expert Korean comedy critic and
↪  rater.
Rate a single Korean joke on seven criteria.
Return STRICT JSON with keys:
 - scores: object with fields funniness,
 ↪  surprise, clarity,
 ↪  cultural_appropriateness, fluency,
 ↪  originality, setup_punchline_coherence
 ↪  each ∈ {1,2,3,4,5}
 - overall: integer ∈ {1,2,3,4,5}
 - rationale: short explanation (<=50 Korean
 ↪  words)
```

User prompt.

```
Rate the following joke on the seven criteria
↪  and an overall score.

{joke}

Output JSON ONLY.
```

E Further on the Controller

This section provides additional implementation details of the controller that are not included in the main text.

Action exploration. The controller follows a bucketed UCB policy. At the beginning of each bucket, actions are forced to be explored once before the standard UCB scoring is applied. This ensures that each refinement operation is observed at

least once before the controller begins exploitation.

Reward update. For each bucket and action pair, the controller maintains the number of selections and the running mean reward. After an action is executed, the reward is used to update the running mean using an incremental update rule:

$$\mu_{t+1} = \mu_t + \frac{r_t - \mu_t}{n_t},$$

where r_t is the observed reward and n_t is the number of times the action has been selected in the bucket. For the reward computation, we set the weights to $\alpha = 1.0$ and $\beta = 1.0$ for improvements in twist and coherence, respectively, and $\gamma = 0.5$ for the semantic repetition penalty. Agent feedback scores are incorporated with weights $w_a = 0.5$, $w_e = 0.5$, and $w_c = 0.2$ for the Audience, Editor, and Cultural agents.

Context discretization. The contextual bandit operates over a small set of discretized states defined by humor signals. The twist score is divided into three ranges (low, medium, high) and repetition into two ranges (low, high), yielding six contextual buckets. Each bucket maintains its own UCB bandit statistics.

Repetition signal. To discourage repeated joke patterns, we compute a semantic repetition penalty based on cosine similarity between the current draft and previously generated jokes. Sentence embeddings are obtained using a Korean sentence encoder, and the maximum similarity with prior outputs is used as the repetition score.

Stopping criterion. Although the controller has a fixed maximum budget of eight refinement steps, generation may terminate earlier. If the reward fails to improve for several consecutive actions, the controller triggers an early stop. This heuristic prevents unnecessary expansions when the joke structure has stabilized.

Practical constraints. To maintain narrative coherence, setup rewriting is restricted after the first refinement step, and punchline generation is executed only once per trajectory. These constraints ensure that the refinement process primarily expands and improves the setup before committing to a final punchline.

F Additional Retrieval Experiments for HDG

To evaluate the impact of retrieval representations in the Human-Data-Grounded (HDG) module, we conducted additional experiments comparing three retrieval strategies:

- **TF-IDF**: lexical similarity using sparse term vectors,
- **Ko-sRoBERTa**[3] sentence embeddings,
- **KoSimCSE** sentence embeddings.

For each method, the system retrieved the top-k most similar setups from the human joke corpus ($k = 4$) and used them as few-shot exemplars for punchline generation. We generated 60 jokes for each retrieval configuration and evaluated them using the same LLM-as-a-Judge protocol described in Section 4.4. The evaluator (GPT-5) rated each joke on seven criteria and an overall score using a 1-5 Likert scale.

Table 10 summarizes the results. Overall, the three retrieval strategies produce similar performance, suggesting that HDG is relatively robust to the choice of embedding representation. However, KoSimCSE achieves the highest overall score and slightly improves fluency and setup-punchline coherence.

Based on these results, we adopt **KoSimCSE embeddings** as the default retrieval representation in the main experiments.

G Significance Tests

To verify whether observed evaluation differences are statistically meaningful, we conduct significance tests for both pairwise preference and Likert-scale ratings. The same statistical procedures are applied to both the human evaluation and the LLM-as-a-judge evaluation.

G.1 Statistical testing procedure.

For pairwise preference comparisons, each evaluation produces a binary outcome indicating which system is preferred. Under the null hypothesis that both systems are equally preferred, the probability of a win is $p = 0.5$. We therefore apply the **exact binomial test** to determine whether the observed win rate significantly differs from chance. We also

[3] https://huggingface.co/jhgan/ko-sroberta-multitask

Retrieval Method	Funny ↑	Surprise ↑	Clarity ↑	Cultural ↑	Fluency ↑	Original ↑	Coherence ↑	Overall ↑
TF-IDF	**3.59**	**2.94**	4.86	**4.98**	4.94	3.08	4.04	3.77
Ko-sRoBERTa	**3.59**	2.91	**4.92**	4.97	4.93	**3.17**	**4.05**	3.77
KoSimCSE (ours)	3.58	2.93	**4.92**	**4.98**	**4.95**	3.12	**4.05**	**3.82**

Table 10: Comparison of retrieval representations used in the HDG module. Scores are averaged Likert ratings (1-5) from GPT-5 evaluations over 60 generated jokes.

report Wilson score confidence intervals for the win rate, which provide reliable coverage for binomial proportions (Wilson, 1927).

For Likert-scale ratings, we compare system score distributions using two complementary non-parametric methods. First, we apply the **Mann-Whitney U test** (Mann and Whitney, 1947), which evaluates whether two independent samples originate from the same distribution without assuming normality. This test is appropriate for ordinal Likert ratings. Second, we estimate **bootstrap confidence intervals** for the mean score difference using 10,000 resamples (Efron and Tibshirani, 1995). Bootstrapping provides a distribution-free estimate of uncertainty and is robust for moderate sample sizes.

All statistical tests are two-sided with significance level $\alpha = 0.05$.

G.2 Pairwise Preference Tests

Table 11 summarizes the statistical tests for pairwise preference evaluations from both human annotators and the LLM-as-a-judge evaluation. Each comparison is evaluated using the exact binomial test under the null hypothesis that both systems are equally preferred ($p = 0.5$). Wilson score confidence intervals are reported for the estimated win rate.

The results show that **K-Humor is significantly preferred over the baseline** in both human and LLM-based evaluations. In human evaluation, K-Humor wins 64.5% of comparisons against the baseline ($p = 5.0 \times 10^{-6}$), while the LLM judge shows a similar trend with a 70.8% win rate ($p = 0.0055$).

When compared with human-written jokes, however, the advantage disappears: both human annotators and the LLM judge do not find statistically significant differences between K-Humor and human jokes ($p = 0.060$ and $p = 0.119$, respectively). This suggests that although the proposed system clearly improves over the baseline, the gap with human-authored humor remains non-negligible.

G.3 Likert Rating Tests

Table 12 reports statistical significance tests for the LLM-as-a-judge Likert ratings, and Table 13 presents the corresponding results for the human evaluation. Mean differences are computed as (K-Humor minus comparison system). Confidence intervals are estimated using bootstrap resampling, and statistical significance is assessed using the Mann-Whitney U test.

For the LLM-as-a-judge evaluation, K-Humor receives significantly higher scores than the baseline across multiple dimensions, including overall rating, funniness, clarity, fluency, and originality. These results indicate that the proposed multi-agent framework consistently improves both structural quality and perceived humor compared with a single-pass baseline.

When compared with human-written jokes, however, the results reveal a mixed pattern. Human jokes significantly outperform K-Humor in dimensions related to humor creativity, such as funniness, surprise, and originality. In contrast, K-Humor achieves higher scores in structural aspects such as clarity, cultural appropriateness, and fluency. This suggests that the proposed system improves several structural and audience-oriented dimensions, although the original cultural appropriateness criterion does not isolate cultural grounding from safety-related considerations. The system still struggles to match human creativity and comedic surprise.

H Dataset Preprocessing

For preprocessing, we remove trailing metadata (text after the "※" marker), split 'setup'/'punchline' primarily at the delimiter "`Answer :`" and otherwise by treating the last line as the 'punchline' and the preceding lines as the 'setup'. We strip bracketed reference tags (`[number]`), drop leading numeric enumerations in headings, and filter out very short items (setup length > 10 chars, punchline > 5 chars).

Evaluator	Comparison	Wins	Losses	Non-Ties	Win Rate	p-value
Human	K-Humor vs Baseline	162	89	251	**0.645**	5.0×10^{-6}
Human	K-Humor vs Human	112	143	255	0.439	0.060
LLM Judge	K-Humor vs Baseline	34	14	48	**0.708**	0.0055
LLM Judge	K-Humor vs Human	31	19	50	0.620	0.119

Table 11: Significance tests for pairwise preference evaluations. Win rates are evaluated using the exact binomial test with Wilson score confidence intervals.

Comparison	Metric	Mean Diff.	95% Bootstrap CI	p-value
vs Baseline	Overall	+0.777	[0.661, 0.890]	6.21×10^{-16}
	Funniness	+0.720	[0.593, 0.852]	1.82×10^{-14}
	Surprise	+0.231	[0.099, 0.366]	7.03×10^{-4}
	Clarity	+0.551	[0.325, 0.792]	1.81×10^{-6}
	Cultural	+0.107	[0.012, 0.216]	0.101
	Fluency	+0.405	[0.240, 0.582]	7.24×10^{-8}
	Originality	+0.503	[0.343, 0.662]	8.09×10^{-9}
	Coherence	+0.118	[-0.069, 0.317]	0.577
vs Human	Overall	-0.489	[-0.611, -0.378]	1.47×10^{-12}
	Funniness	-0.550	[-0.667, -0.433]	3.42×10^{-14}
	Surprise	-1.278	[-1.406, -1.156]	3.74×10^{-22}
	Clarity	+0.278	[0.161, 0.389]	2.71×10^{-6}
	Cultural	+0.472	[0.356, 0.589]	8.09×10^{-10}
	Fluency	+0.217	[0.111, 0.322]	2.80×10^{-5}
	Originality	-0.889	[-0.972, -0.794]	6.07×10^{-22}
	Coherence	-0.833	[-0.944, -0.722]	9.10×10^{-19}

Table 12: Significance tests for LLM-as-a-judge Likert ratings. Mean differences are computed as (K-Humor minus comparison system). Confidence intervals are estimated via bootstrap resampling, and statistical significance is assessed using the Mann-Whitney U test.

I Qualitative Examples

We provide representative examples to illustrate two qualitative patterns observed in the evaluation: the remaining gap between K-Humor and human-written humor, and differences in cultural specificity when the Cultural Agent is removed.

K-Humor vs. human-written humor. A representative K-Humor output develops a familiar Korean holiday scenario in which relatives repeatedly ask about marriage. The joke gradually extends this premise through references to family group chats, marriage-registration forms, and the family registry:

> **K-Humor:** "명절 때마다 친척들이 결혼 언제 하냐고 물어보는데, 내가 대답을 안 하면 다음 명절에 또 업데이트를 요구하더라고요. 이거 자동 갱신 서비스인가요? …아, 그냥 족보에 '미정' 도장 하나 파서 미리 꽉 찍어둘까 봐요."
>
> *(Every holiday, my relatives ask when I am getting married, and if I do not answer, they ask for another update the next holiday. Is this some kind of automatic renewal service? …At this point, I might as well stamp "undecided" next to my name in the family registry.)*

In contrast, a human-written example reaches its punchline through a shorter lexical incongruity:

> **Human-written:** "다음 문제로 민수기의 저자 문제가 나오자 최불암은 자신이 있다는 듯이 정답을 외쳤고 소리쳤다. '민수기는 민수가 썼습니다!!!' 최민수: ?"
>
> *(When asked who wrote the Book of Numbers (*Minsu-gi *in Korean), Choi Bul-am confidently answers, "Minsu wrote it!" Choi Minsu: ?)*

The K-Humor example maintains a coherent premise through incremental exaggeration, whereas the human-written example achieves the humorous turn more abruptly through a compact

Comparison	Metric	Mean Diff.	95% Bootstrap CI	p-value
vs Baseline	Clarity	+0.417	[-0.072, 0.900]	0.137
	Cultural	+0.556	[0.100, 1.011]	**0.028**
	Fluency	+0.406	[-0.011, 0.811]	0.060
	Funniness	+0.178	[-0.289, 0.644]	0.489
	Originality	+0.161	[-0.289, 0.611]	0.471
	Overall	+0.406	[0.022, 0.778]	**0.050**
	Coherence	+0.194	[-0.222, 0.600]	0.399
	Surprise	+0.150	[-0.272, 0.567]	0.555
vs Human	Clarity	-0.489	[-0.922, -0.056]	0.052
	Cultural	+0.300	[-0.167, 0.756]	0.185
	Fluency	-0.378	[-0.789, 0.033]	0.116
	Funniness	-0.544	[-0.989, -0.089]	**0.033**
	Originality	-0.389	[-0.800, 0.022]	0.100
	Overall	-0.433	[-0.822, -0.056]	**0.048**
	Coherence	-0.444	[-0.878, -0.011]	0.067
	Surprise	-0.833	[-1.244, -0.422]	3.7×10^{-4}

Table 13: Statistical significance tests for human-evaluated Likert ratings. Mean differences are computed as (K-Humor minus comparison system). Confidence intervals are obtained using bootstrap resampling and p-values are derived from the Mann-Whitney U test.

semantic reinterpretation. This illustrates a remaining challenge for K-Humor: generating punchlines with the brevity and sharp incongruity often found in human-written jokes.

Effect of the Cultural Agent. Representative outputs from the full and No-Culture conditions also illustrate a difference in cultural specificity. The full system develops a joke around repeated marriage questions during Korean holidays, further grounding the setup through references to *tteokguk* and holiday traditions:

> **Full K-Humor:** "이쯤 되면 결혼 질문이 아니라 명절 전통 의식 같아요-떡국 먹기 전에 '결혼 언제 하냐' 한 번 듣고 시작하는 거죠. ... 요즘은 결혼 허가증보다 명절 생존증명서가 더 필요할 것 같아요."
>
> *(At this point, asking about marriage feels less like a question and more like a holiday ritual-before eating* tteokguk*, you first have to hear "When are you getting married?" ... These days, I feel like I need a holiday survival certificate more than a marriage license.)*

A representative No-Culture output instead develops a more generic everyday scenario around an AI-enabled rice cooker:

> **w/o Cultural Agent:** "요즘엔 AI가 밥도 해준다는데, 혹시 밥솥이 저보다 똑똑해지는 날이 올까요? ... 밥솥은 밥 되면 '띵동!' 하고 알람만 울리는데, 저는 밥 태우면 엄마가 직접 출동해서 '띵동! 네 인생에 경고장 발급이야!' 하시거든요."
>
> *(These days even AI can cook-will the rice cooker eventually become smarter than me? ... The rice cooker just goes "ding-dong" when the rice is ready, but when I burn the rice, my mother shows up herself and says, "Ding-dong! A warning has been issued for your life!")*

Both outputs remain fluent and coherent, but the full-system example is grounded more explicitly in a culturally specific Korean holiday and family context. We therefore interpret the Cultural Agent as primarily contributing to cultural specificity rather than surface fluency.

J Qualitative Failure Analysis

Although K-Humor generally improves several quality dimensions over the baseline, the system still exhibits illustrative failure modes. Below we present representative examples from the generated outputs.

1. Over-extended Setup (Weak Punchline). KR:

> 카페에서 자리 맡아놓고 화장실 갔다 오면, 왜 꼭 내 자리에 누가 앉아있는 것 같을까요?
> 심지어 내 노트북이랑 가방까지 있는데도, 꼭 그 순간만 아무 존재감이 없는 사람처럼 느껴져요.

그래서 자리 맡아놓은 내 물건들이 혹시 나보다 더 인기 많은 건 아닐까, 괜히 질투가 나더라고요.
...
결국엔 카페 직원이 저보다 제 노트북이랑 가방을 더 자주 보는 것 같아서, 다음엔 제 사진도 하나 올려놓고 가야 할까 고민돼요.

EN:

When I leave my seat at a café and come back from the restroom, it always feels like someone might have taken my seat.
Even though my laptop and bag are still there, somehow I feel like the only invisible person in the room.
Sometimes I even wonder if my belongings are more popular than I am.
...
At this point I'm thinking of leaving a photo of myself on the table as well.

Issue: The setup grows progressively longer but the punchline introduces only a mild escalation, resulting in a weak comedic payoff. In this example, repeated setup expansion dilutes the eventual comedic payoff.

2. Literal or Predictable Punchlines. KR:

회식 끝나고 카카오톡 단체방에 올라오는 "오늘 즐거웠어요 " 이 말, 진짜 몇 명이나 진심일까요?
...
솔직히, 진짜 기대되는 건 회식이 아니라 칼퇴 아니었나요.

EN:

After a company dinner, people always post "That was fun today " in the group chat.
But honestly, the thing we are really looking forward to is not the dinner—it is leaving work early.

Issue: The punchline simply restates the implied sentiment in the setup without introducing a surprising twist.

3. Redundant Humor Templates. KR examples:

아메리카노 없으면 사무실 정상까지 못 올라가요.
커피 없으면 회사 정상까지 못 올라요.

EN:

Without an Americano I cannot climb to the summit of the office.
Without coffee I cannot reach the top of the company.

Issue: Multiple generations reuse the same structural metaphor ("coffee + summit"), indicating limited structural diversity.

4. Mild or Under-escalated Humor. KR:

명절 때마다 친척들이 결혼 언제 하냐고 물어보는데,
이쯤 되면 족보에 그냥 '미정' 도장 하나 파서 찍어둘까 봐요.

EN:

Every holiday my relatives ask when I will get married.
At this point I might just stamp "TBD" in the family registry.

Issue: The premise is relatable but the punchline produces only mild exaggeration rather than a strong comedic reversal.

Summary. Across these examples we observe four common limitations:

- overly long setups that dilute punchline impact,
- literal punchlines that restate the premise,
- repeated structural templates across generations,
- and under-escalated humor that lacks a strong twist.

These findings suggest that future work should improve punchline selection and introduce stronger diversity constraints during generation.

K Generated Samples

Table 17 provides representative examples of jokes generated by our system (**K-Humor**), the **single-prompt baseline**, and a **human-written** joke from our reference corpus. For clarity, we include both the original Korean text and its English translation. The translations are not intended to perfectly preserve the humor but to provide non-Korean readers with an approximate sense of the content and style.

Agent	Role	Prompt (Korean; English gloss in parentheses)
Cultural Agent	System	*"당신은 한국 유머 문화에 정통한 전문가입니다... 한국의 사회적 트렌드, 직장 문화, 가족 관계, 미디어(K-pop, 드라마 등), 또는 일상적인 상황 등 다양한 측면을 고려해주세요."* (*You are an expert in Korean humor culture; propose culturally relevant topics and review jokes with attention to social trends, workplace culture, family relations, media, and everyday situations.*)
	User	*"농담을 시작할 문화적으로 적합한 주제 세 가지를 JSON 배열 형식으로 제안해주세요..."* (*Suggest three culturally appropriate joke topics in JSON format.*)
Cultural Agent 2	User	*"아래의 농담 설정을 읽고 문화적으로 어색하거나 공감이 깨지는 지점을 찾아... 문화적 적절성, 공감대, 유머 스타일, 표현의 자연스러움을 분석..."* (*Review the current setup for cultural appropriateness, relatability, humor style, and naturalness of expression, then return a structured JSON verdict and feedback.*)
Wordplay Agent	System	*"당신은 창의적인 한국어 말장난 전문가입니다... 최소 3개의 언어유희 아이디어를 생성하고, 각각을 '유머 깊이', '명확성', '자연스러움' 기준에 따라 1-10점으로 평가하세요..."* (*You are a creative Korean wordplay expert; generate at least three pun ideas, score them for humor depth, clarity, and naturalness, and pass forward the best one.*)
	User	*"문화적 주제를 기반으로... 짧은 일화를 2-3개 생성하세요. 그 일화 속에서 유머의 핵심이 될 수 있는 단어와 이중적 의미를 찾아내고..."* (*Generate 2-3 short anecdotes grounded in the topic, extract a promising wordplay candidate, and return its dual meanings in JSON.*)
Setup	System	*"당신은 실력 있는 한국 코미디 작가입니다... 간결하고(1-2문장) 일관성 있으며, 재미있는 톤을 유지하세요..."* (*You are a skilled Korean comedy writer; write a concise, coherent, conversational setup that creates curiosity for the punchline.*)
	User	*"제공된 언어유희 단어를 먼저 평가하세요... 점수가 6점 이상일 경우에만 해당 언어유희를 활용... 그렇지 않으면 문화적 주제를 바탕으로 자연스러운 농담 설정..."* (*Use the wordplay only when its score is at least 6; otherwise start from the cultural topic alone and return {setup_line, wordplay_word}.*)
Setup 2	User	*"농담 설정의 다음 줄을 작성... 청중의 감정적 피드백, 코미디언의 구조적 피드백, 문화적 피드백을 모두 고려... 상황을 명확히 하거나, 확장하거나, 전환..."* (*Write the next setup line by considering audience reactions, editorial feedback, and cultural feedback; clarify, extend, or transition the setup in JSON format.*)
Audience Agent	System	*"당신은 전개되는 한국 농담에 감정적, 인지적으로 반응하는 청중 시뮬레이션 에이전트입니다... 연령대, 맥락, 지역 또는 방언, 유머 선호도..."* (*You are a simulated Korean audience agent; infer or maintain an audience profile and evaluate the ongoing joke in terms of emotion, anticipation, confusion, and readiness for the punchline.*)
	User	*"지금까지의 농담 설정... JSON 한 개로 응답하세요... audience_profile, emotion, anticipation_level, confusion_level, overall_score, audience_verdict..."* (*Given the current setup, return a JSON response with the audience profile, emotion, numeric ratings, verdict, suggestion, and expected punchline style.*)

Table 14: Complete multi-agent prompts used in the full K-Humor pipeline. Korean prompts are shown with English glosses.

Agent	Role	Prompt (Korean; English gloss in parentheses)
Editor	System	*"당신은 한국 코미디를 전문으로 하는 전문 코미디언이자 스토리 편집자입니다... 논리적 진행, 맥락과의 정렬, 유머 리듬, 감정적 페이싱, 언어유희 평가..."* (*You are a Korean comedy editor; analyze logical progression, contextual fit, pacing, and the quality of the proposed wordplay.*)
	User	*"당신의 임무는 농담에 사용될 언어유희를 검토하고 피드백을 제공하는 것입니다... 유머 잠재력(1-10점)... 억지스러우면 과감히 낮은 점수..."* (*Evaluate the candidate wordplay for humor potential and return a wordplay rating in JSON.*)
Editor 2	User	*"전개 중인 농담을 검토하고 편집 피드백을 제공... 최소 3-4줄의 설정이 구축되기 전에는 'Punchline'을 제안하지 마세요... verdict, editing_suggestion, suggested_punchline_style..."* (*Review the evolving joke and decide whether it needs more setup or is ready for the punchline; include structural advice and an overall score in JSON.*)
Editor 3	User	*"완성된 농담을 검토하고 피드백을 제공... 농담을 어떻게 더 발전시킬 수 있을지 제안..."* (*Review the completed joke and provide final refinement suggestions in JSON.*)
Punchline	System	*"당신은 한국 코미디 펀치라인 작가입니다... 펀치라인은 짧고(1-2줄) 영리하며, 맥락적으로 설정과 연결되어야 합니다... 말장난, 스크립트 반전, 부조화-해결, 과장 또는 역전..."* (*You are a Korean punchline writer; produce a short, clever, setup-consistent punchline using principles such as wordplay, script opposition, incongruity-resolution, and reversal.*)
	User	*"제공된 언어유희 단어가 있다면... 전체 설정, 청중 반응, 코미디언 피드백, few-shot 예시를 참고... {punchline: ...}"* (*Generate a punchline using the current setup, optional wordplay, audience/editor feedback, and retrieved few-shot examples; return {punchline}.*)
Punchline 2	User	*"완성된 농담과 코미디 편집자의 피드백이 주어졌습니다... 수정된 펀치라인만이 아니라 전체 농담을 반환해야 합니다... {full_joke: ...}"* (*Revise the completed joke using editorial feedback and return the entire revised joke, not only the punchline.*)
Comparison Agent	System & User	*"당신은 한국 유머 평론가입니다... 유머 깊이, 명확성, 놀라움, 언어 자연스러움, 독창성, 일관성을 기준으로 두 농담을 평가... 전반적인 승자를 선택하세요."* (*Compare two jokes as a Korean humor critic, score them on humor, clarity, surprise, naturalness, originality, and consistency, and select a winner in JSON.*)

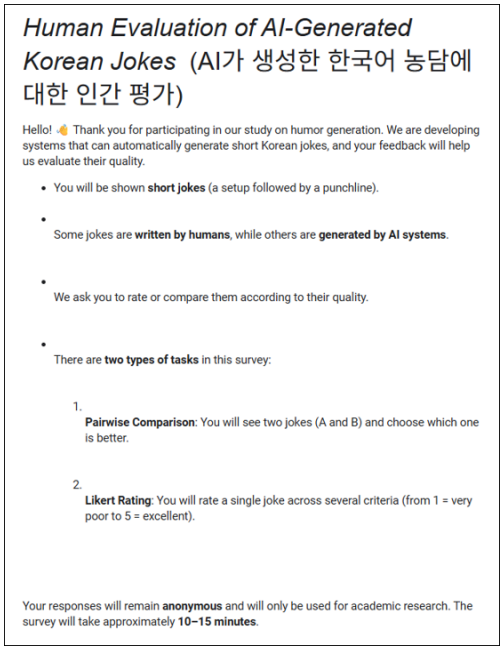
Table 15: Complete multi-agent prompts used in the full K-Humor pipeline (continued).

Role	Prompt (Korean; English gloss in parentheses)
System	*"You are a Korean comedy writer. Generate a single self-contained joke with a setup and a punchline. Return JSON with keys: setup, punchline."* (*Single-shot Korean joke generation; return JSON fields for setup and punchline.*)
User	*"Write a funny Korean joke for audience [audience] in context [context]. Return JSON only."* (*Generate a funny Korean joke for the given audience and context, and return JSON only.*)

Table 16: Baseline single-prompt prompts from the generation script.

Source	Korean (Original)	English (Translation)
K-Humor (Ours)	카페에서 자리 맡으려고 노트북, 책, 심지어 신발까지 올려두는 사람들 보면, 다음엔 뭘 올릴지 궁금하지 않으세요? 어느 순간부터는 자리 맡으려고 도시락통이나 가족사진까지 꺼내놓을까봐 걱정돼요. 이쯤 되면 자리 맡으려고 카페 사장님을 앉혀놓는 사람도 나올 것 같아요. 아니면, 자리 맡기 올림픽 개최되면 누가 뭘 올려야 1등 할지 궁금하잖아요. 결국엔 자리 맡으려고 자기 얼굴 3D 프린터로 뽑아서 두고 갈지 몰라요.	When you see people reserving seats at cafés by putting laptops, books, or even their shoes on the table, don't you start wondering what they'll put there next? At this rate, I'm worried that someday people will start leaving lunchboxes or even family photos just to hold their seats. Eventually someone might even try reserving a seat by placing the café owner there. If there were an Olympics for saving seats at cafés, you'd wonder what someone would have to put on the table to win the gold medal. Honestly, at this point I wouldn't be surprised if someone 3D-printed their own face and left it there to keep the seat.
Baseline (Single-Prompt)	내 친구가 나한테 "넌 왜 맨날 집에만 있냐"고 물어보길래 뭐라고 했게? "난 집돌이 아니라 홈보이야, 글로벌하게!"	My friend asked me, "Why are you always at home?" I said, "I'm not a homebody, I'm a homeboy—global style!"
Human-Written	최불암이 오락실에 가서 스트리트 파이터 2를 열심히 했다. 그리고 나서 이발소에 찾아갔다. "아저씨, 제 머리를 달심처럼 깎아주세요." 이발사는 솜씨좋게 달심처럼 머리를 깎아주었다. 머리를 깎고 나간 최불암. 잠시 후에 돌아와서 하는 말, 왜 팔이나 다리는 안늘어나는겨?	Choi Bul-am went to the arcade and played *Street Fighter II* all day. Then he went to the barbershop. "Sir, please cut my hair like Dhalsim." The barber skillfully gave him a Dhalsim-style haircut. After leaving, Choi Bul-am came back a little later and asked, "Hey, why don't my arms or legs stretch out?"

Table 17: Example jokes from our system (K-Humor), the single-prompt baseline, and the human-written corpus, shown with English translations.



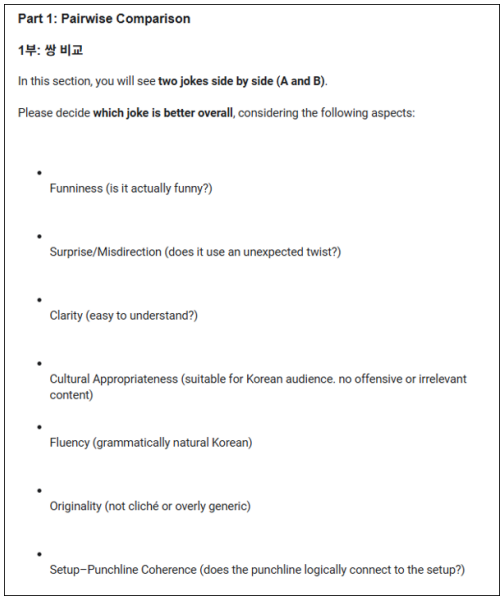
Human Evaluation of AI-Generated Korean Jokes (AI가 생성한 한국어 농담에 대한 인간 평가)

Hello! 👋 Thank you for participating in our study on humor generation. We are developing systems that can automatically generate short Korean jokes, and your feedback will help us evaluate their quality.

- You will be shown **short jokes** (a setup followed by a punchline).
- Some jokes are **written by humans**, while others are **generated by AI systems**.
- We ask you to rate or compare them according to their quality.
- There are **two types of tasks** in this survey:
  1. **Pairwise Comparison**: You will see two jokes (A and B) and choose which one is better.
  2. **Likert Rating**: You will rate a single joke across several criteria (from 1 = very poor to 5 = excellent).

Your responses will remain **anonymous** and will only be used for academic research. The survey will take approximately **10–15 minutes**.

Figure 4: Overview of the human evaluation survey form.



Part 1: Pairwise Comparison

1부: 쌍 비교

In this section, you will see **two jokes side by side (A and B)**.

Please decide **which joke is better overall**, considering the following aspects:

- Funniness (is it actually funny?)
- Surprise/Misdirection (does it use an unexpected twist?)
- Clarity (easy to understand?)
- Cultural Appropriateness (suitable for Korean audience. no offensive or irrelevant content)
- Fluency (grammatically natural Korean)
- Originality (not cliché or overly generic)
- Setup–Punchline Coherence (does the punchline logically connect to the setup?)

Figure 5: Example of a pairwise comparison question (K-Humor vs. baseline or human).

Part 2: Likert Scale Rating

2부: 리커트 척도 평가

In this section, you will rate **single jokes** on a scale of **1 (very poor) to 5 (excellent)** across several dimensions. Please be honest and consistent.

이 섹션에서는 개별 농담을 여러 기준에 따라 **1점(매우 나쁨)에서 5점(매우 좋음)**까지 평가하게 됩니다.

정직하고 일관되게 답변해 주시기 바랍니다.

Q21. Rate this joke 1 (very poor) to 5 (excellent):

회식 끝나고 집에 가는 길, 항상 생각해요. 왜 회식은 늘 '마지막 한 잔'이 세 번이나 나올까요?
심지어 그 '마지막 한 잔'이 끝나면, 누군가는 꼭 '진짜 마지막'이라면서 소주병을 다시 따요.
그렇게 집에 가려다 보면, 어느새 택시비보다 회식비가 더 아까워지는 거 있죠.
그래서 요즘은 회식 자리에서 '진짜 마지막'이 나오면, 슬슬 집에 갈 준비를 시작해요—근데 그때부터 또 안주가 새로 나와요.
결국 집에 도착하면, 회식이 끝난 게 아니라 내일 출근이 시작된 거더라고요.
그래서 요즘은 회식 자리에서 '마지막'이란 말 나오면, 그냥 조용히 물만 따라 마셔요—내일 아침에 해장국 먹을 체력이라도 남겨야 하니까요.
근데 진짜 회식의 마지막은, 계산서 돌 때 화장실 가는 내 달리기 실력이라는 걸 최근에 깨달았습니다.

	1	2	3	4	5
Funniness (how funny is it?)	◯	◯	◯	◯	◯
Surprise/Misdirection (does it have an unexpected twist?)	◯	◯	◯	◯	◯
Clarity (easy to understand?)	◯	◯	◯	◯	◯
Cultural Appropriateness (is it inoffensive and contextually appropriate?)	◯	◯	◯	◯	◯
Fluency (natural Korean expression?)	◯	◯	◯	◯	◯
Originality (is it creative, not cliché?)	◯	◯	◯	◯	◯
Setup–Punchline Coherence (does the punchline fit the setup logically?)	◯	◯	◯	◯	◯
Overall Quality	◯	◯	◯	◯	◯

Figure 6: Example of a Likert-scale rating question.