

Resources and Evaluation in Computational Humour Generation: A Systematic Survey across Languages and Cultures

Sukairaj Hafiz Imam

sukhimam00@gmail.com

King Mongkut's University of Technology Thonburi <https://orcid.org/0009-0002-1693-3764>

Systematic Review

Keywords: Humour generation, Language resources, Cultural evaluation, Large language models, FAIR principles

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11258261/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Resources and Evaluation in Computational Humour Generation: A Systematic Survey across Languages and Cultures

Sukairaj Hafiz Imam¹  · Nuttanart Muansuwan¹  · Booncharoen Sirinaovakul¹

© The author(s) ••••

Abstract

Computational humour generation requires outputs that are well-formed, relevant, and funny to a specific audience. Existing reviews treat resources and evaluation separately, so it remains unclear whether datasets represent the languages, cultures and humour forms for which performance is claimed and whether evaluations support those claims. In this paper, we present a systematic survey of how humour-generation research builds, shares and evaluates its resources. We analysed 74 studies published between 2020 and 2026, selected from 1,226 records across five databases, coded each with a 57-field framework, and assessed the released resources against the FAIR (Findable, Accessible, Interoperable, Reusable) principles. Our results show that large language models appeared in 73.0% of studies, yet resources were publicly available in only 28.4%, only one study named a data licence, and none of the 19 released resources had a persistent identifier. English-only work made up half the corpus, only four studies evaluated cultural appropriateness directly, and only 39.2% reported statistical testing. The main obstacle is therefore no longer a lack of data but the difficulty of finding, reusing, and verifying the data that exist, which calls for licensed, findable resources and evaluation grounded in the intended audience.

Keywords: Humour generation, Language resources, Cultural evaluation, Large language models, FAIR principles

1. Introduction

Computational humour research encompasses four principal functions: detecting or recognising humorous content, understanding or explaining why it is humorous, generating new humorous material, and evaluating the quality and appropriateness of humour. Detection identifies humour and its attributes (Ren et al. 2024), understanding models mechanisms such as incongruity, ambiguity and pragmatic intent (Zangari et al. 2025), generation produces humorous content (Poobanchuen et al. 2025), and evaluation determines whether outputs are funny, relevant, appropriate and safe (Agrawal and Mamidi 2026). Although these functions are closely connected, generation presents a distinctive challenge because a system must create rather than merely identify or interpret humour. Its outputs must be linguistically well-

✉ S.H. Imam
sukairaj.imam@kmutt.ac.th
N. Muansuwan
nuttanart.mua@kmutt.ac.th
B. Sirinaovakul
booncharoen.sir@kmutt.ac.th

¹ Department of Computer Engineering, Faculty of Engineering, King Mongkut's University of Technology Thonburi

formed, contextually relevant, sufficiently unexpected and recognisably amusing to an intended audience (Loakman, Thorne, and Lin 2025).

Humour generation may depend on lexical ambiguity (Mittal, Tian, and Peng 2022), phonetic resemblance (Sarrof 2025), pragmatic inversion (Mishra 2025), expectation violation (Hossain et al. 2020), or interactional timing and delivery (Kiderle et al. 2025). These mechanisms can operate separately or together, so a single prompt may admit many substantially different responses. Unlike tasks evaluated mainly through factual correspondence or semantic similarity, humour generation requires fluency, creativity, contextual reasoning and anticipation of how an audience will interpret the output.

Language, culture, context and audience further determine what counts as humorous. Linguistic structure shapes the ambiguities, phonological similarities and figurative constructions available to a generator (Inácio and Oliveira 2025), while cultural knowledge supplies shared references, traditions and social expectations (Xian et al. 2025). Context affects whether an utterance is read literally, playfully, sarcastically or offensively (Quan, Ramakrishnan, and Chin 2025; Dogra et al. 2026), and audience characteristics influence perceived funniness, relevance and acceptability (Zhang et al. 2024; Garimella et al. 2020). These dimensions are related but not interchangeable: grammatical generation in a target language does not demonstrate cultural knowledge, and multilingual generation alone does not establish cultural appropriateness. The distinction matters most for low-resource languages, regional varieties and code-mixed communication, where suitable resources and culturally matched evaluators remain scarce.

Approaches to humour generation have changed substantially. Early systems relied largely on templates, lexical resources and explicit linguistic rules (Yamane, Mori, and Harada 2021; Yoshikawa and Iwakura 2020). These were followed by task-specific neural and encoder–decoder models for puns (Tian, Sheth, and Peng 2022), headline rewriting (Login, Baranov, and Braslavski 2022), dialogue (Seki 2022; Chauhan et al. 2023), and image-conditioned captioning (Li et al. 2023; Wang et al. 2021). More recently, pretrained and instruction-tuned models have enabled zero- and few-shot generation (Agrawal and Mamidi 2026), supervised adaptation (Ravi et al. 2024; Shimozaki et al. 2024), parameter-efficient fine-tuning (Poobanchuen et al. 2025), retrieval-augmented generation (Hong et al. 2026; Shen et al. 2026; Eskandari et al. 2026), and preference optimisation (Brüggemann and Hou 2026). LLMs also act as co-creators (Cao et al. 2025), candidate selectors (Kiderle et al. 2025; Arora, Arowolo, and Hoblitzell 2026), critics (Ravi et al. 2024; Suppa et al. 2026), and output judges (Sakabe et al. 2026; Tang 2026a). This expansion has broadened generation settings but made it harder to tell whether humour knowledge comes from resources, model parameters, prompts, retrieval components or evaluation feedback.

Resources and evaluation should therefore be examined as connected stages of a humour-generation lifecycle (Fig. 1). Resource collection determines which languages, cultural settings, humour forms and audiences are represented (Zhang et al. 2024). Construction and annotation determine whether systems receive humour labels, paired transformations, contextual information, constraints or preference signals (Tikhonov and Ivanov 2026). Training, prompting and retrieval then determine whether humour knowledge is encoded in model parameters (Mishra 2025; Inácio and Oliveira 2025), supplied through instructions and demonstrations (Abdulmujeeb and Aryal 2026; Schechter, Barzel, and Liebeskind 2026), or obtained from external sources (Suppa et al. 2026). Evaluation defines successful generation through automatic measures (Chen et al. 2024b; Zhang et al. 2026), human judgements (Winters and Van der Stockt 2025; Gorenz and Schwarz 2024), or LLM-based assessment (Sakabe et al. 2026; Eskandari et al. 2026). Weaknesses can propagate across these stages: a culturally narrow resource constrains what a model can learn, and evaluation based only on fluency or lexical overlap may fail to expose that limitation.

Existing reviews have examined humour detection, generation, multimodal humour and evaluation, but each covers only part of this lifecycle (Section 2). In particular, resource construction (Hossain et al. 2020; Zhong et al. 2024), annotation transparency (Tian, Sheth, and Peng 2022; Asapu et al. 2025), availability and licensing (Zhang et al. 2024; Zhong

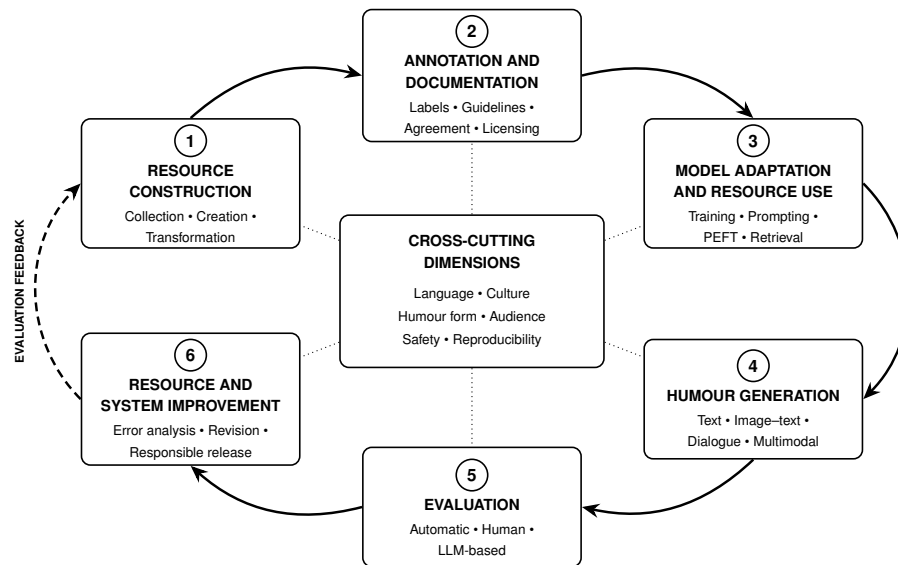


Figure 1. Resource–evaluation lifecycle for computational humour generation. Resources move through construction, annotation, model adaptation, generation and evaluation. Evaluation findings feed back into resource and system development (dashed arrow). Cross-cutting dimensions (centre) apply at every stage.

et al. 2024), resource–model relationships (Li et al. 2023; Zhang et al. 2024), evaluator reporting (Hong et al. 2026; Castro et al. 2026), and cultural and safety evaluation (Xian et al. 2025; Dogra et al. 2026) have not been synthesised as interconnected dimensions. It therefore remains unclear whether current resources represent the languages, cultures and humour forms for which performance is claimed, and whether prevailing evaluations provide reliable evidence for those claims.

This systematic survey addresses this gap by analysing 74 computational humour-generation studies published between 2020 and 2026. It investigates the following questions:

1. What resources support computational humour generation, and what are their principal characteristics?
2. How are these resources constructed and annotated, and which languages, cultures and humour forms do they cover?
3. Which modelling, prompting and adaptation approaches use these resources, and in what ways?
4. What evaluation practices are used for humour-generation systems, and to what extent are they reliable and culturally appropriate?
5. Which resource and evaluation gaps should guide future research?

The survey makes four contributions. First, it provides a structured account of the datasets, benchmarks, lexical and knowledge sources, preference data and experimental materials that support humour generation. Second, it assesses the findability, accessibility, interoperability and reusability of the released resources against the FAIR principles. Third, it relates resource construction, annotation and linguistic and cultural coverage to modelling, prompting, fine-tuning, PEFT, retrieval and preference-based adaptation. Fourth, it examines the validity and reproducibility of automatic, human and LLM-based evaluation, including evaluator reporting, statistical testing, cultural appropriateness and safety. Together, these analyses treat resource availability, resource use and evaluation validity as connected requirements for humour generation across languages and cultures.

2. Related Work

Previous reviews cover computational humour through recognition, generation, multimodality, modelling, and evaluation. Table 1 compares their scope, resources, linguistic coverage, methods, evaluation, LLM use, and reproducibility.

Table 1. Comparison of related computational-humour reviews and the present survey.

Comparison dimension	(54)	(33)	(7)	(41)	(37)	(39)	(31)	Present survey
Primary focus	Recognition	Recognition	Generation	Generation	Cross-task	Multimodal	Evaluation	Generation
Resources and datasets	✓	✓	✓	△	△	✓	△	✓
Resource construction	△	△	△	x	△	△	x	✓
Annotation methodology	△	△	△	x	△	△	△	✓
Languages and multilingual settings	✓	✓	△	△	△	△	x	✓
Cultural coverage and grounding	△	△	△	✓	△	△	x	✓
Models and adaptation	✓	✓	✓	✓	✓	✓	✓	✓
LLM-era methods	△	△	x	✓	✓	✓	✓	✓
Automatic and human evaluation	✓	✓	✓	△	△	✓	✓	✓
LLM-based evaluation	x	x	x	△	△	△	✓	✓
Cultural and safety evaluation	x	x	△	△	△	△	△	✓
Reproducibility and availability	△	△	x	x	△	△	△	✓
Integrated resource–evaluation analysis	x	x	x	x	x	x	x	✓
FAIR assessment of released resources	x	x	x	x	x	x	x	✓

Note. ✓ indicates substantial or systematic coverage, △ indicates partial coverage or discussion without a central systematic focus, and x indicates little or no systematic coverage.

Table 1 shows that earlier reviews cover particular tasks, datasets, models or evaluation methods in depth, but give only partial attention to resource construction, annotation methodology, availability and reproducibility. LLM-based, cultural and safety evaluations are examined centrally only in the most recent reviews, and none analyse resources and evaluation together or assess the reusability of released resources. The present survey fills this gap by treating resource development, model use and evaluation as stages of a single lifecycle (Fig. 1), by assessing released resources against the FAIR principles, and by asking whether the available evidence supports claims about humour quality across languages, cultures and audiences.

3. Review Methodology

3.1. Review Design and Scope

This study is a systematic survey of resources and evaluation practices in computational humour generation. A systematic approach was needed because the evidence is spread across natural language processing, artificial intelligence, computational creativity, multimedia and human–computer interaction. Reporting follows PRISMA 2020 for information sources, eligibility and study selection (Page et al. 2022). The aim was descriptive and interpretive synthesis rather than statistical meta-analysis.

3.2. Information Sources and Search Strategy

Searches were run between 1 August and 5 September 2026 in five sources, using two complementary Boolean queries: Query 1 targeted resources and their construction, and Query 2 targeted modelling and evaluation.

Query 1: Resources and construction

```
("humor generation" OR  
"humour generation" OR  
"humorous text generation" OR  
"joke generation" OR  
"pun generation")  
AND  
(dataset OR datasets OR  
corpus OR corpora OR  
resource OR resources OR  
benchmark OR benchmarks OR  
annotation OR annotations)
```

Query 2: Models and evaluation

```
("humor generation" OR  
"humour generation" OR  
"humorous text generation" OR  
"joke generation" OR  
"pun generation")  
AND  
(model OR models OR prompting OR  
"fine-tuning" OR "fine tuning" OR  
evaluation OR evaluations OR  
metric OR metrics OR culture OR  
cultural OR multilingual OR  
safety)
```

The searches returned 1,226 database records: 650 from Google Scholar, 232 from the ACL Anthology, 167 from the ACM Digital Library, 97 from IEEE Xplore and 80 from Scopus. Backward and forward citation searching of relevant studies and earlier reviews added 13 unique records, which were assessed against the same eligibility criteria.

3.3. Eligibility and Study Selection

Studies were included if they (i) were published between 2020 and 2026, (ii) substantively addressed computational humour generation, (iii) presented or evaluated a computational model, method, system, dataset, benchmark or framework, and (iv) provided enough English-language information for assessment and extraction. Journal, conference and workshop papers were eligible, as were shared-task overview and system papers reporting distinct contributions. Studies were excluded if humour was incidental to another problem, if they addressed only detection, understanding, explanation or evaluation, if they presented no substantive generation method, or if they were duplicates or superseded versions of another contribution.

One reviewer screened titles, abstracts and full texts, and a second reviewer verified the decisions against the criteria. Disagreements were resolved by discussion and re-examination of the publication. After 490 duplicates were removed, 736 records were screened by title and abstract and 659 were excluded. The remaining 77 database records and the 13 citation-search records gave 90 full-text reports, of which 16 were excluded, leaving 74 included studies (Fig. 2).

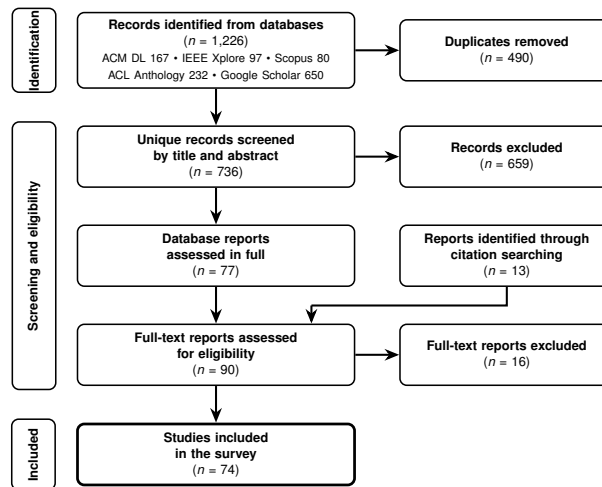


Figure 2. Study-selection process, following the PRISMA 2020 flow structure.

3.4. Data Extraction and Coding

A 30-field extraction form was built around the five research questions, and the extracted evidence was then normalised into 27 controlled variables for quantitative and thematic synthesis, giving a 57-field analytical framework (Table 2).

Table 2. Data-extraction framework: 30 extraction fields and 27 normalised variables.

Domain	Extraction fields	Normalised variables
Bibliographic	Identifier, title, authors, year, venue and type, DOI or URL	Publication type
Eligibility	Eligibility status, exclusion reason	–
Generation scope	Task, humour form, modality and domain	–
Resources	Name and purpose, construction and source, size, availability and licence	Resource type, availability status, licence status, licence name
Language and culture	Language, language and resource setting, cultural grounding, culture-specific expressions	Primary languages, language setting
Annotation	Annotation method	Agent, guidelines, agreement, annotation reporting
Models and adaptation	Architecture, adaptation strategy, prompting or retrieval	Model scale, adaptation category, PEFT, prompting, retrieval
Evaluation	Evaluation design, automatic metrics, human evaluation, LLM as judge, cultural and safety evaluation	Human evaluation status, LLM role, cultural evaluation level, safety evaluation level, statistical testing
Reporting quality	Findings and limitations, evidence location	Resource description, evaluator reporting, resource availability, code availability, limitations reported, evidence confidence

3.5. Data Normalisation and Synthesis

Free-text extractions were normalised into controlled categories. Ambiguous cases and conflicts between codes and their free-text evidence were resolved by re-reading the full publication. Coding distinguished generator adaptation from training auxiliary classifiers, rankers or evaluators. LLM output judging from candidate selection, synthetic annotation or reward construction, and retrieval augmentation from generation without an external resource or maintained memory. LLMs were defined as GPT-3-era or later models. Article licences did not count as data licences. Resources were partially available when source data were public but processed data, annotations or outputs were not. Cultural evaluation was indirect when native or target-audience raters assessed outputs without an explicit cultural criterion. Safety evaluation covered offensiveness, stereotyping, toxicity, moral acceptability and suitability for sensitive audiences. Statistical testing included significance tests, correlations and confidence intervals. Counts and percentages use all 74 included studies as the denominator. The synthesis combined descriptive statistics, thematic analysis and cross-tabulation.

3.6. FAIR Assessment of Released Resources

To assess reusability, we built an inventory of every resource that an included study created or released, counting each once, and added SemEval-2020 Task 7 as a reused reference benchmark. Each resource was scored on eight indicators derived from the FAIR principles (Wilkinson et al. 2016) and the RDA FAIR Data Maturity Model (RDA FAIR Data Maturity Model Working Group 2020): persistent identifier, descriptive metadata, retrievability, persistence, open format, documented schema and language metadata, data licence, and provenance (Table 3). Each indicator was scored 0 (not met), 1 (partially met) or 2 (met), based on the resource’s online location (its code repository or web page) and the paper that introduced it. Information that could not be verified from these sources scored 0. A second reviewer independently scored a random 20% sample (four resources, 32 scores). Exact agreement was 71.9% (linearly weighted $\kappa = 0.54$). Seven of the nine disagreements concerned one resource whose reported release pointed to an anonymous review link, so the reviewers had scored different objects. After excluding it, agreement was 91.7% ($\kappa_w = 0.81$). Disagreements were resolved by discussion, and the rubric was then re-applied to all resources.

Table 3. FAIR scoring rubric. Each indicator is scored 0 (not met), 1 (partially met) or 2 (met).

Indicator	2: met	1: partially met
F1 Identifier	DOI or equivalent PID	Stable repository URL
F2 Metadata	Dataset card or datasheet	README or paper description only
A1 Access	Direct open download	Registration, request or partial download
A2 Persistence	Resolves and archived with a PID	Resolves, not archived
I1 Format	Open standard (JSON, CSV, TSV)	Archive or mixed formats
I2 Schema	Field schema and language codes	Fields or labels described, no codes
R1.1 Licence	Named data licence	Software licence or use statement
R1.2 Provenance	Source, collection and origin documented	Partial provenance

4. Results and Synthesis

4.1. Overview of the Evidence Base

The final evidence base comprised 74 studies published between 2020 and 2026. Publication activity was low and stable until 2023, with 16 studies in 2020–2023, and then rose sharply. 58 studies (78.4%) appeared between 2024 and 2026 (Fig. 3A). Most of the evidence therefore postdates the spread of instruction-following language models.

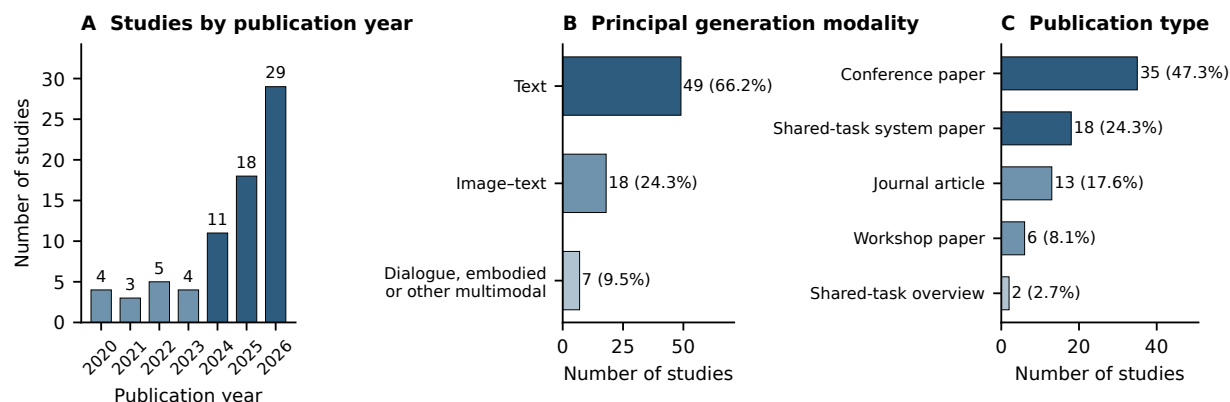


Figure 3. Characteristics of the 74 included studies. (A) Publication year. (B) Principal generation modality (mutually exclusive). (C) Publication type.

Conference and shared-task papers made up most of the corpus (Fig. 3C). SemEval-2026 Task 1 (MWAHAHA) was the largest single source, with 19 studies, including the task overview (Castro et al. 2026). The remaining studies appeared in ACL-family, IEEE and ACM venues, computer-vision and computational-creativity conferences, and journals in artificial intelligence, language technology and human–computer interaction.

Most studies used fully automatic generation (69; 93.2%), in which humans only created resources or evaluated outputs. Three studies (4.1%) examined human–AI co-creation (Mirowski et al. 2024; Wu, Weber, and Müller 2025; Cao et al. 2025), and two (2.7%) used an interactive tool to support human creators (Hossain et al. 2020; Kariyawasam et al. 2024). These categories describe how outputs were produced, not how they were evaluated.

Text was the principal modality in 49 studies (66.2%; Fig. 3B), covering, for example, jokes (He and Mei 2025), puns (Mittal, Tian, and Peng 2022), humorous headlines (Hossain et al. 2020; Weller, Fulda, and Seppi 2020), sarcasm (Mishra 2025), satire (Gorenz and Schwarz 2024), stories (Garimella et al. 2020) and educational content (Fischer, Zylowski, and Wölfel 2025; Liu et al. 2025). Image–text generation accounted for 18 studies (24.3%), including memes (Wang et al. 2021), cartoon captions (Zhang et al. 2024; Önal et al. 2026), image-based Oogiri responses (Zhong et al. 2024) and GIF-conditioned humour (Zare et al. 2026). The remaining seven studies (9.5%) addressed dialogue (Shimozaki et al. 2024), speech and virtual characters (Kiderle et al. 2025), live performance (Winters and Van der Stockt 2025), and multimodal content combining text, images and audio-video (Tang 2026b). Despite growing multimodal work, the evidence base remains mainly textual. The humour forms covered are analysed in Section 4.3.3.

4.2. Language Resources for Humour Generation

The 74 studies used resources at several stages of the generation lifecycle: model training and adaptation, prompt construction, retrieval, candidate ranking and evaluation. Individual studies often combined several external and newly created resources, while prominent benchmarks were reused by many systems. We therefore distinguish study-level resource use from the recurring named resources, which are counted once in the resource inventory (Section 4.2.3).

4.2.1. Resource Types and Purposes

The resources fall into four functional groups. First, *training and adaptation resources* include humorous texts, non-humorous contrasts, contextual inputs or paired transformations for supervised learning, for example humour/non-humour collections for parameter-efficient adaptation (He and Mei 2025), sarcasm data (Mishra 2025), Thai jokes (Poobanchuen et al. 2025) and literal-to-humorous paraphrase pairs (Ravi et al. 2024). Several studies added general-language or conversational corpora, such as the One Billion Word corpus for pun generation (Mittal, Tian, and Peng

2022) and dialogue data combined with a newly transcribed Manzai corpus (Shimozaki et al. 2024). A generation resource can therefore combine humour examples with auxiliary linguistic or contextual data rather than form a single self-contained corpus.

Second, *benchmarks* specified generation inputs and allowed systems to be compared. MWAHAHA was the most reused: all 19 SemEval-2026 Task 1 studies (25.7%) used it, including the task overview (Castro et al. 2026). It supplied headline, word-pair and GIF inputs for constrained generation in English, Spanish and Chinese (Schechter, Barzel, and Liebeskind 2026; Zare et al. 2026; Kongqiang, Peng, and Qingli 2026; Camacho Gonzalez et al. 2026). This gave the recent literature a common basis for comparison, but it also means that a quarter of the corpus shares a single task design. Other recurring benchmarks were Oogiri-GO (Zhong et al. 2024; Sakabe et al. 2026), the SemEval-2017 Task 7 pun data (Mittal, Tian, and Peng 2022; Chen et al. 2024a) and New Yorker cartoons and caption-contest data (Kariyawasam et al. 2024; Zhang et al. 2024).

Third, *preference and audience-response resources* recorded how people judged humour. FunLines collected humorous headline edits with repeated funniness ratings through an online game (Hossain et al. 2020), and the New Yorker Cartoon Caption Contest resource pairs captions with large-scale crowd ratings (Zhang et al. 2024). Other studies built pairwise preference data for candidate ranking (Tikhonov and Ivanov 2026), collected live audience ratings (Winters and Van der Stockt 2025) or recorded location-specific judgements (Garimella et al. 2020). These resources served as evaluation evidence and, increasingly, as supervision for rankers, reward models and preference optimisation (Section 4.4.3).

Fourth, *experimental resources* were built mainly for evaluation rather than training, such as prompt-generated joke sets (Jentsch and Kersting 2023), human–AI co-created memes (Wu, Weber, and Müller 2025), jokes generated across model scales (Agrawal and Mamidi 2026) and preregistered comparisons of human and AI humour (Cao et al. 2025; Gorenz and Schwarz 2024). One embodied system used no dedicated corpus and evaluated generation within a live interaction pipeline (Kiderle et al. 2025).

4.2.2. Resource Scale

Resource size and construction were adequately reported in 64 studies (86.5%). Scale varied by several orders of magnitude. Controlled studies used small sets, such as 30 Portuguese headlines (Inácio and Oliveira 2025), 50 Mad Lib stories (Garimella et al. 2020) or 54 live improvised jokes with 2,160 audience ratings (Winters and Van der Stockt 2025), while shared-task evaluation typically used 300 inputs per language (Castro et al. 2026). Medium-sized task-specific resources included 2,000 Hindi–English code-mixed puns (Asapu et al. 2025), 8,000 Chinese humour instructions (Chen 2026), 12,118 constrained-humour instances (Abbas and Abdul Rauf 2026) and 27,451 rhetorically annotated humorous texts (Chen et al. 2025). The largest resources came from web platforms and audience activity: OxfordTVG-HIC contains about 2.9 million image–caption pairs (Li et al. 2023), the New Yorker resource more than 2.2 million captions with over 250 million ratings (Zhang et al. 2024), and a Japanese dialogue system retained 368,589 training pairs from more than 15.6 million tweet-retweet pairs (Seki 2022). Some large collections were reused rather than created; for example, the Short Jokes collection of 231,657 jokes scraped from joke websites was reused for GPT-2 fine-tuning (Akbar et al. 2021). Scale alone did not determine suitability: large scraped collections offered coverage, whereas smaller resources supplied controlled constraints, cultural grounding, detailed ratings or matched human comparisons.

4.2.3. Availability, Licensing and FAIR Compliance

At the study level, resource accessibility was limited and uneven (Table 4). Resources were publicly available in 21 studies (28.4%) and partially available in 22 (29.7%), typically because the source data were public but the processed data, annotations or outputs were not. Availability was not reported in 24 studies (32.4%). Licensing was reported far less often than availability. Only one study (1.4%) named a licence for its data (Zhang et al. 2024); two named a licence

Table 4. Study-level availability and licensing ($N = 74$). Article licences were not treated as data or code licences.

Reporting category	<i>n</i>	%
<i>Resource availability</i>		
Publicly available	21	28.4
Partially available or unclear scope	22	29.7
Not reported	24	32.4
Unclear	5	6.8
Planned release	1	1.4
Restricted or proprietary	1	1.4
<i>Licensing</i>		
Named data licence	1	1.4
Code licence only	2	2.7
Not reported	56	75.7
Unclear	14	18.9
Publication or reuse restriction only	1	1.4
<i>Code</i>		
Available	26	35.1
Not reported	19	25.7
Unclear or not applicable	29	39.2

that covered code only, and 56 (75.7%) reported none. Licences stated for the published article were not counted as data licences. Code was available for 26 studies (35.1%).

Because a study can report a release that is hard to find, open or reuse, we assessed the released resources themselves against the FAIR principles (Wilkinson et al. 2016). The 19 resources created or released by included studies, and SemEval-2020 Task 7 as a reused reference benchmark, were scored using the procedure in Section 3.6 and the rubric in Table 3 (Fig. 4).

The 19 study resources scored a mean of 8.2 out of 16 (median 9; range 2–14). *Findability* was the weakest principle. No study resource had a persistent identifier; nine were hosted only on GitHub and seven had no stable identifier at all, such as a Google Drive folder, a registration-only platform or an anonymous view-only link (Li et al. 2023; Castro et al. 2026; Gorenz and Schwarz 2024). Only two provided a dataset card (Zhong et al. 2024; Zhang et al. 2024). *Accessibility* was stronger: 13 resources (68.4%) could be downloaded directly and three required registration or a request. Three could not be located despite a reported release, because of a broken repository link, an untraceable Hugging Face record or an anonymous review link (Sakabe et al. 2026; Winters and Van der Stockt 2025; Tang 2026b). None was archived under a persistent identifier.

Interoperability was mixed. Twelve resources (63.2%) used open formats such as JSON, CSV or TSV, but only two described their fields together with machine-readable language metadata (Zhong et al. 2024; Zhang et al. 2024). *Reusability* was limited mainly by licensing. Only two resources carried a named data licence: the New Yorker caption rankings (CC BY-NC 4.0) and Oogiri-GO (MIT) (Zhang et al. 2024; Zhong et al. 2024). Six relied on a software licence for the whole repository or an informal use statement, and eleven (57.9%) had no licence. Provenance was fully documented for nine resources (47.4%), but the documentation seldom addressed the rights attached to content scraped from humour platforms.

4.3. Resource Construction, Annotation and Coverage

Figure 5 summarises the language settings, humour forms, annotation agents and annotation reporting across the 74 studies.

	F1 PID	F2 Metadata	A1 Access	A2 Persistence	I1 Format	I2 Schema	R1.1 Licence	R1.2 Provenance	
New Yorker Caption Ranking	1	2	2	1	2	2	2	2	14/16
Oogiri-GO	1	2	2	1	2	2	2	1	13/16
SemEval-2020 Task 7 / Humicroedit	2	2	2	2	0	2	0	2	12/16
Hinglish homophone and survey data	1	1	2	1	2	1	1	2	11/16
Humor-Parameter-Ablation data	1	1	2	1	2	1	1	2	11/16
RAGthoven joke retrieval corpus	1	1	2	1	2	1	0	2	10/16
ChatGPT joke sets	1	1	2	1	2	1	1	1	10/16
Humor Arena pairs and system artefacts	1	1	2	1	2	1	1	1	10/16
ChinesePun	1	1	2	1	2	1	1	1	10/16
HECoP	1	1	2	1	2	1	0	1	9/16
SLPG constrained-humour dataset	1	1	2	1	2	1	0	1	9/16
FunLines	1	1	2	1	0	1	0	2	8/16
OneLinerJoke (OLJ)	1	0	2	1	2	1	0	1	8/16
MWAHAHA inputs	0	1	1	1	2	1	0	1	7/16
OxfordTVG-HIC	0	1	1	1	0	1	1	1	6/16
Human vs ChatGPT joke study data	0	1	2	1	0	0	0	2	6/16
ARHAHA 2026 inputs	0	1	1	1	0	1	0	1	5/16
Six-style Chinese joke/video corpus	0	1	0	0	0	1	0	2	4/16
Multi-dimensional Oogiri corpus	0	0	0	0	0	0	0	2	2/16
Live improv jokes and ratings	0	0	0	0	0	0	0	2	2/16

0 Not met 1 Partially met 2 Met

Figure 4. FAIR assessment of released humour-generation resources (checked 25 September 2026). Each indicator is scored 0 (not met), 1 (partially met) or 2 (met)

4.3.1. Collection and Construction

Studies reused existing resources when they needed large training sets, established test sets or auxiliary linguistic knowledge. Pun generators combined humour examples with sense inventories, general corpora and pun benchmarks (Mittal, Tian, and Peng 2022; Tian, Sheth, and Peng 2022; Zeng et al. 2024), and multimodal systems drew on existing cartoon, meme, image–caption and Oogiri collections (Zhong et al. 2024; Li et al. 2023; Zhang et al. 2024). Reuse extended coverage, but each system also inherited the annotation conventions, biases and licensing conditions of several independent sources.

New resources were built when existing data did not represent the intended language, humour form, culture or interaction context. Data came from Reddit (Tikhonov and Ivanov 2026), Twitter/X (Seki 2022), news headlines (Hossain et al. 2020), comedy and cartoon platforms (Poobanchuen et al. 2025; Zhang et al. 2024), video transcripts (Shimozaki et al. 2024), crowdsourcing (Wu, Weber, and Müller 2025) and direct participant production (Mirowski et al. 2024; Gorenz and Schwarz 2024). Language- and culture-specific collection produced locally sourced Thai jokes (Poobanchuen et al. 2025), Hindi–English code-mixed puns (Asapu et al. 2025), Arabic constrained-joke inputs (Almasoud et al. 2026) and a transcribed Japanese Manzai corpus (Shimozaki et al. 2024), while other resources recorded humour as situated

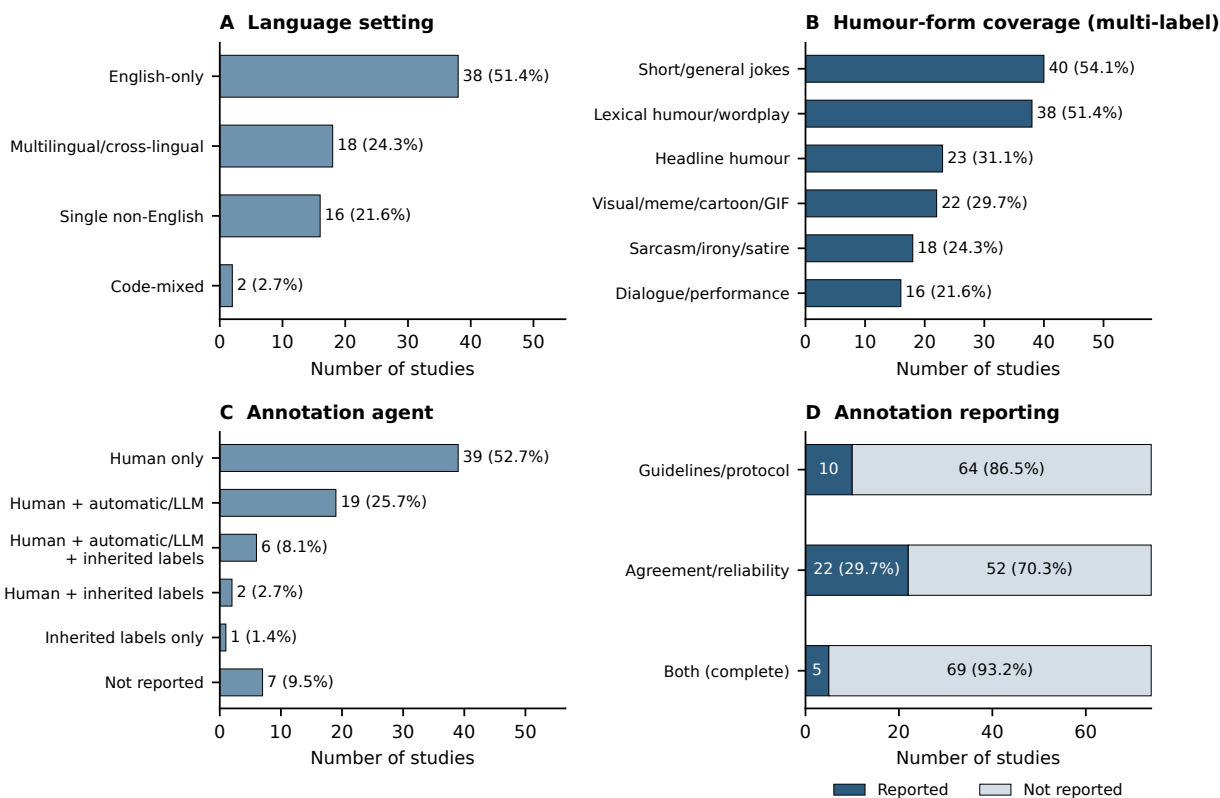


Figure 5. Coverage and annotation across the 74 included studies. (A) Language setting (mutually exclusive). (B) Humour-form coverage (multi-label; percentages do not sum to 100%). (C) Annotation agent. (D) Reporting of annotation guidelines, agreement, and both.

activity in comedy workshops, live improvisation or location-matched experiments (Mirowski et al. 2024; Winters and Van der Stockt 2025; Garimella et al. 2020).

Synthetic and hybrid construction was increasingly common. Teacher models produced literal-to-humorous training pairs (Ravi et al. 2024), multilingual reward data mixed human and synthetic labels (Brüggemann and Hou 2026), and generate-and-select pipelines built candidate pools before ranking them with human preferences, classifiers or LLMs (Tikhonov and Ivanov 2026; Arora, Arowolo, and Hoblitzell 2026). Other resources combined human-authored humour with synthetic augmentation or LLM-assisted filtering (Zhong et al. 2024; Abbas and Abdul Rauf 2026; Chen 2026). These methods expanded scale, but instance-level provenance was often missing: in the FAIR assessment, provenance was fully documented for fewer than half of the released resources (Section 4.2.3).

4.3.2. Annotation

Human judgement remained central (Fig. 5C). Human-only annotation was the largest category (39 studies, 52.7%), followed by human judgements combined with automatic or LLM-assisted procedures (19, 25.7%). Eight studies also used labels inherited from existing datasets, one relied on inherited labels alone (Akbar et al. 2021), and seven did not identify an annotation agent. Annotated constructs went well beyond binary humorousness, covering funniness intensity (Hossain et al. 2020; Winters and Van der Stockt 2025), pun validity and dual-sense compatibility (Tian, Sheth, and Peng 2022; Asapu et al. 2025), relevance and constraint satisfaction (Inácio and Oliveira 2025; Castro et al. 2026), humour style and rhetorical devices (Quan, Ramakrishnan, and Chin 2025; Chen et al. 2025), pairwise preference (Tikhonov and Ivanov 2026; Zhang et al. 2024) and, for multimodal data, image–text correspondence and

caption appropriateness (Kariyawasam et al. 2024; Zhang et al. 2026). Likes, contest votes and live audience responses served as implicit preference signals (Zhong et al. 2024; Winters and Van der Stockt 2025; Zhang et al. 2024). They supported large-scale collection but did not necessarily measure the same construct as an explicit funniness judgement. Annotation reporting was weak (Fig. 5D). Guidelines or protocols were reported in 10 studies (13.5%), agreement statistics in 22 (29.7%) and both in only five (6.8%). Reported measures included Cohen’s and Fleiss’ κ , Krippendorff’s α , intraclass correlation and pairwise agreement (Hossain et al. 2020; Tian, Sheth, and Peng 2022; Asapu et al. 2025; Castro et al. 2026; Chen et al. 2025). These figures refer to resource annotation; agreement among evaluators of generated outputs is examined separately in Section 4.5.2. Differences in scales, rater populations and aggregation procedures limited comparison across resources.

4.3.3. *Linguistic, Cultural and Humour-Form Coverage*

English-only research made up just over half of the corpus (38 studies, 51.4%; Fig. 5A), followed by multilingual or cross-lingual studies (18, 24.3%), single non-English languages (16, 21.6%) and code-mixed generation (2, 2.7%). Beyond English, the corpus covered Chinese, Spanish, Japanese, Hindi, Portuguese, Thai, Arabic and Dutch. Multilingual work depended heavily on MWAHAHA’s parallel English, Spanish and Chinese settings (Castro et al. 2026), and including several languages did not always mean building balanced resources: some systems kept English-centred prompts or knowledge sources, or used translated rather than natively collected humour. Code-mixed evidence was limited to Hindi–English puns (Sarrof 2025; Asapu et al. 2025).

Short jokes and lexical humour were the dominant forms (Fig. 5B). Lexical humour included homographic and homophonic puns (Mittal, Tian, and Peng 2022; Tian, Sheth, and Peng 2022; Chen et al. 2024a), cross-lingual phonetic wordplay (Sarrof 2025; Asapu et al. 2025), headline substitution (Hossain et al. 2020; Weller, Fulda, and Seppi 2020) and word-pair constraints (Abbas and Abdul Rauf 2026; Castro et al. 2026). Visual humour covered memes, cartoon captions and GIF-conditioned responses (Chen et al. 2024b; Zhang et al. 2024; Baruah 2026), and dialogue and performance settings included stand-up monologues, virtual-character jokes and live improvisation (Hong et al. 2026; Kiderle et al. 2025; Winters and Van der Stockt 2025).

Cultural grounding ranged from explicit representation to association with language or source. The most explicit cases encoded established traditions or audiences: Japanese Oogiri, tatoe-tsukkomi and Manzai with its Boke and Tsukkomi roles (Zhong et al. 2024; Seki 2022; Shimozaki et al. 2024), Chinese xiehouyu allegorical sayings (Xu 2024), regional Portuguese material (Inácio and Oliveira 2025; Moreira, Lima Inácio, and Gonçalo Oliveira 2025), locally collected Thai comedy (Poobanchuen et al. 2025), regional humour across Asian countries (Xian et al. 2025), Arabic cultural constraints (Almasoud et al. 2026) and location-specific preferences in India and the United States (Garimella et al. 2020). Elsewhere, culture was usually inferred from language, platform or country of collection rather than recorded as labels, metadata or audience profiles. Culturally grounded forms therefore remained concentrated in a few traditions, and how they were evaluated is examined in Section 4.5.4.

4.4. Models, Adaptation and Resource Use

Systems moved from task-specific statistical and neural architectures to pretrained generators and, more recently, instruction-tuned large language models (LLMs), defined here as GPT-3-era or later models; GPT-2, BERT and earlier pretrained models were coded as other architectures. Many systems combined a generator with classifiers, retrievers, reward models or rule-based validators, so adaptation strategies were coded as multi-label (Fig. 6B). Throughout, we distinguish adaptation of the principal generator from training of auxiliary classifiers, rankers or discriminators.

4.4.1. *Modelling Paradigms*

Earlier studies adapted models to narrowly defined humour forms. Examples include N-gram generation filtered by a recurrent moral-judgement classifier (Yamane, Mori, and Harada 2021), an SVM discriminator combined with

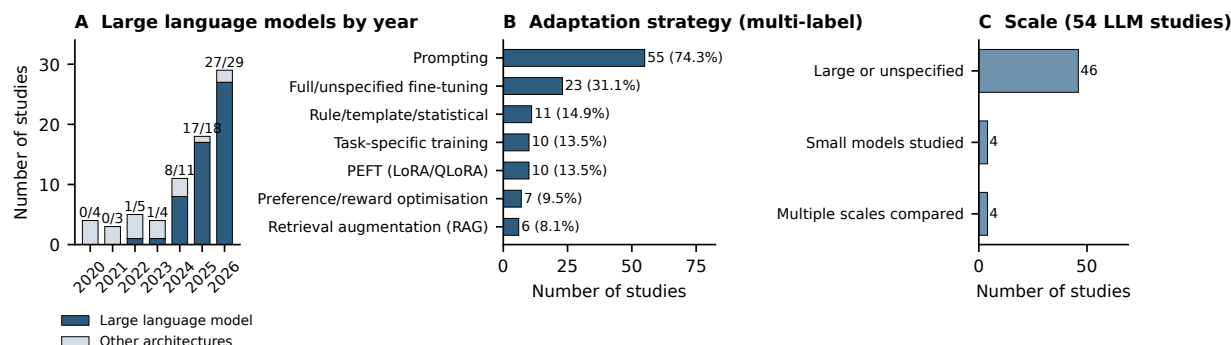


Figure 6. Models and adaptation in the 74 included studies. (A) Studies using a large language model (GPT-3-era or later) by publication year. (B) Adaptation strategies (multi-label; percentages do not sum to 100%). (C) Model scale among the 54 LLM studies.

phonetic and lexical generation (Yoshikawa and Iwakura 2020), a shallow Transformer that rewrote headlines as humorous variants (Weller, Fulda, and Seppi 2020), GPT-2 fine-tuned for joke or pun generation (Su 2022; Tian, Sheth, and Peng 2022; Akbar et al. 2021), and specialised architectures for memes and captions (Sanobar et al. 2025; Wang et al. 2021; Li et al. 2023). Multimodal and conversational systems added visual encoders, acoustic features and dialogue histories (Chen et al. 2024b; Chauhan et al. 2023; Zhai and Wibowo 2023).

LLMs took part in 54 studies (73.0%) as generators, co-creators, critics, judges or candidate selectors. They appeared in none of the 2020–2021 studies but in 27 of the 29 studies from 2026 (Fig. 6A). Recent systems used GPT-family, Llama, Qwen, Gemini, Claude, DeepSeek and Mistral models, often arranged in multi-stage pipelines for association planning, candidate generation, constraint checking and final selection (Schechter, Barzel, and Liebeskind 2026; Shen et al. 2026; Suppa et al. 2026; Tikhonov and Ivanov 2026). Most LLM studies used large or unspecified models (46; Fig. 6C). Proprietary frontier models dominated prompt-based generation and co-creation (Mirowski et al. 2024; Sakabe et al. 2026; Cao et al. 2025), whereas four studies explicitly examined small models, namely LoRA- or QLoRA-tuned Phi-2 and 3B-parameter Llama and Qwen models (Mishra 2025; Poobanchuen et al. 2025; Abbas and Abdul Rauf 2026; Chen 2026). Another four compared model scales (Agrawal and Mamidi 2026; Yamane 2024).

4.4.2. Adaptation Strategies

Prompting was the most common strategy (55 studies, 74.3%), ranging from zero-shot instructions to few-shot demonstrations, personas, structured planning, self-refinement and generate-and-select procedures. Controlled zero-shot prompts supported comparisons across models and scales (Agrawal and Mamidi 2026; Gorenz and Schwarz 2024; Yamane 2024), while structured prompts encoded humour styles, theories, target words or headlines (Abdulmujeeb and Aryal 2026; Schechter, Barzel, and Liebeskind 2026; Inácio and Oliveira 2025). One study used prompting alone to produce multiple joke styles and then convert them into images, audio narration and social-media content (Tang 2026b). Pipelines added association generation, internal voting, self-critique or LLM judges to select among candidates (Kiderle et al. 2025; Shen et al. 2026; Suppa et al. 2026; Bubus et al. 2026). Prompting was especially common in shared-task systems, which typically treated organiser inputs as constraints rather than training data.

Weight adaptation was less common. Full or unspecified fine-tuning of GPT-2, T5, BART, dialogue and multimodal models appeared in 23 studies (31.1%) (Li et al. 2023; Ravi et al. 2024; Su 2022; Tian, Sheth, and Peng 2022; Shimozaki et al. 2024). PEFT, mostly LoRA or QLoRA, appeared in 10 (13.5%) (He and Mei 2025; Mishra 2025; Poobanchuen et al. 2025; Xian et al. 2025); in one pipeline LoRA adapted a discriminator while the generator remained prompt-based (Zhang et al. 2026). Ten studies trained task-specific models outside the pretrained fine-tuning setting, including auxiliary selectors such as a moral-judgement classifier applied to statistical candidates (Yamane, Mori,

and Harada 2021). Preference or reward optimisation (RLHF/PPO, DPO, curriculum DPO, GRPO and discriminator-derived rewards) appeared in seven studies (9.5%), supervised by human preferences, synthetic comparisons or learned humour scores (Ravi et al. 2024; Zhang et al. 2024; Brüggemann and Hou 2026; Chen et al. 2024a). Retrieval augmentation appeared in six (8.1%), drawing on community discussions (Hong et al. 2026), current news (Shen et al. 2026), humour exemplars (Suppa et al. 2026), lexical knowledge graphs (Aali 2026), culture-conditioned indexes (Zare et al. 2026) and Wikipedia (Eskandari et al. 2026).

4.4.3. *Relationships Between Resources and Models*

Resource structure shaped modelling choices more than resource size did (Table 5). Large paired or labelled corpora supported supervised training, small task-specific collections supported PEFT or few-shot prompting, and evaluation inputs without training data led to prompting and generate-and-select pipelines. Modality determined architecture, and preference resources enabled ranking and reward modelling. Language- and culture-specific resources split between the two approaches: locally collected Thai jokes and regional Asian humour were encoded through LoRA or supervised adaptation (Poobanchuen et al. 2025; Xian et al. 2025), whereas multilingual shared-task systems conditioned general models through language-specific prompts (Aali 2026; Arampatzis and Arampatzis 2026). Humour knowledge was therefore distributed across model parameters, prompts, retrieved material and ranking components. Where it resides determines what a resource must document for the resulting system to be reproducible.

4.5. Evaluation of Generated Humour

Studies combined automatic, human and LLM-based evaluation, but these approaches measured different constructs and were rarely applied through a common protocol (Fig. 7). Funniness was the main outcome, but many studies also rated fluency, relevance, surprise, originality, style conformity, cultural appropriateness or offensiveness, so reported scores were seldom comparable across systems.

4.5.1. *Automatic Evaluation*

Classification and ranking measures were the most frequent automatic measures (Fig. 7A). Most estimated whether a trained classifier recognised an output as humorous (He and Mei 2025; Akbar et al. 2021), not whether an audience found it funny. The Elo and Bradley–Terry scores reported by shared-task papers were a different case, because they aggregate human pairwise votes (Castro et al. 2026). Reference-overlap metrics such as BLEU and ROUGE were rare (Wang et al. 2021; Chauhan et al. 2023), which is consistent with the open-endedness of humour. Semantic relevance, diversity, fluency and constraint checks each captured one property (Mishra 2025; Chen et al. 2024a; Abdulmujeeb and Aryal 2026), and no composite measure of generation quality emerged.

4.5.2. *Human Evaluation*

Human judgement remained the main evidence of experienced humour, used in 62 studies (Fig. 7B). Protocols ranged from Likert and binary ratings to pairwise comparison, best–worst selection, source identification and live-audience response (Hossain et al. 2020; Sakabe et al. 2026; Winters and Van der Stockt 2025). Evaluator pools ranged from professional comedians and trained annotators (Hong et al. 2026; Mirowski et al. 2024) to crowdworkers, students and target-country participants (Inácio and Oliveira 2025; Sarrof 2025; Xian et al. 2025). Shared-task system papers usually inherited the organisers’ evaluation and described its raters only briefly (Abdulmujeeb and Aryal 2026; Schechter, Barzel, and Liebeskind 2026). Agreement among evaluators was reported inconsistently (Hong et al. 2026; Tian, Sheth, and Peng 2022), and statistical testing appeared in only 29 studies (39.2%), so most system comparisons rest on aggregate means without an estimate of uncertainty.

4.5.3. *LLM-Based Evaluation*

LLMs judged final outputs in 15 studies (20.3%), using rubric scoring, pairwise preference or simulated audiences (Ravi et al. 2024; Sakabe et al. 2026; Tang 2026a; Zhang et al. 2024). In eight further roles, LLMs selected or ranked

Table 5. Resource characteristics and associated modelling choices. Rows are not mutually exclusive; individual systems often combine several configurations.

Resource configuration	Modelling choice and role of the resource	Example studies
Large paired or labelled corpora	Supervised training or fine-tuning. Pairs or labels teach humour transformations through model parameters	Mishra 2025; Li et al. 2023; Weller, Fulda, and Seppi 2020
Small task-specific collections	PEFT, controlled fine-tuning or few-shot prompting, and adapt pretrained models with limited data	Poobanchuen et al. 2025; Xian et al. 2025; Abbas and Abdul Rauf 2026; Chen 2026
Evaluation inputs without training data	Zero- or few-shot prompting and generate-and-select pipelines, and inputs act as constraints, humour knowledge stays in the pretrained model	Abdurmujeeb and Aryal 2026; Parthiban and R 2026; Schechter, Barzel, and Liebeskind 2026; Arampatzis and Arampatzis 2026
Image-caption, meme or GIF resources	Vision-conditioned decoders or multimodal LLMs and align visual context with textual incongruity	Zhong et al. 2024; Wang et al. 2021; Zhang et al. 2026; Baruah 2026
Dialogue, performance or interaction resources	Context-aware dialogue models, memory and multimodal delivery; represent history, speaker roles, audience response and prosody	Kiderle et al. 2025; Seki 2022; Winters and Van der Stockt 2025; Chauhan et al. 2023
Ratings, pairwise comparisons or preferences	Candidate ranking, reward modelling and preference optimisation; turn judgements into selection scores or rewards	Ravi et al. 2024; Tikhonov and Ivanov 2026; Zhang et al. 2024; Brüggemann and Hou 2026; Chen et al. 2024a
Lexical, factual or cultural knowledge	Retrieval augmentation, symbolic constraints or hybrid pipelines; supply word senses, current information and cultural evidence missing from model parameters	Sarrof 2025; Shen et al. 2026; Aali 2026; Zare et al. 2026
Synthetic instructions, labels or preference pairs	Distillation, synthetic supervision and model-based filtering also expand scarce human-authored data	Ravi et al. 2024; Arora, Arowolo, and Hoblitzell 2026; Brüggemann and Hou 2026; Chen 2026

candidates inside the generation pipeline (Shen et al. 2026; Bubus et al. 2026), and in others they supplied rewards or synthetic labels (Fig. 7C). Such pipelines shape the outputs and are not independent evaluations. Where LLM judgements were compared with human ratings, agreement was weak and model-specific biases appeared (Sakabe et al. 2026; Cao et al. 2025; Dogra et al. 2026). For example, Oogiri judges favoured novelty and rated even unrelated responses as funny, whereas human raters prioritised empathy (Sakabe et al. 2026).

4.5.4. Cultural Appropriateness and Safety

Only four studies evaluated cultural appropriateness directly, through target-community raters and explicit cultural criteria (Xian et al. 2025; Almasoud et al. 2026; Garimella et al. 2020; Shimozaki et al. 2024) (Fig. 7D). Eighteen addressed it indirectly by using native or target-audience raters without a cultural criterion, so they usually assumed cultural fit based on who the raters were rather than measuring it. Seven studies operationalised safety by rating offensiveness, stereotyping, toxicity, moral acceptability or suitability for sensitive audiences (Dogra et al. 2026; Almasoud et al. 2026; Teh et al. 2025; Yamane, Mori, and Harada 2021). Fifteen relied on preventive filtering, which controls content but does not show how audiences perceive the outputs that remain.

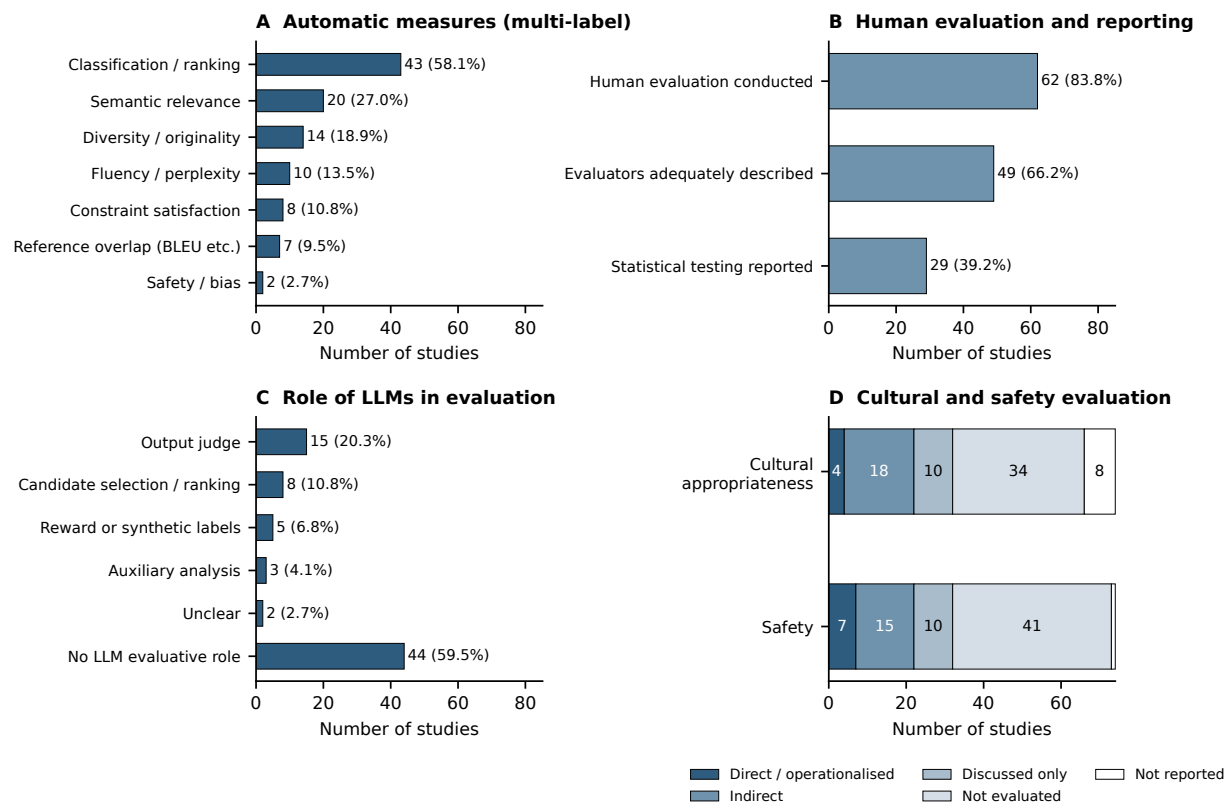


Figure 7. Evaluation of generated humour in the 74 included studies. (A) Automatic measures (multi-label). (B) Human evaluation, evaluator reporting and statistical testing. (C) Evaluative roles of LLMs (multi-label). (D) Level of cultural and safety evaluation.

5. Discussion

5.1. Principal Findings

The field is growing quickly, but its resource and evaluation foundations have not kept pace. The clearest change is the move from rule-based, statistical and task-specific neural systems to general-purpose LLMs, which took part in 54 of the 74 studies (73.0%), with prompting used in 55 (74.3%). This shift has widened what systems can do, including multilingual, multimodal and co-creative generation, and continues the trajectory described in earlier reviews (Amin and Burghardt 2020; Loakman, Thorne, and Lin 2025). However, general models have not removed the need for specialised resources. Humour knowledge is now spread across model parameters, prompts, retrieved material, constraints, preference data and ranking components. What a system can generate, and for whom it is funny, still depends on where these components come from and whom they represent. Progress therefore cannot be judged by model scale alone.

The second finding is that the main obstacle has shifted from dataset scarcity to lifecycle opacity. Earlier surveys pointed to a lack of datasets (Amin and Burghardt 2020). Today more data exist, but they are hard to find, cite and reuse. Resources were publicly available in only 21 studies (28.4%), and only one study named a licence for its data. The FAIR assessment confirms this at the level of the resources themselves: no released resource had a persistent identifier, three reported releases could not be located, and only two carried a named data licence (Section 4.2.3). Because systems now chain several sources and synthetic steps, reproducing a result means reconstructing the whole

chain of collection, transformation, annotation, prompting and evaluation, not downloading one corpus. In most cases the documentation needed to do this is missing.

These trends create a tension. Capable models can generate humour from a few examples or instructions, which reduces the need for large training sets. The same flexibility makes it easy to overlook whether prompts, retrieval sources and evaluation samples represent the intended humour form, language or audience. Shared benchmarks such as MWAHAHA, Oogiri-GO and the SemEval pun data make comparison possible (Castro et al. 2026; Zhong et al. 2024; Mittal, Tian, and Peng 2022), but a quarter of the corpus now shares one task design, which risks narrowing what counts as successful humour generation to the constraints that design encodes. The field needs both common benchmarks for comparison and locally grounded resources that capture linguistic, cultural and interactional variation.

5.2. Resource–Evaluation Misalignment

Across RQ1–RQ4, the central problem is a mismatch between what resources contain and what evaluations claim. Resources hold jokes, edited headlines, puns, captions or audience preferences, yet evaluations often claim to measure humour quality, creativity or appropriateness. Each common measure supports a narrower claim. A classifier trained on labelled jokes shows that an output resembles its positive class, not that an audience finds it funny. The problem is sharpest when the classifier and generator learn from the same collection, as when a joke classifier and a joke generator were both trained on the Short Jokes corpus (Akbar et al. 2021). Semantic similarity shows topical relevance, constraint checks show that required words appear, and diversity rewards variation; none of them measures humour appreciation. Evaluation is valid only when each claim is matched to a measure of that construct: automatic measures for fluency, relevance, diversity and constraints, and audience judgements for funniness.

Preference data create a related circularity. Ratings, contest votes and pairwise comparisons are valuable supervision for rankers and reward models (Tikhonov and Ivanov 2026; Zhang et al. 2024; Brüggemann and Hou 2026; Chen et al. 2024a). But when the same kind of signal is used both to train a model and to judge it, improvement may show alignment with those preferences rather than broader humour competence. Natural signals such as likes and votes also reflect platform visibility and community habits, not only funniness. The key questions are whose preferences were collected, under what conditions, for which humour forms, and whether the evaluation audience differs from the training data.

Synthetic data can close this loop entirely. Model-generated instructions, candidates, preference pairs and labels expand coverage cheaply (Ravi et al. 2024; Arora, Arowolo, and Hoblitzell 2026; Brüggemann and Hou 2026; Chen 2026), but they may carry the style and cultural assumptions of the model that produced them. When another LLM then judges the outputs, creator, generator and evaluator may share training data and biases, so consistent results may reflect agreement among models rather than with audiences. In one study, the same Llama-2 model produced the training data, supplied the preference feedback and judged the final outputs (Ravi et al. 2024). Synthetic resources are most informative when their provenance is separated from human-authored material and when evaluation includes independent human evidence. Otherwise, datasets grow in size while the range of humour judgements behind them stays narrow.

5.3. Linguistic and Cultural Limitations

The corpus extends beyond English to Chinese, Spanish, Japanese, Hindi, Portuguese, Thai, Arabic, Dutch and code-mixed Hindi–English, helped by shared tasks and multilingual models (Castro et al. 2026; Arampatzis and Arampatzis 2026). Yet English-only work still makes up half of the corpus, and multilingual generation, native-language resource construction and cultural grounding are different achievements. A model can produce fluent text in several languages while relying on English-centred prompts, translated examples or cultural associations absorbed during pretraining. Language is a carrier of culture but not a proxy for it: a national or linguistic label hides regional, generational and social variation, and bilingual audiences may read a joke through several cultural frames.

The evaluation evidence shows the same gap: direct cultural evaluation was rare, and in most multilingual work cultural fit was assumed from who the raters were rather than measured (Section 4.5.4). The strongest cases tied resources to identifiable traditions, such as Japanese Oogiri and Manzai or Chinese allegorical sayings (Section 4.3.3), but even these remain isolated resource families rather than parts of a balanced cross-cultural benchmark, echoing the coverage gaps reported for multimodal humour (Liang et al. 2026). The corpus therefore shows more clearly that systems can produce humour-formatted text in many languages than that they can produce humour appropriate for particular communities. Until resources record provenance, intended audience, language variety and humour mechanism, cross-lingual comparisons may reflect differences in resource availability or translation quality rather than in generation capability.

5.4. Evaluation Validity and Comparability

Human judgement remains the most direct evidence of experienced humour, but using humans does not by itself make an evaluation valid. Studies differed in sample size, rater expertise, language background, recruitment and aggregation, so two “funniness” scores may measure different experiences. Fewer than two in five studies reported statistical testing, and agreement statistics were inconsistent (Section 4.5.2). Low agreement need not mean a faulty instrument, because people genuinely differ in what they find funny. Without rater-level data, uncertainty estimates and a stated aggregation method, however, real audience variation cannot be told apart from noise. Reporting score distributions, subgroup patterns and effect sizes would be more informative than a single mean.

LLM judges add scale, not validity. Many studies used LLMs inside the pipeline to select candidates or supply rewards rather than to judge final outputs independently (Section 4.5.3), and where judges were compared with humans, they diverged or amplified biases (Sakabe et al. 2026; Cao et al. 2025; Dogra et al. 2026). Consistency across repeated calls shows procedural stability, not that the judge measures human funniness. LLM judges are most defensible as complementary tools for screening, rubric-based diagnosis and flagging outputs for human review, provided that the model, prompt, rubric and decoding settings are reported and agreement with a human sample from the relevant audience is established.

Safety needs the same audience-based evidence. Studies that measured safety-related outcomes remained a small minority, and most relied on filtering alone (Section 4.5.4). Filtering controls content, but it cannot show that the remaining outputs are acceptable to their audience, because the same joke can be harmless in one community and offensive in another. The evidence therefore supports a layered design rather than any single metric: automatic measures for formal properties, LLM judges for scalable screening, and humans from the intended audience for funniness, cultural fit and acceptability. Each layer must be independent of the signals used to optimise the system, and each study should state what was measured, by whom, under which conditions, and for which audience.

5.5. Evidence Gaps and Future Research Directions

The gaps identified across RQ1–RQ4 go beyond the number of humour datasets. They concern whether resources can be found and reused, whether systems keep a traceable link to the resources behind them, and whether evaluations support their claims across languages and cultures. Table 6 links each gap to its evidence and to a concrete recommendation.

6. Conclusion

This survey examined 74 computational humour-generation studies published between 2020 and 2026 to characterise the resources, construction practices, models and evaluation methods that support the field. Two connected findings emerge. First, the shift to general-purpose large language models has not removed the dependence on resources. Humour knowledge is now spread across model parameters, prompts, retrieved material, preference data and ranking components, and each of these depends on data whose origin and coverage shape what a system can generate and for whom it is funny. Second, the main obstacle has shifted from dataset scarcity to lifecycle opacity. More humour data

Table 6. Evidence gaps and recommended directions for computational humour generation ($N = 74$ studies; FAIR results refer to the 19 released resources in Section 4.2.3).

Dimension	Corpus evidence	Gap and consequence	Recommendation	Studies
Findability and licensing	Resources were public in 21 studies (28.4%), and one named a data licence. No released resource had a persistent identifier, and three could not be located (Fig. 4).	Access is mistaken for reusability. Results cannot be reproduced, and legal reuse is unclear.	Deposit versioned data with a DOI and a dataset card. State a data licence and the terms inherited from each source.	Zhang et al. 2024; Zhong et al. 2024; Sakabe et al. 2026; Winters and Van der Stockt 2025; Tang 2026b
Annotation transparency	Guidelines were reported in 10 studies (13.5%), agreement in 22 (29.7%), and both in only five.	Labels cannot be compared across resources. Real audience disagreement cannot be told apart from unreliable annotation.	Publish guidelines, rating instruments, rater recruitment and demographics, aggregation methods and agreement statistics.	Asapu et al. 2025; Tian, Sheth, and Peng 2022; Almasoud et al. 2026; Chen et al. 2025; Jentzsch and Kersting 2023
Linguistic coverage	English-only studies formed 51.4% of the corpus. Code-mixed generation appeared in two studies.	Claims of general capability may not transfer to low-resource languages, regional varieties or code-mixed communication.	Build community-led resources for under-represented languages. Collect native humour rather than translating it, and record the language variety.	Sarrof 2025; Asapu et al. 2025
Cultural grounding	Four studies evaluated culture directly, and 18 inferred it from rater identity. Grounded humour forms came from a few traditions.	Fluent target-language output is mistaken for cultural competence. Cross-lingual comparisons may reflect resource differences rather than capability.	Record provenance, intended audience and humour tradition. Evaluate cultural fit with raters who match the claimed audience, using explicit criteria.	Xian et al. 2025; Almasoud et al. 2026; Garimella et al. 2020; Shimozaki et al. 2024; Zhong et al. 2024; Xu 2024
Synthetic provenance and reproducibility	Model-generated data were common. Code was available in 26 studies (35.1%), and provenance was fully documented for 9 of 19 released resources.	Human and synthetic content are mixed. Biases pass from data creation to evaluation, and gains cannot be traced to pipeline components.	Label the provenance of each instance. Report teacher models, prompts and filters, and release retrieval indexes, decoding settings, seeds and selection components.	Ravi et al. 2024; Arora, Arowolo, and Hoblitzell 2026; Brüggemann and Hou 2026
Human and LLM evaluation	Human evaluation appeared in 62 studies, and statistical testing in 29 (39.2%). LLM judges were used in 15 and diverged from humans where compared.	Differences between systems may reflect rating design or judge preferences rather than humour quality.	Adopt a minimum reporting standard covering constructs, samples, uncertainty and agreement. Calibrate LLM judges against audience-matched human ratings.	Sakabe et al. 2026; Cao et al. 2025; Dogra et al. 2026
Safety	Seven studies measured safety-related outcomes, and 15 relied on filtering alone.	Outputs may pass fluency and funniness criteria yet be unsuitable for their audience.	Treat safety as an evaluation construct, with audience-specific acceptability ratings, subgroup analysis and reporting of problematic outputs.	Dogra et al. 2026; Almasoud et al. 2026; Teh et al. 2025

exist than ever, but they are hard to find, cite and reuse. Only one study named a licence for its data, and none of the 19 released resources had a persistent identifier. Annotation guidelines were reported in 13.5% of studies and statistical testing in 39.2%. These gaps matter most for multilingual and culturally grounded work: English-only research still made up half of the corpus, only four studies evaluated cultural appropriateness directly, and in 18 others cultural fit was assumed from who the raters were rather than measured. Progress therefore depends on documentation as much as on models. Resources should be deposited with persistent identifiers, dataset cards and explicit licences. Annotation and evaluation protocols should be published in full, synthetic data should be labelled as such, and the roles of human and LLM evaluators should be kept distinct. Resources for under-represented languages and humour traditions should

be built with the communities they represent. The field does not need a single universal measure of humour. It needs evidence that states clearly what was measured, by whom or by what, under which conditions and for which audience. Only then can genuine advances in humour generation be distinguished from gains that reflect model scale, benchmark conventions or agreement among automated components.

Acknowledgements

The first author gratefully acknowledges King Mongkut's University of Technology Thonburi for the Petchra Pra Jom Klao Ph.D. Research Scholarship, and the Department of Computer Engineering, Faculty of Engineering, for supporting his doctoral studies.

Declarations

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. S.H. Imam is supported by the Petchra Pra Jom Klao Ph.D. Research Scholarship of King Mongkut's University of Technology Thonburi (Grant No. ****).

Competing Interests

Two authors (N.M. and B.S.) are co-authors of one included study (Poobanchuen et al. 2025). Its screening, data extraction and coding were carried out by the first author under the same criteria applied to all other studies. The authors declare no other competing interests.

Use of Artificial Intelligence

The authors used Claude (Anthropic) and Grammarly to assist with language editing and polishing. All AI-assisted outputs were reviewed, verified, and, where necessary, corrected by the authors. The authors take full responsibility for the content of this article, including its accuracy, integrity and the interpretation of the evidence.

Author Contributions

Sukairaj Hafiz Imam: Conceptualisation, Methodology, Investigation, Data curation, Formal analysis, Visualisation, Writing (original draft), Writing (review and editing).

Nuttanart Muansuwan: Conceptualisation, Methodology, Validation, Supervision, Writing (review and editing).

Booncharoen Sirinaovakul: Conceptualisation, Supervision, Writing (review and editing).

All authors read and approved the final manuscript and agree to be accountable for all aspects of the work.

References

- Aali, Y.: 2026, yasaminal at Semeval2026: Constraint-Aware Humor Generation with Knowledge Graph Guidance. In: Kochmar, E., Ghosh, D., North, K., Komachi, M., Zampieri, M. (eds.) *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, Association for Computational Linguistics, San Diego, California, USA, 2409. ISBN 979-8-89176-414-9. DOI. <https://aclanthology.org/2026.semeval-1.303/>.
- Abbas, I., Abdul Rauf, S.: 2026, SLPGFJWUInsa at SemEval-2026 Task 1: Enhancing Linguistic Creativity for English Text-Based Humor. In: Kochmar, E., Ghosh, D., North, K., Komachi, M., Zampieri, M. (eds.) *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, Association for Computational Linguistics, San Diego, California, USA, 1832. ISBN 979-8-89176-414-9. DOI. <https://aclanthology.org/2026.semeval-1.231/>.
- Abdulmujeeb, L., Aryal, S.: 2026, Howard University-AI4PC at SemEval-2026 Task 1: Exploring Prompt Strategies for Automatic Humor Generation. In: Kochmar, E., Ghosh, D., North, K., Komachi, M., Zampieri, M. (eds.) *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, Association for Computational Linguistics, San Diego, California, USA, 2527. ISBN 979-8-89176-414-9. DOI. <https://aclanthology.org/2026.semeval-1.319/>.
- Agrawal, J., Mamidi, R.: 2026, Does Bigger Mean Funnier? Evaluating Humor Generation Across the Qwen3 Model Family. In: Miller, T., Amir, O., Rayz, J., Dong, T., Hempelmann, C.F. (eds.) *Proceedings of the 2nd Workshop on Computational Humor (CHum 2026)*, Association for Computational Linguistics, Online, 81. ISBN 979-8-89176-431-6. DOI. <https://aclanthology.org/2026.chum-1.7/>.
- Akbar, N.A., Darmayanti, I., Fati, S.M., Muneer, A.: 2021, Deep learning of a pre-trained language model's joke classifier using GPT-2. *Journal of Hunan University Natural Sciences* **48**.

- Almasoud, A.M., Al-Khalifa, H., Alqifari, R.F., Alangari, N., Albahlal, M.M.: 2026, ARHAHA 2026: The Shared Task on Arabic Humor Automatic Generation. In: Al-Khalifa, H., El-Haj, M., Ezzini, S. (eds.) *The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks*, Association for Computational Linguistics, Palma, Mallorca (Spain), 141. DOI. <https://aclanthology.org/2026.osact-1.16/>.
- Amin, M., Burghardt, M.: 2020, A Survey on Approaches to Computational Humor Generation. In: DeGaetano, S., Kazantseva, A., Reiter, N., Szpakowicz, S. (eds.) *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, International Committee on Computational Linguistics, Online, 29. <https://aclanthology.org/2020.latechclfl-1.4/>.
- Arampatzis, G., Arampatzis, A.: 2026, DUTH at SemEval-2026 Task 1: Prompt-Based Zero-Shot Large Language Models for Constrained Multilingual Humor Generation. In: Kochmar, E., Ghosh, D., North, K., Komachi, M., Zampieri, M. (eds.) *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, Association for Computational Linguistics, San Diego, California, USA, 579. ISBN 979-8-89176-414-9. DOI. <https://aclanthology.org/2026.semeval-1.83/>.
- Arora, U., Arowolo, M.O., Hoblitzell, A.: 2026, BAHABA at SemEval-2026 Task 1: Multi-Style Humor Generation via Comedian-Inspired Prompting and LLM Self-Assessment. In: *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, Association for Computational Linguistics, 785.
- Asapu, L., Kodali, P., Dua, A., Rajesh Kavitha, K., Shrivastava, M.: 2025, Bridging Laughter Across Languages: Generation of Hindi-English Code-mixed Puns. In: Hempelmann, C.F., Rayz, J., Dong, T., Miller, T. (eds.) *Proceedings of the 1st Workshop on Computational Humor (CHum)*, Association for Computational Linguistics, Online, 32. <https://aclanthology.org/2025.chum-1.5/>.
- Baruah, A.: 2026, ABARUAH at SemEval-2026 Task 1: Leveraging High-Resolution VLMs and Reasoning LLMs for Multimodal Humor Generation. In: Kochmar, E., Ghosh, D., North, K., Komachi, M., Zampieri, M. (eds.) *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, Association for Computational Linguistics, San Diego, California, USA, 3536. ISBN 979-8-89176-414-9. DOI. <https://aclanthology.org/2026.semeval-1.436/>.
- Brüggenmann, K., Hou, L.: 2026, Team TüLK at SemEval-2026 Task 1: Humor Generation with Qwen and Group Relative Policy Optimization. In: Kochmar, E., Ghosh, D., North, K., Komachi, M., Zampieri, M. (eds.) *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, Association for Computational Linguistics, San Diego, California, USA, 463. ISBN 979-8-89176-414-9. DOI. <https://aclanthology.org/2026.semeval-1.67/>.
- Bubus, B., Soyal, N., Schmitt, V., Feldhus, N., Solopova, V.: 2026, XplainNLP at SemEval-2026 Task 1: BVAHAHA - Benign Violation Algorithm for Humor and Harmless Absurdity. In: Kochmar, E., Ghosh, D., North, K., Komachi, M., Zampieri, M. (eds.) *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, Association for Computational Linguistics, San Diego, California, USA, 1501. ISBN 979-8-89176-414-9. DOI. <https://aclanthology.org/2026.semeval-1.195/>.
- Camacho Gonzalez, A.T., Cruz, X., Godínez-Aldana, N., Palacios-Patiño, L., Rangel, R., Meza Ruiz, I.V.: 2026, L52+-IIMAS-UNAM at SemEval-2026 Task 1 (MWAHAHA): Joke Selection Through a Multi-Stage Prompt-Engineering and Heuristic Pipeline. In: Kochmar, E., Ghosh, D., North, K., Komachi, M., Zampieri, M. (eds.) *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, Association for Computational Linguistics, San Diego, California, USA, 2301. ISBN 979-8-89176-414-9. DOI. <https://aclanthology.org/2026.semeval-1.291/>.
- Cao, Y., Cao, J., Hou, Y., Ji, L.-J.: 2025, How humorous is AI? Exploring ChatGPT's role in humor generation and human-AI interaction. *Computers in Human Behavior Reports* **20**, 100807. DOI. <https://www.sciencedirect.com/science/article/pii/S2451958825002222>.
- Castro, S., Chiruzzo, L., Góngora, S., Deng, N., Rahili, S., Sastre, I., Rosá, A., Amoroso, V., Rey, G., Moncecchi, G., Meaney, J.A., Prada, J.J., Mihalcea, R.: 2026, SemEval-2026 Task 1: MWAHAHA, Models Write Automatic Humor And Humans Annotate. In: Kochmar, E., Ghosh, D., North, K., Komachi, M., Zampieri, M. (eds.) *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, Association for Computational Linguistics, San Diego, California, USA, 3797. ISBN 979-8-89176-414-9. DOI. <https://aclanthology.org/2026.semeval-1.454/>.
- Chauhan, D.S., Singh, G.V., Ekbal, A., Bhattacharyya, P.: 2023, MHaDiG: A Multilingual Humor-aided Multi-party Dialogue Generation in multimodal conversational setting. *Knowledge-Based Systems* **278**, 110840. DOI. <https://www.sciencedirect.com/science/article/pii/S0950705123005907>.
- Chen, C.: 2026, deepgpt at SemEval-2026 Task 1: A Chinese Humor Generation System via Instruction-Masked QLoRA and Reverse Constraint Data Mixing. In: Kochmar, E., Ghosh, D., North, K., Komachi, M., Zampieri, M. (eds.) *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, Association for Computational Linguistics, San Diego, California, USA, 1116. ISBN 979-8-89176-414-9. DOI. <https://aclanthology.org/2026.semeval-1.152/>.
- Chen, M., Wang, C., Hui, L., Wu, Y., Sha, Y.: 2025, Can MLLMs Tell Jokes Based on Images? A Visual Context-Driven Humor Generation Framework. In: *2025 IEEE International Conference on Multimedia and Expo (ICME)*, 1. DOI.
- Chen, Y., Yang, C., Hu, T., Chen, X., Lan, M., Cai, L., Zhuang, X., Lin, X., Lu, X., Zhou, A.: 2024a, Are U a Joke Master? Pun Generation via Multi-Stage Curriculum Learning towards a Humor LLM. In: Ku, L.-W., Martins, A., Srikumar, V. (eds.) *Findings of the Association for Computational Linguistics: ACL 2024*, Association for Computational Linguistics, Bangkok, Thailand, 878. DOI. <https://aclanthology.org/2024.findings-acl.51/>.
- Chen, Y., Yan, S., Zhu, Z., Li, Z., Xiao, Y.: 2024b, XMeCap: Meme Caption Generation with Sub-Image Adaptability. In: *Proceedings of the 32nd ACM International Conference on Multimedia, MM '24*, Association for Computing Machinery, New York, NY, USA, 3352–3361. ISBN 9798400706868. DOI. <https://doi.org/10.1145/3664647.3681332>.
- Dogra, A., Ghosal, S.S., Deshpande, A., Kalyan, A., Manocha, D.: 2026, Engagement Undermines Safety: How Stereotypes and Toxicity Shape Humor in Language Models. In: Demberg, V., Inui, K., Marquez, L. (eds.) *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Rabat, Morocco, 7971. ISBN

- 979-8-89176-380-7. DOI: <https://aclanthology.org/2026.eacl-long.373/>.
- Eskandari, S., Mousavi, S.A., Rahimi, A., Pouresmaeil, M., Vitaggio, M., Savelli, C., Coppola, R., Giobergia, F.: 2026, MINDS at SemEval-2026 Task 1: Enhancing Humor Generation through RAG and Synthetic DPO Alignment. In: Kochmar, E., Ghosh, D., North, K., Komachi, M., Zampieri, M. (eds.) *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, Association for Computational Linguistics, San Diego, California, USA, 2967. ISBN 979-8-89176-414-9. DOI: <https://aclanthology.org/2026.semeval-1.372/>.
- Fischer, K., Zyłowski, T., Wölfel, M.: 2025, Making Knowledge Acquisition Fun! Generating Content-Related Humor for Learning with Large Language Models. In: *2025 International Conference on Education Technology and Computers (ICETC)*, 251. DOI.
- Garimella, A., Banea, C., Hossain, N., Mihalcea, R.: 2020, “Judge me by my size (noun), do you?” YodaLib: A Demographic-Aware Humor Generation Framework. In: Scott, D., Bel, N., Zong, C. (eds.) *Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2814. DOI: <https://aclanthology.org/2020.coling-main.253/>.
- Gorenz, D., Schwarz, N.: 2024, How funny is ChatGPT? A comparison of human-and AI-produced jokes. *Plos one* **19**, e0305364.
- He, J., Mei, A.: 2025, Advancing Computational Humor: LLaMa-3 Based Generation with DistilBert Evaluation Framework. In: *ITM Web of Conferences* **70**, 03024. EDP Sciences.
- Hong, S., Li, L., Rong, E.Z., Shen, C., Lu, Z.: 2026, Multi-Agent Comedy Club: Investigating Community Discussion Effects on LLM Humor Generation. In: Liakata, M., Moreira, V.P., Zhang, J., Jurgens, D. (eds.) *Findings of the Association for Computational Linguistics: ACL 2026*, Association for Computational Linguistics, San Diego, California, United States, 2981. ISBN 979-8-89176-395-1. DOI.
- Hossain, N., Krumm, J., Sajed, T., Kautz, H.: 2020, Stimulating Creativity with FunLines: A Case Study of Humor Generation in Headlines. In: Celikyilmaz, A., Wen, T.-H. (eds.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, Online, 256. DOI: <https://aclanthology.org/2020.acl-demos.28/>.
- Inácio, M.L., Oliveira, H.G.: 2025, A full pipeline for context-aware pun generation. In: *Proceedings of the 16th International Conference on Computational Creativity*, 1.
- Jain, S.: 2025, JokeEval: Are the Jokes Funny? Review of Computational Evaluation Techniques to Improve Joke Generation. In: *KDD 2025 Workshop on Evaluation and Trustworthiness of Agentic and Generative AI Models*, Toronto, Canada. <https://www.amazon.science/publications/jokeeval-are-the-jokes-funny-review-of-computational-evaluation-techniques-to-improve-joke-generation>.
- Jentsch, S., Kersting, K.: 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*, 325.
- Kalloniatis, A., Adamidis, P.: 2024, Computational humor recognition: a systematic literature review. *Artificial Intelligence Review* **58**, 43. DOI: <https://doi.org/10.1007/s10462-024-11043-3>.
- Kariyawasam, H., Niwarthana, A., Palmer, A., Kay, J., Withana, A.: 2024, Appropriate Incongruity Driven Human-AI Collaborative Tool to Assist Novices in Humorous Content Generation. In: *Proceedings of the 29th International Conference on Intelligent User Interfaces, IUI '24*, Association for Computing Machinery, New York, NY, USA, 650–659. ISBN 9798400705083. DOI: <https://doi.org/10.1145/3640543.3645161>.
- Kiderle, T., Dobre, C.G., Nasir, J., Gonzalez Diaz, C., Ritschel, H., Klein, S., Mertes, S., André, E.: 2025, Multimodal Generation of Contextualized Jokes for a Real-Time Virtual Character. In: *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents, IVA '25*, Association for Computing Machinery, New York, NY, USA. ISBN 9798400715082. DOI: <https://doi.org/10.1145/3717511.3749299>.
- Kongqiang, W., Peng, Z., Qingli, T.: 2026, wangkongqiang at SemEval-2026 Task 1: MWAHAHA- Competition on Humor Generation. In: Kochmar, E., Ghosh, D., North, K., Komachi, M., Zampieri, M. (eds.) *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, Association for Computational Linguistics, San Diego, California, USA, 80. ISBN 979-8-89176-414-9. DOI: <https://aclanthology.org/2026.semeval-1.12/>.
- Lemmens, J., De Marez, V.: 2026, Computational Humor Modeling: A Survey on the State of the Art. *ACM Comput. Surv.* **58**. DOI: <https://doi.org/10.1145/3778357>.
- Li, R., Sun, S., Elhoseiny, M., Torr, P.: 2023, OxfordTVG-HIC: Can Machine Make Humorous Captions from Images? In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 20236. DOI.
- Liang, T., Hu, Z., Liu, D., Li, J., Yin, Y.: 2026, *Computational Humor with Multimodal LLMs: Methods, Datasets, Evaluation, and Challenges*. <https://arxiv.org/abs/2607.19011>.
- Liu, L., Ibrahim, R., Ismail, N.A., Chan, W.H.: 2025, Humor Style Transfer for Virtual AI Teachers Using StyleGPT4o for Comedian Guided Script Rewriting in Virtual Educational Environments. In: *2025 5th International Conference on Computer Graphics, Image and Virtualization (ICCGIV)*, 166. DOI.
- Loakman, T., Thorne, W., Lin, C.: 2025, Who’s Laughing Now? An Overview of Computational Humour Generation and Explanation. In: Flek, L., Narayan, S., Phng, L.H., Pei, J. (eds.) *Proceedings of the 18th International Natural Language Generation Conference*, Association for Computational Linguistics, Hanoi, Vietnam, 780. <https://aclanthology.org/2025.inlg-main.45/>.
- Login, N., Baranov, A., Braslavski, P.: 2022, Jokingbird: Funny Headline Generation for News. In: *Analysis of Images, Social Networks and Texts: 10th International Conference, AIST 2021, Tbilisi, Georgia, December 16–18, 2021, Revised Selected Papers*, Springer, Cham, 1622–1636. ISBN 978-3-031-16499-6. DOI.
- Mirowski, P., Love, J., Mathewson, K., Mohamed, S.: 2024, A Robot Walks into a Bar: Can Language Models Serve as Creativity SupportTools for Comedy? An Evaluation of LLMs’ Humour Alignment with Comedians. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24*, Association for Computing Machinery, New York, NY, USA, 1622–1636. ISBN 9798400704505.

- DOI. <https://doi.org/10.1145/3630106.3658993>.
- Mishra, A.: 2025, Investigating Structured Fine-Tuning for Sarcasm Generation in Small Language Models. Available at SSRN 5970754. DOI.
- Mittal, A., Tian, Y., Peng, N.: 2022, AmbiPun: Generating Humorous Puns with Ambiguous Context. In: Carpuat, M., de Marneffe, M.-C., Meza Ruiz, I.V. (eds.) *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Seattle, United States, 1053. DOI. <https://aclanthology.org/2022.naacl-main.77/>.
- Moreira, A., Lima Inácio, M., Gonçalo Oliveira, H.: 2025, Text Rewriting with Transformers for Humor Generation in Portuguese. In: *Progress in Artificial Intelligence: 24th EPIA Conference on Artificial Intelligence, EPIA 2025, Faro, Portugal, October 1–3, 2025, Proceedings, Part II*, Springer-Verlag, Berlin, Heidelberg, 435–447. ISBN 978-3-032-05178-3. DOI. https://doi.org/10.1007/978-3-032-05179-0_33.
- Önal, U., Sariel, S., Sezgin, M., Akleman, D., Akleman, E.: 2026, Hybrid humor: Investigating AI's potential in cartoon caption writing. *AI Mag.* **47**. DOI. <https://doi.org/10.1002/aaai.70064>.
- Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., et al.: 2022, The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Revista panamericana de salud publica* **46**, e112.
- Parthiban, H., R, P.: 2026, Humor Generation – Text-based Humor Generation (English). In: Kochmar, E., Ghosh, D., North, K., Komachi, M., Zampieri, M. (eds.) *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, Association for Computational Linguistics, San Diego, California, USA, 735. ISBN 979-8-89176-414-9. DOI. <https://aclanthology.org/2026.semEval-1.104/>.
- Poobanchuen, P., Chaithong, S., Limseelo, C., Suwannahong, K., Muansuwan, N., Sirinaovakul, B.: 2025, Thai Humor Generation by Small Language Models. In: *2025 22nd International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 1. DOI.
- Quan, K., Ramakrishnan, P., Chin, J.: 2025, Can AI Take a Joke—Or Make One? A Study of Humor Generation and Recognition in LLMs. In: *Proceedings of the 2025 Conference on Creativity and Cognition, CC '25*, Association for Computing Machinery, New York, NY, USA, 431–437. ISBN 9798400712890. DOI. <https://doi.org/10.1145/3698061.3734388>.
- Ravi, S., Huber, P., Shrivastava, A., Shwartz, V., Einolghozati, A.: 2024, Small But Funny: A Feedback-Driven Approach to Humor Distillation. In: Ku, L.-W., Martins, A., Srikumar, V. (eds.) *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Bangkok, Thailand, 13078. DOI. <https://aclanthology.org/2024.acl-long.706/>.
- RDA FAIR Data Maturity Model Working Group: 2020, FAIR Data Maturity Model: Specification and Guidelines. RDA recommendation, Research Data Alliance. DOI.
- Ren, C., Guo, Z., Zhang, P., Gao, Y.: 2024, Humor detection using deep learning in 10 years: A survey. *Métodos numéricos para cálculo y diseño en ingeniería: Revista internacional* **40**, 1.
- Sakabe, R., Kim, H., Hirasawa, T., Komachi, M.: 2026, Assessing the capabilities of llms in humor: A multi-dimensional analysis of oogiri generation and evaluation. In: *Proceedings of the AAAI Conference on Artificial Intelligence* **40**, 32867.
- Sanober, Z., Samreen, S.A., Nishat, Z., Pasha, M., Farooq, M.U., Tabassum, N.: 2025, Edgy Humor: Image-Based Meme Generation Using Deep Learning. In: *2025 IEEE International Conference for Women in Innovation, Technology Entrepreneurship (ICWITE)*, 1. DOI.
- Sarraf, Y.R.: 2025, Homophonic Pun Generation in Code Mixed Hindi English. In: Hempelmann, C.F., Rayz, J., Dong, T., Miller, T. (eds.) *Proceedings of the 1st Workshop on Computational Humor (CHum)*, Association for Computational Linguistics, Online, 23. <https://aclanthology.org/2025.chum-1.4/>.
- Schechter, B., Barzel, S., Liebeskind, C.: 2026, JCT at SemEval-2026 Task 1: Let the Best Joke Win - A Generate - and-Rank Approach to Constrained Humor. In: Kochmar, E., Ghosh, D., North, K., Komachi, M., Zampieri, M. (eds.) *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, Association for Computational Linguistics, San Diego, California, USA, 89. ISBN 979-8-89176-414-9. DOI. <https://aclanthology.org/2026.semEval-1.13/>.
- Seki, Y.: 2022, Dialogue System to Make Humor by Tatoo-Tsukkomi. In: *2022 Tenth International Symposium on Computing and Networking Workshops (CANDARW)*, 147. DOI.
- Shen, W., Huang, L., He, J., Li, L., Zhang, J.: 2026, ICT-NLP at SemEval-2026 Task 1: Humor Generation via RAG-based Augmentation and Multi-LLM Internal-External Voting. In: Kochmar, E., Ghosh, D., North, K., Komachi, M., Zampieri, M. (eds.) *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, Association for Computational Linguistics, San Diego, California, USA, 1965. ISBN 979-8-89176-414-9. DOI. <https://aclanthology.org/2026.semEval-1.247/>.
- Shimozaki, A., Tsuge, Y., Kitamura, T., Umetani, T., Nadamoto, A.: 2024, Comparative Analysis with Multiple Large-Scale Language Models for Automatic Generation of Funny Dialogues. In: *Database Systems for Advanced Applications. DASFAA 2024 International Workshops: BDMS, GDMA, BDQM and ERDSE, Gifu, Japan, July 2-5, 2024, Proceedings*, Springer-Verlag, Berlin, Heidelberg, 187–202. ISBN 978-981-96-0913-0. DOI.
- Su, Y.: 2022, Computer-generated Humour Based on GPT-2. In: *2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA)*, 890. DOI.
- Suppa, M., Ondrejová, V., Ganajová, L., Karetka, G., Skala, D.: 2026, RAGthoven at SemEval-2026 Task 1: A Multi-Stage Pipeline Walks Into a Benchmark and Barely Clears the Bar. In: Kochmar, E., Ghosh, D., North, K., Komachi, M., Zampieri, M. (eds.) *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, Association for Computational Linguistics, San Diego, California, USA, 3343. ISBN 979-8-89176-414-9. DOI. <https://aclanthology.org/2026.semEval-1.416/>.
- Tang, M.: 2026a, Are LLMs qualified evaluators and generators of one-liner jokes. *Neural Networks* **202**, 109077. DOI.

- <https://www.sciencedirect.com/science/article/pii/S089360802600537X>.
- Tang, M.: 2026b, A multi-modal framework for generating and evaluating diverse jokes using large language models. *Engineering Applications of Artificial Intelligence* **179**, 115206. DOI. <https://www.sciencedirect.com/science/article/pii/S0952197626014909>.
- Teh, A.T.-R., Toh, X.Y., Seow, C.X., Rakib, T.b.A., Soon, L.-K.: 2025, Personalising AI-Generated Jokes: an Empirical Study. In: *TENCON 2025 - 2025 IEEE Region 10 Conference (TENCON)*, 229. DOI.
- Tian, Y., Sheth, D., Peng, N.: 2022, A Unified Framework for Pun Generation with Humor Principles. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2022*, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3253. DOI. <https://aclanthology.org/2022.findings-emnlp.237/>.
- Tikhonov, A., Ivanov, A.: 2026, Imfaoooo at SemEval-2026 Task 1: Humor Is an Audience. Preference Modeling for Constrained Humor Generation. In: Kochmar, E., Ghosh, D., North, K., Komachi, M., Zampieri, M. (eds.) *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, Association for Computational Linguistics, San Diego, California, USA, 101. ISBN 979-8-89176-414-9. DOI. <https://aclanthology.org/2026.semeval-1.15/>.
- Wang, L., Zhang, Q., Kim, Y., Wu, R., Jin, H., Deng, H., Luo, P., Kim, C.-H.: 2021, Automatic Chinese Meme Generation Using Deep Neural Networks. *IEEE Access* **9**, 152657. DOI.
- Weller, O., Fulda, N., Seppi, K.: 2020, Can Humor Prediction Datasets be used for Humor Generation? Humorous Headline Generation via Style Transfer. In: Klebanov, B.B., Shutova, E., Lichtenstein, P., Muresan, S., Wee, C., Feldman, A., Ghosh, D. (eds.) *Proceedings of the Second Workshop on Figurative Language Processing*, Association for Computational Linguistics, Online, 186. DOI. <https://aclanthology.org/2020.figlang-1.25/>.
- Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., et al.: 2016, The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018. DOI.
- Winters, T., Van der Stockt, S.: 2025, Evaluating humor generation in an improvisational comedy setting. *Computational Linguistics in the Netherlands Journal* **14**, 505.
- Wu, Z., Weber, T., Müller, F.: 2025, One Does Not Simply Meme Alone: Evaluating Co-Creativity Between LLMs and Humans in the Generation of Humor. In: *Proceedings of the 30th International Conference on Intelligent User Interfaces, IUI '25*, Association for Computing Machinery, New York, NY, USA, 1082–1092. ISBN 9798400713064. DOI. <https://doi.org/10.1145/3708359.3712094>.
- Xian, A.F.C., Ting, N.C., Cheng, A.K.S., Tan, W.Y., Chanthran, M.R., Soon, L.-K., Lee, M.: 2025, Laughing Across Borders: A Culturally-Aware Joke Generator for Asian Regions. In: *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 921. DOI.
- Xu, R.: 2024, Exploring Chinese Humor Generation: A Study on Two-part Allegorical Sayings. In: *2024 International Joint Conference on Neural Networks (IJCNN)*, 1. DOI.
- Yamane, H.: 2024, Generic Joke Generation with Moral Constraints. In: Wand, M., Malinovská, K., Schmidhuber, J., Tetko, I.V. (eds.) *Artificial Neural Networks and Machine Learning – ICANN 2024*, Springer Nature Switzerland, Cham, 340. ISBN 978-3-031-72344-5.
- Yamane, H., Mori, Y., Harada, T.: 2021, Humor meets morality: Joke generation based on moral judgement. *Information Processing & Management* **58**, 102520. DOI.
- Yoshikawa, T., Iwakura, R.: 2020, Study on Development of Humor Discriminator for Dialogue System. *Journal of Advanced Computational Intelligence and Intelligent Informatics* **24**, 422–435. DOI. <http://dx.doi.org/10.20965/jaciii.2020.p0422>.
- Zangari, A., Marcuzzo, M., Albarelli, A., Pilehvar, M.T., Camacho-Collados, J.: 2025, Pun Unintended: LLMs and the Illusion of Humor Understanding. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Suzhou, China, 27936. ISBN 979-8-89176-332-6. DOI. <https://aclanthology.org/2025.emnlp-main.1419/>.
- Zare, M.E., Abbasi, T., Veisi, H., Ala, S., Naderi, H.: 2026, YNAAZ at SemEval-2026 Task 1: Bridging the Semantic-Visual Gap: Multimodal Humor Generation. In: Kochmar, E., Ghosh, D., North, K., Komachi, M., Zampieri, M. (eds.) *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, Association for Computational Linguistics, San Diego, California, USA, 1308. ISBN 979-8-89176-414-9. DOI. <https://aclanthology.org/2026.semeval-1.171/>.
- Zeng, J., Yang, L., Kang, J., Diao, Y., Yang, Z., Lin, H.: 2024, “Barking up the Right Tree”, a GAN-Based Pun Generation Model through Semantic Pruning. In: Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, ELRA and ICCL, Torino, Italia, 2119. <https://aclanthology.org/2024.lrec-main.191/>.
- Zhai, C., Wibowo, S.: 2023, A WGAN-Based Dialogue System for Embedding Humor, Empathy, and Cultural Aspects in Education. *IEEE Access* **11**, 71940. DOI.
- Zhang, J., Jain, L., Guo, Y., Chen, J., Zhou, K.L., Suresh, S., Wagenmaker, A., Sievert, S., Rogers, T., Jamieson, K., et al.: 2024, Humor in ai: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning. *Advances in Neural Information Processing Systems* **37**, 125264.
- Zhang, J., Luo, S., Zhang, R., Su, Q.: 2026, Humorchain: Theory-guided multi-stage reasoning for interpretable multimodal humor generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19176.
- Zhong, S., Huang, Z., Gao, S., Wen, W., Lin, L., Zitnik, M., Zhou, P.: 2024, Let’s think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13246. IEEE.