Navigating the Joke Space: Towards Automated Originality Assessment of AI-Generated Humor

Ori Amir[1,2,3], Huyen Dieu Ngo[1], Joe Toplyn[3,4], Kevin P. Hickerson[5]

[1]Fulbright University Vietnam [2]HaHator, LLC
[3]Semantic AI and Creativity Lab, East Texas A&M University
[4]Twenty Lane Media, LLC [5]California Institute of Technology

Correspondence: ori.amir@fulbright.edu.vn oriacadem@gmail.com

Abstract

This study validates automated, corpus-based methods for quantifying joke originality using "topic handles" — key nouns or noun phrases capturing a joke's script opposition and logical mechanism (per the General Theory of Verbal Humor). Using a reference corpus of one million jokes in English from Reddit, we compute Pointwise Mutual Information (PMI) in three variants (raw co-occurrence, semantic-cluster smoothing, and word-decomposition) and two embedding-based measures (handle-level conceptual distance and full-text corpus novelty via Sentence-BERT). We evaluate these measures on 400 LLM-generated jokes (200 each from GPT-4o and GPT-5.4) and 80 jokes from the Witscript-powered JEST benchmark, rated by three professional comedians for originality and funniness. Corpus novelty and concept distance between the most semantically distant handle pair both correlated significantly with human originality ratings (ρ = .37); PMI-based measures showed weaker but significant associations (ρ = .23–.25) on the most original handle pair. A Lasso-based composite of the three strongest predictors achieved ρ = .40 (cross-validated), capturing 82% of the theoretically predictable variance given inter-rater agreement. These results demonstrate that handle-based PMI and semantic novelty metrics offer practical, quantitative tools for assessing originality in AI-generated humor, advancing objective evaluation of computational creativity.

1 Introduction

Large Language Models (LLMs) can generate jokes on demand, producing humor that rivals the quality of professional comedy writers, as evidenced by audience reactions in comedy club settings (Toplyn and Amir, 2025). However, the originality of LLM-generated jokes remains uncertain (Jentzsch and Kersting, 2023; Veale, 2024). While traditional plagiarism detection can identify verbatim matches, LLMs excel at rephrasing, prompting critics to label them as plagiarism machines (Chomsky et al., 2023) or "stochastic parrots" (Bender et al., 2021). Unlike rule-based joke generation algorithms, which follow transparent processes (Binsted and Ritchie, 1997), LLM-based systems function as opaque "black boxes," obscuring the mechanisms behind their joke creation. Numerous studies investigating the comedic capabilities of LLMs often implicitly assume that the generated jokes are original. However, there is evidence that successful LLM humor is often not original (Jentzsch and Kersting, 2023; Veale, 2024).

To address this issue, Amir (2025) proposed characterizing the "Joke Space" — the set of all possible verbal jokes in English — as a framework for evaluating originality. Even when considering only the core elements of a joke, which make a joke truly unique, the space of all possible jokes is estimated to be at minimum 1 to 32 billion jokes, several orders of magnitude larger than the total number of jokes any LLM or human comedian could plausibly have been exposed to (estimated as 150 million for modern LLMs; Figure 1). This suggests that, in principle, LLMs could be prompted to generate entirely novel jokes. Despite this vast potential, however, both AI systems and human comedians frequently produce non-original jokes. This tendency can be attributed either to prior exposure to similar jokes, or to a focus on highly salient topics (Amir and Biederman, 2016; Amir et al., 2022; Shahaf et al., 2015) or emotionally charged content (Hempelmann and Ruch, 2005).

Determining whether any individual joke is original therefore requires identifying its "essence": the core elements that define its humor independent of surface wording. Drawing on the General Theory of Verbal Humor (GTVH; Attardo and Raskin, 1991), Amir (2025) defines a joke's essence as its Script Opposition (SO) and Logical Mechanism

Proceedings of the 2nd Workshop on Computational Humor (CHum 2026), pages 95–105

(LM), with other elements such as language and narrative strategy representing optimization rather than essence. At the implementation level, this essence can be approximated by a joke's topic handles — two key nouns or short noun phrases whose pairing results in the comedic effect (Toplyn, 2014), and the associative link connecting them. Drawing on the definition of originality as statistical infrequency (Dumas et al., 2021; Runco et al., 2024), a joke is more original when its handles co-occur more rarely in natural language generally and in joke corpora in particular. Despite this body of prior work, quantitative corpus-based originality assessment of individual AI-generated jokes remains understudied (Jentzsch and Kersting, 2023; Loakman et al., 2025; Veale, 2024). Most research on AI creativity has instead relied on general frameworks such as the Torrance Test or the Alternative Uses Test, which measure originality as statistical rarity relative to a reference population (Lamb et al., 2018; Runco et al., 2024). These approaches, however, are domain-general and may fail to capture the essential elements of a joke that determine whether two jokes are distinct or, instead, share a common premise based on a particular combination of script opposition and logical mechanism (Ruch et al., 1993; Cain et al., 2024).

Toplyn (2014) proposed a practical operationalization of this framework, originally designed for human comedy writers and later adapted for generative computational humor work (e.g., Toplyn, 2023). The approach involves extracting (usually two) handles from the joke topic and identifying unexpected or pseudo-logical associative links between them (see also Dean, 2000). The handles are typically the key nouns or verbs in the joke topic that are critical for establishing the script opposition and logical mechanism.

This work explores automated methods for assessing joke originality by integrating widely used PMI and semantic embedding techniques with a topic handle extraction framework. We validate these methods against originality judgments by professional comedians, and examine how the relationship between originality and funniness varies across joke sources and generation methods.

2 Methods

2.1 Data and Resources

As a reference corpus, we used the SocialGrep/one-million-reddit-jokes dataset, comprising one million jokes in English drawn from joke subreddits and covering a wide variety of styles (SocialGrep, 2021). For experimentation, 400 AI-generated jokes were produced using the OpenAI API: 200 by GPT-4o and 200 by GPT-5.4. Jokes were generated in batches of 10 using the following system prompt: "You are a professional comedian. You provide raw joke content without any headers or metadata." The user prompt randomly varied the comedic style across five categories (absurdist, satirical, dry wit, wordplay, and observational), with the style selected uniformly at random for each batch, to encourage output diversity. The full user prompt was:

> Generate n original and clever jokes in the style of 'style'. STRICT FORMATTING RULES: (1) Do NOT include any topic names, categories, or labels. (2) 'title' MUST be the setup or question only. (3) 'selftext' MUST be the punchline or answer only. (4) Do NOT use common clichés. Return ONLY a raw JSON array of objects.

All calls used a temperature of 1.0 to maximize output diversity. The model was instructed to return jokes as structured JSON objects with a title field (setup) and selftext field (punchline); the two fields were concatenated to form the full joke text. To prevent repetition, jokes were deduplicated based on normalized full-text content (lowercased, whitespace-collapsed) before inclusion; batches were re-requested as needed until 200 unique jokes per model were collected. An additional 80 items from the JEST benchmark (Toplyn and Amir, 2026) were used for validation. The JEST benchmark was generated using Witscript (Toplyn, 2023), a software system for automated joke writing: 60 items were produced by Witscript in response to short texts and selected to achieve approximately equal representation of funniness ratings of 1 (Somewhat Funny), 2 (Funny), and 3 (Very Funny). The remaining 20 items were generated by the OpenAI model GPT-5.1 to match the benchmark jokes in style and length but to contain no humor; each was assigned a funniness rating of 0 (Not Funny). While the benchmark's funniness categories were by design, the category assignments were confirmed by ratings from experienced professional comedy writers, ensuring that the stratification reflects genuine expert judgment rather than design alone. GPT-4o was selected as it is the most studied LLM in the context of computational humor

and was also the model underlying Witscript, the system used to generate the JEST benchmark. GPT-5.4, the most recent OpenAI model at the time of data collection, was included to assess whether newer models produce more original jokes.

2.2 Data Processing and Semantic Indexing

The title and body fields of each reference corpus entry were concatenated into a single text sequence. All jokes were lowercased and tokenized using a regular expression that retains word tokens of three or more characters; standard stop words (e.g., 'the,' 'and,' 'with') were removed. An inverted index was constructed mapping each token to the set of joke IDs in which it appears. To ensure statistical stability and reduce sparsity, only tokens with a corpus frequency of five or more were retained in the working vocabulary. All vocabulary items were encoded using all-mpnet-base-v2 for subsequent semantic clustering and comparison (Reimers and Gurevych, 2019).

All jokes in the dataset were encoded as dense vector representations using all-mpnet-base-v2. These embeddings were L2-normalized so that all vectors share a uniform magnitude, which allows cosine similarity to be computed efficiently via dot product. To enable fast retrieval of the most semantically similar jokes for any given input, we indexed the normalized embeddings using the FAISS library with an IndexFlatIP configuration (Johnson et al., 2019).

2.3 Topic Handle Extraction

Topic handles were defined as key nouns or short noun phrases representing the core reusable subject matter of a joke, grounded in Toplyn (2014) finding that jokes typically link two discrete concepts pseudo-logically. Handle extraction was performed by prompting GPT-5.4 (temperature 0.1) using the following system prompt:

> You are an expert comedy analyst. Extract the topic handles from a given joke. A topic handle is a key noun or short noun phrase representing the joke's core, reusable subject matter. Handles must be concrete and specific (e.g., 'dog', 'depressed penguin'). Do not include generic nouns (e.g., 'man', 'thing', 'person') or verbs/adjectives that only enable punchline logic (e.g., 'smell', 'think'). Return ONLY a JSON object: 'handles': ['handle1', 'handle2', ...] . All handles must be lowercase strings.

The output was stripped of whitespace and stop words. Multi-word handles were additionally decomposed into their component words (for example, 'depressed penguin' into 'depressed' and 'penguin'), yielding both compound and atomic handles for each joke.

Two examples illustrate the method. For the GPT-5.4 joke "I tried becoming more spontaneous. I've scheduled another attempt for Thursday," the extractor returned: 'spontaneity,' 'schedule,' 'Thursday' — capturing the core script opposition cleanly, with 'Thursday' as minor noise that strengthens the joke through specificity but is not necessary for its logic. For the JEST joke "Giraffes only sleep 30 minutes a day... takes 25 minutes to fluff the neck pillow," it returned: 'giraffes,' 'neck pillow,' 'neck,' 'pillow' — where 'neck pillow' and 'giraffe' are the appropriate key handles, with 'neck' and 'pillow' as the decomposed components.

It is worth noting that this extraction method differs from Toplyn's (2021; 2023) Witscript approach in two respects. First, whereas Toplyn extracts handles from the joke setup during the writing process, here extraction is performed on the full joke text after generation. Second, the prompt used here typically returned more than two handle candidates per joke. For the handle-based measures (PMI variants and handle-level conceptual distance), subsequent analysis therefore enumerated all pairs of handles and computed each measure for every pair, then selected the most original pair per joke. This is conceptually consistent with Witscript's algorithm, in which one criterion for selecting topic handles is maximizing the semantic distance between them (Toplyn, 2021). The corpus novelty measure, which operates on the full joke text rather than handle pairs, was not subject to this pairwise procedure. Our goal was to demonstrate that even this relatively basic extraction approach can capture meaningful originality signals, with methods leveraging deeper joke understanding, a skill that modern LLMs exhibit to a useful extent (Narad et al., 2025), as a natural direction for future work.

2.4 PMI-Based Familiarity Measures

Based on the hypothesis that jokes utilizing less common concept combinations are more original, we adopt Pointwise Mutual Information (PMI) to quantify the frequency of handle co-occurrence

in the reference corpus. For each pair of handles, a raw PMI score is computed from exact co-occurrence counts. Let D(h) denote the set of jokes containing handle h, and N = 1,000,000 be the total number of jokes. The raw PMI is defined as:

$$\mathrm{PMI}(h_1, h_2) = \log_2 \left(\frac{|D(h_1) \cap D(h_2)| \cdot N}{|D(h_1)| \cdot |D(h_2)|} \right) \quad (1)$$

Higher PMI values indicate more frequent co-occurrence, implying a more familiar conceptual pairing and lower originality. Because exact handle co-occurrence is often sparse (especially for specific or multi-word concepts), we introduce a cluster-based PMI variant. Vocabulary items are first grouped into 15 semantic clusters using K-means over pre-trained sentence embeddings. Each handle is mapped to its corresponding cluster, and PMI is computed over the union of joke sets associated with each cluster.

To further address sparsity in multi-word handles, a decomposed PMI score is computed as the mean of raw PMI across all pairs of component words. This preserves partial lexical associations when the full phrase is absent or too rare in the corpus. For each joke, all unordered handle pairs are enumerated and raw, cluster-based, and decomposed PMI scores are computed for each pair.

2.5 Embedding-Based Semantic Measures

Because PMI is insensitive to semantic meaning, an embedding-based method is used to quantify originality through conceptual distance and holistic similarity. Two distinct embedding-based measures were used. At the handle level, conceptual distance between two handles is computed as one minus their cosine similarity:

$$d(h_1, h_2) = 1 - \cos(\mathrm{emb}(h_1), \mathrm{emb}(h_2)) \quad (2)$$

A larger distance indicates a more semantically dissimilar and thus more novel conceptual pairing.

At the joke level, the full-text similarity score s_full is determined as the maximum cosine similarity between the joke and any joke in the reference corpus (of over one million jokes), retrieved via the pre-built FAISS index. A high s_full indicates that the joke is semantically redundant with existing material; a low score suggests greater originality. For analysis, this score is transformed to a corpus novelty measure as 1 - s_full, so that higher values consistently indicate greater originality.

2.6 Human Ratings

Three raters evaluated the jokes for originality and funniness. All three raters are professional comedians, each with at least a decade of experience performing and/or writing comedy in the United States. Each joke was rated on two 4-point scales (0–3). Originality was defined as how original the joke felt relative to the rater's own prior exposure to similar humor: 0 indicated an exact repeat (the rater had heard this specific joke before), 1 indicated low originality (not the exact joke, but something very similar, with the same setup and punchline), 2 indicated a familiar premise (the joke was new to the rater but used a recognizable formula or trope), and 3 indicated full originality (both the premise and execution felt fresh). Funniness was rated on a scale from 0 (not funny; likely to get no reaction) to 3 (very funny; likely to get a laugh), with intermediate values of 1 (almost funny; might get a smile) and 2 (funny; likely to get a chuckle). Raters were instructed that some texts might not be intended to be funny and to read each text carefully before rating. Each rater evaluated 80 jokes from each of the three sources: GPT-4o, GPT-5.4, and the JEST benchmark (i.e., 240 jokes per rater, 480 ratings per rater across both dimensions).

2.7 Statistical Analysis

Inter-rater agreement was assessed on the 80 human-rated jokes per source, as well as pooled across all 240 jokes. Fleiss' κ with quadratic weighting served as the primary measure of absolute rating agreement. Quadratic-weighted Fleiss' κ is the multi-rater generalization of Cohen's weighted κ, penalizing larger disagreements proportionally to the square of the distance between categories, which is appropriate for ordinal scales with more than two raters. Mean pairwise Spearman ρ (the average of all three pairwise rank correlations) served as a complementary measure of relative agreement, capturing the degree to which raters agreed on the rank ordering of jokes independently of differences in scale use. Bootstrapped 95% confidence intervals and p-values were derived from a single resampling pass (2,000 item-level resamples) using the recentering method.

Differences in mean originality and funniness ratings across the three joke sources were tested using Kruskal-Wallis H tests on item-level means, with post-hoc pairwise Mann-Whitney U tests and Holm-Bonferroni correction. Within each source,

the relationship between originality and funniness at the item level was quantified using Spearman ρ, and pairwise differences between the three resulting correlations were tested using both Fisher's z-transformation and a bootstrap test of the difference in ρ, with Holm correction across the three pairs.

To assess the validity of the proposed automated originality measures, Spearman rank-order correlations with human originality and funniness ratings were examined. Spearman's ρ was used throughout as all measures departed significantly from normality (Shapiro-Wilk W ranging from .48 for PMI Cluster Max to .91 for Corpus Novelty, all $p < .001$). All measures were oriented prior to analysis so that higher values consistently indicate greater originality: the full-text BERT similarity score was transformed as 1 - BERT_full to produce a corpus novelty score; the three PMI-based scores were negated (× -1) so that rarer concept pairings yield higher values; the handle-level BERT distance measure required no transformation as it already reflects conceptual distance. For the handle-pair measures (PMI variants and handle-level conceptual distance), scores were summarized across pairs by selecting the most original pair per joke — defined as the pair with the lowest raw PMI (greatest rarity) for PMI measures, and the pair with the greatest cosine distance for the conceptual distance measure. Where all handle pairs for a given joke were absent from the corpus (rendering PMI technically undefined), the PMI score was imputed to the theoretical maximum originality value, corresponding to the minimum possible co-occurrence frequency in a corpus of this size ($-\log_2(N) \approx 15.3$ after negation). Jokes for which no handle pairs could be formed (single-handle jokes) were excluded from PMI-based analyses. Correlation analyses were restricted to the subset of jokes for which all three human raters provided ratings (n = 240: 80 JEST benchmark, 80 GPT-4o, 80 GPT-5.4). Ninety-five percent confidence intervals for all correlations were computed via the Fisher z-transformation.

To assess whether combining measures improves prediction, a composite originality score was constructed using Lasso regression, a penalized regression method that automatically shrinks redundant predictor coefficients to zero. The regularization strength (alpha) was selected via 10-fold cross-validation. Predictive validity of the composite was evaluated using both 10-fold and leave-one-out cross-validation, with out-of-fold Spearman correlations serving as the performance metric. Jokes missing any measure (primarily single-handle jokes lacking pair-based scores) were excluded from the composite analysis (n = 235).

2.8 Implementation

Both the computation of the originality measures and the subsequent analyses were implemented in Python. All semantic operations were performed using the all-mpnet-base-v2 Sentence-Transformer model. All 400 LLM-generated jokes were produced with the OpenAI API.

3 Results

3.1 Inter-Rater Agreement

Pooled across all 240 jokes (80 per source), inter-rater agreement on originality was low. Quadratic-weighted Fleiss' κ was .187 (95% CI [.10, .26], $p < .001$) and mean pairwise Spearman ρ was .240 (95% CI [.15, .32], $p < .001$). Pairwise Spearman correlations ranged from $\rho = .083$ (95% CI [-.05, .21], $p = .209$) to $\rho = .372$ (95% CI [.25, .48], $p < .001$).

Agreement on funniness was substantially higher. Quadratic-weighted Fleiss' κ was .45 (95% CI [.36, .53], $p < .001$) and mean pairwise Spearman ρ was .477 (95% CI [.40, .56], $p < .001$). Pairwise Spearman correlations ranged from $\rho = .353$ (95% CI [.24, .47], $p < .001$) to $\rho = .648$ (95% CI [.55, .73], $p < .001$).

3.2 Originality and Funniness Ratings by Joke Source

Mean originality ratings differed significantly across the three sources (Kruskal-Wallis $H(2) = 45.80$, $p < .001$). The JEST benchmark jokes received the highest mean originality rating M = 2.48 (SD = 0.45), followed by GPT-4o jokes M = 2.09 (SD = 0.48) and GPT-5.4 jokes M = 1.92 (SD = 0.55). Post-hoc pairwise comparisons showed that all three sources differed significantly from one another after Holm correction (Benchmark vs. GPT-4o: $U = 4653.5$, $p < .001$; Benchmark vs. GPT-5.4: $U = 4980.0$, $p < .001$; GPT-4o vs. GPT-5.4: $U = 3857.5$, $p = .022$).

Mean funniness ratings did not differ significantly across sources (Kruskal-Wallis $H(2) = 4.74$, $p = .094$), with means of M = 1.50 (SD = 1.06) for Benchmark jokes, M = 1.24 (SD = 0.60) for GPT-4o jokes, and M = 1.21 (SD = 0.62) for GPT-5.4 jokes. Levene's test indicated significant het-

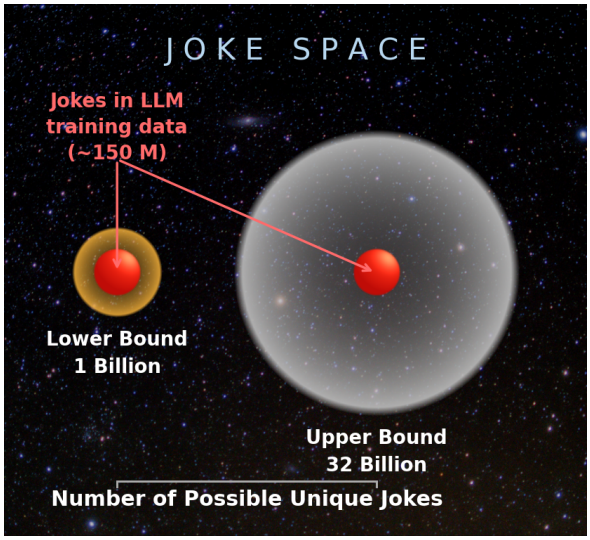

Figure 1: The "Joke Space" versus LLM training coverage. Sphere ("planet") volumes are proportional to estimated unique English joke counts, derived from vocabulary combinatorics under the General Theory of Verbal Humor: noun pairs linked by 5 association bridges each yield the Lower Bound (1B jokes, 14k-noun adult vocabulary) and Upper Bound (32B jokes, 80k WordNet synsets). Both dwarf the estimated jokes in LLM training data (150M; red core, sized identically in both spheres), suggesting LLMs have encountered only a small fraction of all possible jokes and thus have meaningful potential for generating genuinely novel humor. Full derivation is in Amir (2025). Note: all spheres are to scale (the red core represents 15% of the Lower Bound planet's volume and less than 0.5% of the Upper Bound's) though volume differences are difficult to appreciate in a 2D projection.

erogeneity of variance in funniness ratings across sources ($F(2, 237) = 30.69$, $p < .001$), with Benchmark jokes showing considerably greater spread than the two AI-generated sets (as expected per the benchmark's design – see Introduction).

3.3 Relationship Between Originality and Funniness Within Each Source

The relationship between item-level originality and funniness differed markedly across sources (see Figure 2). For Benchmark jokes the correlation was positive (ρ = .554, $p < .001$), for GPT-4o jokes it was negative (ρ = -.247, $p < .001$), and for GPT-5.4 jokes it was weakly positive (ρ = .239, $p < .001$). All three pairwise differences between these correlations were statistically significant after Holm correction by both Fisher's z and bootstrap tests: Benchmark vs. GPT-4o ($\Delta\rho$ = .80, $p < .001$), Benchmark vs. GPT-5.4 ($\Delta\rho$ = .32, p = .018), and GPT-4o vs. GPT-5.4 ($\Delta\rho$ = -.49, p = .004).

3.4 Correlations Between Automated Measures and Human Originality Ratings

Correlation analyses were restricted to the 240 jokes rated by all three raters (80 per source). Within this subset, the mean human originality rating was M = 2.16 (SD = 0.55) and the mean human funniness rating was M = 1.32 (SD = 0.79), both on a 0–3 scale.

All five automated measures correlated positively and significantly with human originality ratings (Figure 3). The two strongest predictors were Corpus Novelty (ρ = .372, 95% CI [.26, .48], $p < .001$) and Concept Distance Max (ρ = .369, 95% CI [.25, .47], $p < .001$), which were nearly identical in magnitude despite capturing distinct aspects of originality: Corpus Novelty reflects how semantically distant an entire joke is from the nearest joke in the reference corpus, while Concept Distance Max reflects the semantic distance between the two most distant handles within the joke itself. Their moderate inter-correlation (ρ = .392, 95% CI [.28, .50], $p < .001$) suggests they capture partially overlapping but meaningfully distinct signals.

The three PMI-based measures showed weaker but consistently significant associations, all in the range ρ = .23–.25: PMI Raw Max (ρ = .231, 95% CI [.11, .35], $p < .001$), PMI Decomposed Max (ρ = .232, 95% CI [.11, .35], $p < .001$), and PMI Cluster Max (ρ = .248, 95% CI [.12, .36], $p < .001$). All three index the co-occurrence rarity of the joke's most original handle pair in the reference corpus. As detailed in Section 3.6, PMI Raw Max and PMI Decomposed Max are near-identical by construction, while PMI Cluster Max captures related but partially independent information through semantic cluster smoothing rather than exact token co-occurrence.

3.5 Correlations Between Automated Measures and Human Funniness Ratings

None of the five automated measures of originality showed a significant correlation with human funniness ratings (all $p > .08$; Figure 3). Corpus Novelty showed a small negative trend (ρ = -.059) and PMI Cluster Max a small positive one (ρ = .073), but neither approached significance. This pattern of null results suggests the automated framework captures something specific to originality rather than to joke quality more broadly.

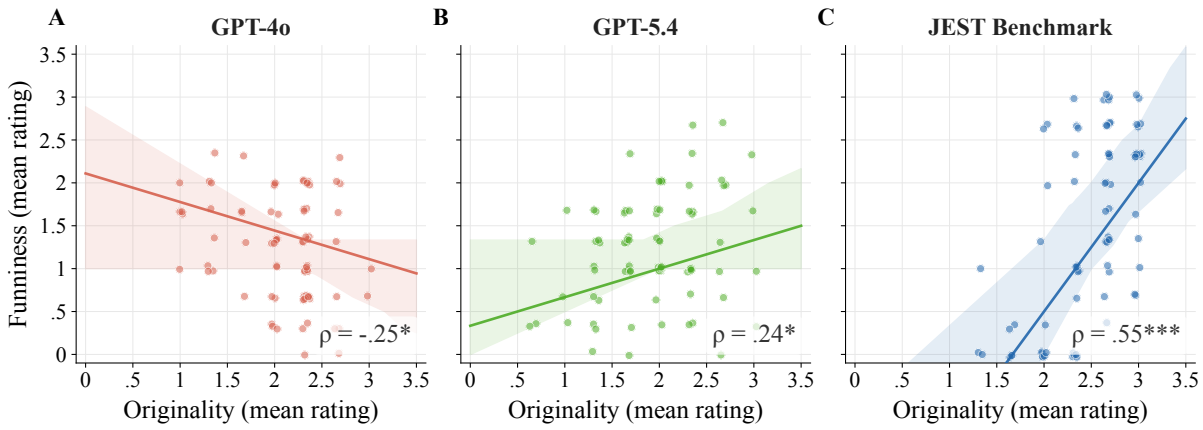

Figure 2: Scatterplots showing the relationship between mean originality and mean funniness ratings by three expert raters for jokes from each source. (A) GPT-4o, (B) GPT-5.4, (C) JEST benchmark. Each data point represents a single joke rated by all three raters, with coordinates reflecting the mean rating across raters. The line shows the Theil-Sen fit with 95% bootstrap confidence band (1,000 resamples). ρ = Spearman rank correlation. * p < .05, *** p < .001.

3.6 Inter-correlations Among Automated Measures

The inter-correlations among automated measures are shown in Figure 3. PMI Raw Max and PMI Decomposed Max were near-perfectly correlated (ρ = .903, 95% CI [.88, .92], p < .001). The two measures return identical values when all handles are single words. The small residual difference arises from the minority of jokes containing at least one two-word handle, which tend to be rarer or absent in the reference corpus. PMI Cluster Max showed moderate correlations with both PMI Raw Max (ρ = .243, 95% CI [.12, .36], p < .001) and PMI Decomposed Max (ρ = .276, 95% CI [.15, .39], p < .001), confirming that the cluster-based smoothing approach captures related but not interchangeable information. Corpus Novelty correlated moderately with Concept Distance Max (ρ = .392, 95% CI [.28, .50], p < .001) and PMI Cluster Max (ρ = .304, 95% CI [.18, .42], p < .001), but showed little relationship with PMI Raw Max (ρ = .098, p = .128) or PMI Decomposed Max (ρ = .088, p = .170), indicating that full-text semantic novelty and handle-pair co-occurrence rarity capture largely independent aspects of originality. Concept Distance Max correlated moderately with all PMI measures (ρ = .222–.440, all p < .001), suggesting partial convergence between semantic distance and co-occurrence rarity, while retaining sufficient independence to contribute distinct information.

3.7 Composite Originality Score

To assess whether combining automated measures improves prediction of human originality ratings, we constructed a composite score using Lasso regression with the alpha parameter selected via 10-fold cross-validation. Applied to the five automated originality measures, Lasso assigned non-zero weights to only three predictors: Corpus Novelty (51%), Concept Distance Max (43%), and PMI Decomposed Max (6%), zeroing out PMI Raw Max and PMI Cluster Max entirely — consistent with their redundancy with PMI Decomposed Max and weaker independent contribution respectively. The in-sample composite correlated with human originality at ρ = .417 (95% CI [.31, .52], p < .001, n = 235 — excluding single-handle jokes). Cross-validation within the 235-joke subset, for which all human ratings and all originality measures were available, yielded ρ = .401 (10-fold) and ρ = .388 (LOO), confirming that the composite generalizes beyond the training sample, though the gain over the best individual measure (Concept Distance Max, ρ = .369) is modest.

To examine whether joke length confounds the composite, word count was added as an additional predictor. Length correlated substantially with human originality ratings (ρ = .358, p < .001), and the Lasso assigned it 27% of the composite weight, partially displacing Corpus Novelty (51% → 41%) and Concept Distance Max (43% → 31%). The cross-validated composite improved marginally (LOO ρ = .39 → .41). However, from a GTVH perspective, joke length should not affect the core script opposi-

tion or logical mechanism that defines originality, and conceptual novelty is independent of how elaborately it is expressed. While we acknowledge that the craft of comedy writing involves careful word selection, empirical evidence suggests professional comedians would consider a reworded joke with the same premise (operationalized as these two structural components) as joke theft (Cain et al., 2024). Length may either create an illusion of originality in human raters or correlate with aspects of originality not directly captured by our measures, but neither justifies its inclusion in a theoretically grounded assessment tool. We therefore retain the length-free composite as our primary measure.

Applying the Spearman-Brown formula to the mean pairwise inter-rater agreement (ρ = .24, k = 3) yields an estimated reliability of the mean originality rating of ρ = .49, the theoretical ceiling for predicting it. The cross-validated composite (ρ = .40) reaches approximately 82% of this ceiling, suggesting the automated measures capture most of the reliably predictable variance in human originality judgments.

4 Discussion

This work explores automated measures of humor originality assessment. Several of the measures correlated significantly with originality judgments of experienced human comedy writers. The hope is that such measures will prove useful for judging the likelihood that human- or AI-generated jokes are novel.

4.1 Relationship between Perceived Originality and Funniness

Three joke sources were used: naïve prompting of GPT-4o and GPT-5.4 to generate original jokes, and the JEST benchmark, where humor was generated in response to short texts using the commercially available Witscript application (accessed November 2025) and curated to achieve a wide distribution of funniness ratings. Figure 2 shows a significant pattern where, in GPT-4o, a negative correlation emerged between perceived originality and funniness by human raters. That pattern reversed for GPT-5.4 and peaked with the strongest positive correlation for the Witscript-generated JEST benchmark. That more original jokes would be perceived as funnier is consistent with the long-held view among some humor researchers as well as comedy writers that surprise is either an essential element of humor or, if not necessary, at least enhances amusement (Hurley et al., 2011; Suls, 1972). However, these results are also consistent with the view that surprise or novelty is not a sufficient condition for humor. The negative correlation between originality and funniness observed in the earliest model (GPT-4o) is consistent with Veale's (2024) findings that this model, when prompted plainly to generate jokes, produces either funny but unoriginal jokes clearly retrieved from its training set or unfunny attempts at original humor. GPT-5.4's moderate positive correlation suggests it may generate somewhat better original jokes. However, even a superficial review of its outputs reveals heavy reliance on variations of classical joke themes. For instance, five of the 200 jokes were variations of puns involving a mirror reflecting: "Why did the mirror become a therapist? It had a gift for helping people reflect" and "Why did the mirror get promoted? It really knew how to reflect well on the company."

Since the JEST benchmark used an LLM-powered application that is algorithmically designed for joke generation, as well as specific setups (short texts) in the prompts, it does not rely as much on preexisting jokes. Consequently, its jokes were rated reliably more original than the jokes produced by the general-purpose models in response to generic prompts (i.e., "generate 10 original jokes").

4.2 Humor Detection vs. Originality

A common criticism of LLM creativity, especially in domains like humor where novelty and surprise are highly valued, is that the statistical pattern completion the model is based on renders it a "stochastic parrot" rather than a source of true creativity (Bender et al., 2021). Classical machine learning approaches for automated humor detection concluded that the best way to recognize humor is to find words that appear more often in jokes compared to non-jokes (sometimes referred to as the "bag of words" approach; Mihalcea and Strapparava, 2006). However, as we have seen, words that are infrequently used in jokes also signal their novelty. Modern approaches for humor detection make use of semantic embedding models with broader context representations (Tasnia et al., 2023). However, as our results show, semantic embedding distance also signals originality. There is thus an inherent tension: the statistical features that make a joke recognizable as humor (e.g., frequent concept pairings and semantic proximity to known jokes) are precisely the features that make it unoriginal.

	1	2	3	4	5	6	7	M	SD
1. Human Originality	—							2.16	0.55
2. Human Funniness	0.289***	—						1.32	0.79
3. Corpus Novelty	0.372***	-0.059	—					0.36	0.10
4. PMI Raw Max	0.231***	0.104	0.098	—				1.49	2.12
5. PMI Cluster Max	0.248***	0.073	0.304***	0.243***	—			0.16	0.20
6. PMI Decomposed Max	0.232***	0.085	0.088	0.903***	0.276***	—		1.69	2.62
7. Concept Distance Max	0.369***	0.092	0.392***	0.222***	0.440***	0.230***	—	0.86	0.10

Figure 3: Spearman rank-order correlation matrix for human ratings and automated originality measures (n = 240, jokes rated by all three raters). Cell shading reflects the magnitude and direction of each correlation (blue = positive, white = 0, red = negative). The lower triangle displays the Spearman ρ coefficient with significance stars and 95% confidence intervals computed via Fisher z-transformation. All measures entering the correlations are oriented so that higher values indicate greater originality; M and SD for PMI measures are shown in the raw (un-negated) scale, where lower values indicate greater originality. Corpus Novelty is the complement of maximum cosine similarity between the joke and the reference corpus (1 - BERT_full); PMI Raw, PMI Cluster, and PMI Decomposed quantify the co-occurrence rarity of the most original handle pair per joke (i.e., the pair with the lowest co-occurrence in the reference corpus); Concept Distance is the embedding-based semantic distance of the most semantically distant handle pair (1 - cosine similarity). Undefined PMI values for jokes where all handle pairs were absent from the corpus are imputed to the theoretical maximum originality score. Human Originality and Human Funniness are mean rater scores on a 0–3 scale. * p < .05, ** p < .01, *** p < .001.

An LLM trained to produce text that looks like humor is therefore also trained, implicitly, to produce humor that looks like existing humor.

4.3 Towards Automated Originality Assessment of AI-Generated Humor

We explored several variant methods for probabilistic automated joke originality assessment. The methods can be reduced to identifying how rarely a joke's most distinctive handle pair co-occurs in a large corpus of jokes, or how semantically distant that pair is. The measures vary in specificity (cluster vs. raw PMI) and sensitivity to context (PMI vs. embedding). A Lasso-based composite of the three strongest predictors: Corpus Novelty, Concept Distance Max, and PMI Decomposed Max — modestly outperformed individual measures on cross-validated prediction of human originality ratings (ρ = .40), suggesting that the measures capture partially complementary signals. Remarkably, when considering the modest human experts' inter-rater reliability, this composite automated measure captures 82% of the reliably predictable variance in their originality judgments. These methods are ideal for cases in which it is clear from the context that a text segment is an instance of humor. The conflict described in Section 4.2 between humor detection and novelty judgment remains to be resolved.

Limitations

All but one of the proposed methods rely on the assumption that jokes can be reduced to a few concepts extractable as "handles" and linked together. While the positive correlation between these measures and human originality ratings is promising, it is possible that this approach only works for a subset of humor.

Since we do not have access to all the humor ever produced nor to the full set of jokes included in the training set of most modern LLMs, a ground truth originality score is impossible to obtain. We must therefore rely on the judgments of human comedy writers. However, even the most experienced comedians have likely been exposed to a smaller set of jokes than modern LLMs (Amir, 2025; Brawer and Amir, 2021), so they are likely to rate some jokes as original that are nevertheless available in an LLM training set.

This work relies exclusively on the SocialGrep (2021) dataset. While this dataset includes over a million highly diverse short jokes sourced from

Reddit, it still likely fails to capture the full richness of human humor. This is fine as a proof of concept, but for real-world applications a broader and more diverse set of reference humor should be used.

References

Ori Amir. 2025. Are AI-generated jokes truly original? Charting the "Joke Space". In *Proceedings of the 16th International Conference on Computational Creativity*, pages 302–308, Campinas, Brazil. Association for Computational Creativity.

Ori Amir and Irving Biederman. 2016. The neural correlates of humor creativity. *Frontiers in Human Neuroscience*, 10:597.

Ori Amir, Konrad J. Utterback, Justin Lee, Kevin S. Lee, Suehyun Kwon, Dave M. Carroll, and Alexandra Papoutsaki. 2022. The elephant in the room: Attention to salient scene features increases with comedic expertise. *Cognitive Processing*, 23(2):203–215.

Salvatore Attardo and Victor Raskin. 1991. Script theory revis(it)ed: Joke similarity and joke representation model. *Humor: International Journal of Humor Research*, 4(3-4):293–347.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Kim Binsted and Graeme Ritchie. 1997. Computational rules for generating punning riddles. *Humor: International Journal of Humor Research*, 10(1):25–76.

Jacob Brawer and Ori Amir. 2021. Mapping the "funny bone": Neuroanatomical correlates of humor creativity in professional comedians. *Social Cognitive and Affective Neuroscience*, 16(9):915–925.

Kathleen Cain, Steven Gimbel, Lindsay Howard, Brittany Maronna, and Sean Beirne. 2024. Joke synonymy sensitivity among working comedians and the General Theory of Verbal Humor. *Humor: International Journal of Humor Research*, 37(4):513–528.

Noam Chomsky, Ian Roberts, and Jeffrey Watumull. 2023. The false promise of ChatGPT. *The New York Times*.

Greg Dean. 2000. *Step by step to stand-up comedy*. Heinemann, Portsmouth, NH.

Denis Dumas, Peter Organisciak, and Michael Doherty. 2021. Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts*, 15(4):645.

Christian F. Hempelmann and Willibald Ruch. 2005. 3WD meets GTVH: Breaking the ground for interdisciplinary humor research. *Humor: International Journal of Humor Research*, 18(4):353–387.

Matthew M. Hurley, Daniel C. Dennett, and Reginald B. Adams, Jr. 2011. *Inside jokes: Using humor to reverse-engineer the mind*. MIT Press.

Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! Humor is still challenging Large Language Models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 325–340, Toronto, Canada. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Carolyn Lamb, Daniel G. Brown, and Charles L. Clarke. 2018. Evaluating computational creativity: An interdisciplinary tutorial. *ACM Computing Surveys*, 51(2):1–34.

Tyler Loakman, William Thorne, and Chenghua Lin. 2025. Who's laughing now? An overview of computational humour generation and explanation. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 780–794, Hanoi, Vietnam. Association for Computational Linguistics.

Rada Mihalcea and Carlo Strapparava. 2006. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence*, 22(2):126–142.

Reuben Narad, Siddharth Suresh, Jiayi Chen, Pine S. L. Dysart-Bricken, Bob Mankoff, Robert Nowak, Jifan Zhang, and Lalit Jain. 2025. Which LLMs get the joke? Probing non-STEM reasoning abilities with HumorBench. *Preprint*, arXiv:2507.21476.

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.

Willibald Ruch, Salvatore Attardo, and Victor Raskin. 1993. Toward an empirical verification of the General Theory of Verbal Humor. *Humor: International Journal of Humor Research*, 6(2):123–136.

Mark A. Runco, Burak Turkman, Selcuk Acar, and Ahmed M. Abdulla Alabbasi. 2024. Examining the idea density and semantic distance of responses given by AI to tests of divergent thinking. *The Journal of Creative Behavior*, 59(3):e1528.

Dafna Shahaf, Eric Horvitz, and Robert Mankoff. 2015. Inside jokes: Identifying humorous cartoon captions. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1065–1074.

SocialGrep. 2021. One million Reddit jokes [Dataset]. Hugging Face. https://huggingface.co/datasets/SocialGrep/one-million-reddit-jokes (accessed April 2025).

Jerry M. Suls. 1972. A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. In Jeffrey H. Goldstein and Paul E. McGhee, editors, *The psychology of humor: Theoretical perspectives and empirical issues*, volume 1, pages 81–100. Academic Press, New York.

Radiathun Tasnia, Nabila Ayman, Afrin Sultana, Abu N. Chy, and Masaki Aono. 2023. Exploiting stacked embeddings with LSTM for multilingual humor and irony detection. *Social Network Analysis and Mining*, 13(1):43.

Joe Toplyn. 2014. *Comedy Writing for Late-Night TV*. Twenty Lane Media.

Joe Toplyn. 2021. Witscript: A system for generating improvised jokes in a conversation. In *Proceedings of the 12th International Conference on Computational Creativity*, pages 22–31, Mexico City, Mexico (Virtual). Association for Computational Creativity.

Joe Toplyn. 2023. Witscript 3: A hybrid AI system for improvising jokes in a conversation. *Preprint*, arXiv:2301.02695.

Joe Toplyn and Ori Amir. 2025. Can AI make us laugh? Comparing jokes generated by Witscript and a human expert. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 71–78, Online. Association for Computational Linguistics.

Joe Toplyn and Ori Amir. 2026. JEST: A benchmark for rating the funniness of short texts. In *Proceedings of the 17th International Conference on Computational Creativity*, Coimbra, Portugal. Association for Computational Creativity. In press.

Tony Veale. 2024. From symbolic caterpillars to stochastic butterflies: Case studies in re-implementing creative systems with LLMs. In *Proceedings of the 15th International Conference on Computational Creativity (ICCC)*, Jönköping, Sweden. Association for Computational Creativity.