Who's Laughing Now? An Overview of Computational Humour Generation and Explanation

Tyler Loakman[1], William Thorne[1], Chenghua Lin[2]
[1]Department of Computer Science, The University of Sheffield, UK
[2]Department of Computer Science, The University of Manchester, UK
tcloakman1@sheffield.ac.uk
wthorne1@sheffield.ac.uk
chenghua.lin@manchester.ac.uk

Abstract

The creation and perception of humour is a fundamental human trait, positioning its computational understanding as one of the most challenging tasks in natural language processing (NLP). As an abstract, creative, and frequently context-dependent construct, humour requires extensive reasoning to understand and create, making it a pertinent task for assessing the common-sense knowledge and reasoning abilities of modern large language models (LLMs). In this work, we survey the landscape of computational humour as it pertains to the generative tasks of creation and explanation. We observe that, despite the task of understanding humour bearing all the hallmarks of a foundational NLP task, work on generating and explaining humour beyond puns remains sparse, while state-of-the-art models continue to fall short of human capabilities. We bookend our literature survey by motivating the importance of computational humour processing as a subdiscipline of NLP and presenting an extensive discussion of future directions for research in the area that takes into account the subjective and ethically ambiguous nature of humour.

1 Introduction

Humour serves as a foundational element of human communication, acting as a way through which to express emotion, build interpersonal relationships, and experience levity and entertainment (Ritschel and André, 2018). However, humour may arise from myriad sources (Dynel, 2009), including simple, innocuous wordplay such as puns, all the way to deeply contextualised topical references that require layered reasoning to both create and interpret (Highfield, 2015; Laineste, 2002). The position of humour as a distinctly human experience, in addition to the challenges its processing presents, even to humans (Bell and Attardo, 2010; Mak and Carpenter, 2007; Wierzbicki and Young, 1978), makes generative computational humour tasks such as joke creation and explanation a formidable domain for analysing the common sense reasoning capabilities of modern large language models (LLMs).

1.1 Existing Surveys

Several surveys have examined different aspects of computational humour. Early foundational work by Ritchie (2001) provides a comprehensive overview of the emerging field of computational humour. More recent reviews have focused on specific subdomains of computational humour. Ramakristanaiah et al. (2021) and Kalloniatis and Adamidis (2024) survey humour recognition and detection, while Kenneth et al. (2024) review humour style classification. Ganganwar et al. (2024) explore sarcasm and humour detection in code-mixed Hindi, and Nijholt et al. (2017) cover the unique domain of humour in human-computer interaction. The most comparable works to ours are Amin and Burghardt (2020), who present a survey of text-based computational humour generation, and Nguyen and Ng (2024), who survey works on meme understanding. Our survey provides an up-to-date account of the field in the age of LLMs, an additional focus on humour explanation, and suggestions for future work based on real-world ethical considerations, rather than solely technical innovations.

1.2 Survey Outline

This survey is structured around the two primary generative tasks in computational humour. In §3 we explore humour generation, broken down by humour type, whilst in §4 we explore humour explanation, broken down into explanation through classification (see §4.1) and natural language explanation (see §4.2). Finally, in §5 we provide future directions for research in generative computational humour, taking into account the complex ethics of a potentially offensive language form, and the nature of generating creative text.

2 The Role of Humour Research in NLP

Significant effort and resources are currently being invested into enhancing the reasoning capabilities

Proceedings of the 18th International Natural Language Generation Conference, pages 780–794
October 29–November 2, 2025.

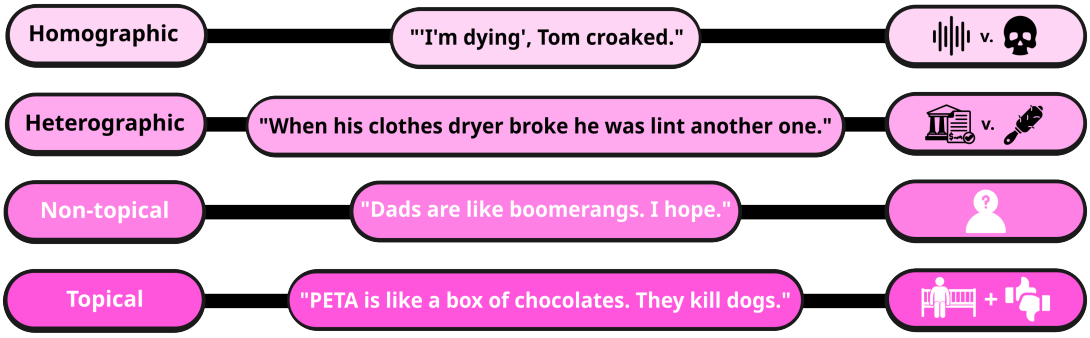

Figure 1: Examples of 4 broad textual joke types from Loakman et al. (2025b). Homographic and Heterographic refer to types of puns, whilst Topical and Non-Topical relate to incongruity-based humour, themed around common sense and contemporary news, respectively. The *homographic* pun exploits the dual meaning of "croaked" as both a style of speech and a euphemism for dying; the *heterographic* pun relies on the phonetic similarity between "lint" and "leant"; the *non-topical* joke plays on the trope of an absentee father (not) returning like a boomerang; and the *topical* joke refers to the animal welfare organisation PETA's high euthanasia rates and a reference to the movie Forrest Gump. Each joke type plays on polysemy, phonetics, social constructs, and esoteric knowledge, respectively.

of language models (Wu et al., 2024b; Yin et al., 2024; Servantez et al., 2024). Reasoning or thinking models embed established prompting techniques such as *chain-of-thought* (Wei et al., 2023) or *tree of thought* (Yao et al., 2023) into the generation process, typically producing a verbose stream of consciousness in order to elicit the necessary steps to solve complex tasks. Scaling test-time compute in this way has seen a significant improvement on many benchmarks (Geiping et al., 2025; Snell et al., 2024); however, the current status quo relies heavily on more technical and formalised reasoning tasks such as code generation and arithmetic. Whilst results in such domains can be deterministically validated (Wang et al., 2024a,b), and large quantities of samples can be generated (Xu et al., 2025), verbal and common-sense reasoning remains a fundamental task with wider applicability to the end-users of such technologies (Trichelair et al., 2019). We argue, therefore, that computational humour processing is an essential task for addressing the limitations in how models process natural language pragmatics and cultural knowledge.

To emphasise the existing focus on mathematical reasoning, of the benchmarks suites used in recent releases (i.e., Kimi K2[1] and Grok4[2]), nine focus on STEM problem solving: GPQA (Rein et al., 2024), USAMO (Petrov et al., 2025), HMMT,[3] AIME-25, [4] LiveCodeBench (Jain et al., 2024), SWE (Jimenez et al., 2024), OJBench (Wang et al., 2025b); or structured reasoning: ARC-AGI-2 (Chollet et al., 2025), and ACEBench (Chen et al., 2025). On the other hand, only Tau2 (Barres et al., 2025) touched upon communicative competence, albeit through customer service scenarios.

Computational humour presents a uniquely demanding arena for model evaluation that complements the existing array of benchmarks. Lacking from existing options are foundational elements of humour, such as phonetic understanding and pragmatic inference. Consider the joke in which a husband and wife are solving a crossword. The joke centres on the husband giving clues such as "*Emphatic no, five letters*", and "*Pistol, 3 letters*", resulting in guesses of "*never*" and "*gun*". This continues until the string of words guessed by the wife reads "never gun ugh give ewe Up". Whilst meaningless in isolation, the spoken realisation of the sequences equates to "*Never gonna give you up*", an instance of the popular *Rickrolling* internet phenomenon.[5] The creation and comprehension of this joke depends on the recognition of Rick Astley's song of the same title, the cultural phenomenon of *Rickrolling*, and an understanding of the phonetic similarity present in homophones.[6]

Moreover, humour frequently depends on understanding what was *not* said or inferring the opposite of explicit statements (Yus, 2013), which directly conflicts with the semantic-similarity-based retrieval systems that would typically be employed when addressing jokes that reference post-training events (Barnett et al., 2024; Akila and Jayakumar, 2014).

[1] https://moonshotai.github.io/Kimi-K2/
[2] https://x.ai/news/grok-4
[3] https://www.hmmt.org/
[4] https://artofproblemsolving.com

[5] https://en.wikipedia.org/wiki/Rickrolling
[6] This is made especially hard as there is not a 1-to-1 mapping, with "gun" and "ugh" taking the place of "gonna".

It is for these reasons that we assert the importance of humour understanding as a subdomain of NLP in the LLM age. The field of computational humour remains a fruitful area for continued work (Ignat et al., 2024; Lima Inácio and Gonçalo Oliveira, 2023); one that is critically overlooked and under-researched.

2.1 Theories of Humour

Humour, as a uniquely human experience, has been the focus of thinkers for centuries, becoming the subject of heavy debate. As such, several prominent theories have arisen that attempt to explain the perception of humour, a subset of which are presented below:

- **Relief Theory** suggests that humour, and particularly laughter, is the result of releasing psychological energy (Freud, 1963; Spencer, 1875; Kant, 1790).
- **Superiority Theory** posits that the experience of humour and comedy is born out of the perception that one individual is superior to another, thus making the inferior individual the subject of humour (Hobbes, 1660; Plato, 1892; Aristotle, 1902).
- **Incongruity Theory** states that humour is the perception of something that conflicts with established mental patterns and expectations, therefore being incongruous (Morreall, 2024; Tu et al., 2014).
- **Benign Violation Theory** postulates that humour arises from situations that are simultaneously harmless (i.e., benign) and are incongruous with expectations (i.e., violating a norm). (McGraw et al., 2012; McGraw and Warren, 2010)

Whilst the Superiority and Relief theories offer accounts of the experience of humour, they do not present simple interpretations that can be easily formulated linguistically and computationally. It is for this reason that, of the approaches to humour processing that are directly grounded in theory, incongruity and the general sense of norm violation remain essential elements (Tian et al., 2022; He et al., 2019; Valitutti et al., 2013). We present examples of textual linguistic humour in Figure 1 and an example of a multimodal humorous meme in Figure 2, to exemplify the broad range of possible humour forms.

3 Humour Generation

The ability to compose novel jokes requires an implicit understanding of the cognitive mechanisms that underlie humour. As outlined in the *Incongruity* and *Benign Violation* theories of humour, a necessary feature of humorous language is that it violates an expectation within the reader, either in the form of cultural or situational norms (e.g., topical and contextual humour), or linguistic norms (e.g., puns). This core property poses a challenge to computational approaches and to creative language generation more broadly: the language modelling objective aims to maximise the log likelihood of an output sequence, thereby working against the very incongruity that humour demands (Loakman et al., 2025a). The following subsections explore historical attempts to address or incorporate this contention for pun generation.

3.1 Pun Generation

Pun generation has dominated research, spanning from early rule-based approaches (Lessard, 1992) to contemporary neural methods. The historical prevalence of puns is largely a result of their computational tractability and theoretical grounding in Incongruity Theory, providing both an entry point and capacity to produce large quantities of training/evaluation data. Early works, such as JAPE (Binsted, 1996; Binsted and Ritchie, 1994) and STANDUP (Ritchie et al., 2007), leverage WordNet (Fellbaum, 1998) for semantics and Unisyn[7] for phonetic similarities. However, such classical approaches rely heavily on fixed schema, limiting overall creativity. For a deeper coverage of early literature in pun generation, we refer the reader to the prior surveys of Ritchie (2001) and Amin and Burghardt (2020).

The remainder of this section is split according to the distinction of *heterographic* and *homographic* puns (Redfern, 1987), including their combination. See Figure 1 for an example of each type.

3.1.1 Homographic Pun Generation

Homographic puns exploit polysemy, the phenomenon of multiple meanings existing for a single word, to create humour through semantic ambiguity. Early work in automatic pun generation focused on this category due to the accessibility of semantic resources like WordNet (Miller, 1994).

Yu et al. (2018) presents the first neural approach to homographic pun generation, training a conditional encoder-decoder LSTM on unlabelled Wikipedia text to create sentences that could support the semantics of two words simultaneously. This model then uses a constrained beam search algorithm to jointly decode

[7] https://www.cstr.ed.ac.uk/projects/unisyn/.

the two distinct senses of the same word, generating puns without requiring any pun-specific training data.

In a follow-up work, Yu et al. (2020) explore pun generation through a lexically constrained rewriting approach that first identifies constraint words supporting semantic incongruity for a sentence, then rewrites it with explicit positive and negative constraints. Their method achieved state-of-the-art results in both automatic and human evaluations.

Building on this, Luo et al. (2019) introduced PunGAN, a GAN-based model (Goodfellow et al., 2014) that employed a discriminator module with word-sense disambiguation capabilities to assess how well a generated sentence supported the polysemy of the target homographic pun word, aiming to maximise semantic ambiguity. The generator was trained via reinforcement learning using the discriminator's output as a reward signal to encourage the production of sentences that could support two word senses simultaneously. However, the key shortcoming of this approach was its tendency to produce generic outputs that prioritised semantic ambiguity over overall pun quality.

AMBIPUN (Mittal et al., 2022) takes as input a homograph with two distinct word senses and proposes that ambiguity comes from context rather than the pun word itself. The approach first produces a list of related concepts through a reverse dictionary, then utilises one-shot GPT-3 to generate context words from both concepts before generating puns that incorporate these contextual elements. They achieve a 52% success rate in human evaluation, significantly outperforming baselines but still remaining tied to a complex, multi-staged pipeline.

3.1.2 Heterographic Pun Generation

Heterographic puns present additional challenges due to the requirement of modelling phonetic similarity between different surface forms while maintaining semantic coherence. These puns exploit words that sound alike but are spelt differently, requiring systems to understand both phonetic relationships and semantic contexts (Kao et al., 2016).

He et al. (2019) generate heterographic puns through an unsupervised retrieve-and-edit framework based on the *local-global surprisal* principle. Given a pair of homophones (e.g., "died" and "dyed"), they first retrieve candidate sentences containing the alternative word from a corpus, replace the alternative word with the pun word, and insert a semantically related topic word to the start of the pun. The approach is a direct instantiation of Incongruity Theory: the foreshadowing topic word creates a strong association with the pun word in the global context while maintaining the local context of the alternative word in the sentence.

3.1.3 Combined Pun Generation

More recently, Tian et al. (2022) proposed a unified framework that addresses both homographic and heterographic pun generation by incorporating three key linguistic attributes: ambiguity, distinctiveness, and surprise. Their approach demonstrates that principled integration of humour theories can improve generation quality across different pun types.

Chen et al. (2024) propose a multi-stage curriculum learning approach using Direct Preference Optimisation (DPO) (Rafailov et al., 2024) to align models with the ability to create valid linguistic structures in support of both homographic and homophonic pun creation. The authors ultimately released the *ChinesePun* dataset of Chinese humour, presenting one of the few works that does not focus predominantly on English.

Recent systematic evaluation by Xu et al. (2024) provides evidence that large language models struggle considerably more with the generation of heterographic puns than homographic puns. This difficulty likely stems from the need to infer phonetic characteristics of words, which is challenging for models operating primarily on text-based representations (Baluja, 2025). The authors additionally identify what they term "lazy pun generation," whereby models incorporate both senses of the pun word in a single generation (e.g., "The sailor's pay was docked after he struggled to dock on time"). This phenomenon nullifies the intended effects of surprisal, eliminating any ambiguity that would require cognitive effort to resolve.

Sun et al. (2022) demonstrates a departure from standalone pun generation, releasing CUP: a novel task for *context situated* puns. CUP integrates puns more naturally into real-world conversations by demanding contextual awareness. The authors extract keywords from context sentences using RAKE (Rose et al., 2010) and use a T5 model to create sentences that support both meanings of the pun word. Logically, this approach can be viewed as a hybrid of Mittal et al. (2022), He et al. (2019), and Luo et al. (2019).

3.2 Non-Pun Humour Generation

Whilst puns have received the most attention in computational humour research, broader forms of humour generation present a greater challenge due to association with cultural knowledge, timeliness, and more complex incongruities.

Goel et al. (2024) proposed an approach combining template extraction and infilling using BERT (Devlin et al., 2019) with LLMs (specifically GPT-4, OpenAI, 2024 and Zephyr-7B, Tunstall et al., 2023) to generate set-up and punchline style jokes (e.g., "*Why did the chicken cross the road? To get to the other side*"). To achieve this, tokens from jokes are masked, and BERT attention weights are used to determine what elements of a given joke can be masked to maintain the essential structure, whilst removing overly topic-specific terms. As a result, the system learns to create joke structures that can then be filled to create novel joke instances.

On the other hand, Chung et al. (2024) present UNPIE, a benchmark to assess the understanding of Vision Language Models (VLMs) when reasoning about lexical incongruities. To achieve this, they take as input written puns and generate an image to serve as a visual representation of the pun, demonstrating that such visual clues may be beneficial in pun understanding tasks, helping to identify the location of a pun in a given text.

Horvitz et al. (2020) focus on the generation of satirical headlines, a type of language used that is both humorous and dry, being used to criticise individuals and entities on complex topics. In doing so, they tackle the linking of relevant real-world knowledge from sources such as Wikipedia and CNN, to relevant satirical headlines from TheOnion,[8] which are then used to finetune GPT-2 (Radford et al., 2019).

Tian et al. (2021) explore humour through the lens of hyperbole with HYPOGEN. To do so, they curate a dataset of hyperbolic phrases following the "*so [X] that [Y]*" pattern (e.g., "*My personality is so dry that a cactus flourishes inside*"), using variants of CoMET models (Bosselut et al., 2019) to learn the commonsense and counterfactual relationships present in hyperbolic language to assist generation.

Finally, in recent years, there has been a growing body of work in the humour-adjacent domain of tongue twister generation (texts where entertainment and humour arise from mispronunciations stemming from complex phonetics). Such approaches involve training keyword-to-twister and style-transfer models for tongue twister generation, either via training on graphemes (Loakman et al., 2025a, 2023; Keh et al., 2023) or phonemes (Keh et al., 2023). Research in this domain has also highlighted the benefit of incorporating explicit phonetic and phonemic information into the generation of language formats that rely on such characteristics.

[8] https://theonion.com/

4 Humour Explanation

> "*...if it can say why a joke's funny, it really does understand*" - Geoffrey Hinton.[9]

The above quote from Hinton ("the Godfather of AI") refers directly to the use of joke explanation as a milestone achievement in the marketing of Google's PaLM LLM (Chowdhery et al., 2023). Detection approaches, while offering objective and easily automatable evaluation, remain susceptible to statistical flukes. Explanation generation eliminates this vulnerability through a vanishingly low probability of accidentally producing coherent comedic analysis. Moreover, the act of explanation provides valuable insight into the inner workings of what still exists as nearly opaque black boxes. This value comes at the expense of significantly more challenging, expensive and labour-intensive evaluation methodologies.

In its present state, the field of humour explanation remains critically understudied. To provide necessary context, we draw upon select literature covering humour classification and detection.

4.1 Explanation through Classification

Humour explanation, although understudied, is not a completely isolated evaluation of comprehension. Whilst not encompassing the full scope of a natural language explanation, humour *detection* tasks models with identifying the linguistic traits common to humour. As such, we denote humour *detection* a precursor to explanation.

The majority of existing works on humour explanation do not focus on providing textual natural language explanations, instead focusing on other indications of a joke's source, such as word senses or identifying the type of humour being expressed.

Miller et al. (2017) presents an overview of Task 7 from SemEval 2017, which concerned detection and "interpretation" (i.e., classification) of puns. Specifically, the pun interpretation task consisted of assigning word sense keys from WordNet (Fellbaum, 1998) to the punning word contained in a text. One approach to the task, taken by Oele and Evang (2017), splits a given text into 2 parts at all possible locations, with the split where both parts have low semantic similarity being where the two meanings of the pun word are best separated. A Word Sense Disambiguation (WSD) model is then used to retrieve the relevant senses. Similar techniques based on mapping and comparing the

[9] From the 16th June 2025 episode of The Diary of a CEO podcast. See https://youtu.be/giT0ytynSqg?si=O0iN3DM2Fv58fp8o&t=4460.

semantics of different partitions of the text with the pun word are shown in multiple submissions (e.g., Hurtado et al., 2017; Indurthi and Oota, 2017). Whilst assigning sense keys to a pun word offers a minor explanation to a user, the requirement for the pun word to be pre-identified nullifies the appropriateness of such systems in practice.

Palma Preciado et al. (2024) present an overview of the JOKER shared task from CLEF 2024, where Task 2 aimed to classify humorous texts based on their genre and the linguistic techniques used, including irony, sarcasm, exaggeration, incongruity/absurdity, self-deprecation and wit/surprise. In total, they received 54 submissions to the shared task, with techniques ranging from training BERT classifiers (e.g., Narayanan et al., 2024) and ensembles of classic machine learning classifiers (e.g., Bartulović and Váradi, 2024), to zero-shot prompting of LLMs like GPT-4 (e.g., Wu et al., 2024a).[10]

However, explanation through classification presents a series of issues. Firstly, and most obviously, whilst such approaches can demonstrate a model's understanding of humour (to an extent), assigning the correct label to a given joke is unlikely to present a human user with valuable information that would aid their interpretation. Secondly, classification requires the creation of a taxonomy under which to categorise different jokes. Owing to the complexity of humour, high-quality, robust taxonomies are challenging to create. For instance, Task 2 from JOKER (Palma Preciado et al., 2024) has considerable overlap between joke categories. A prerequisite for both irony and sarcasm is the violation of a norm, whilst *incongruity* is presented as a separate category. This is additionally true of "wit" (a subjective judgement of cleverness and novelty) and *surprise*.

4.2 Natural Language Explanation

Whilst explanation through the lens of classification is a valid approach for simple joke formats such as puns, more complex humour formats, such as more esoteric, context-dependent jokes, benefit from natural language explanations (in addition to being more friendly to the proposed end-user). Natural language explanations rectify the coarseness of explanation through classification. Such approaches do not require a predefined taxonomy of humour types to identify, but instead test the overall abilities of models to provide tailored, specific explanations for every humorous text.

[10]See Palma Preciado et al. (2024) for a full overview of submissions.

Xu et al. (2024) generates natural language explanations for puns using chain-of-thought prompting with a range of LLMs. They find that most LLMs are able to identify the punning word in both heterographic and homographic examples, but most models struggle with correctly identifying the alternative intended meaning of heterographic puns, which rely on phonetic similarity. The authors identify a series of mistakes commonly made by models, including failure to recognise the joke as a pun, incorrectly identifying the alternative senses of the pun word, and failing to provide the meanings of the pun word, but without contextualisation into a natural language explanation.

Loakman et al. (2025b) extend these findings and use a range of LLMs on a range of joke formats, including homographic puns, heterographic puns, long-form humour, and topical jokes from Reddit. Their results further confirm that LLMs struggle with heterographic puns (owing to their lack of phonetic knowledge), but additionally demonstrate that longer incongruity-based jokes and jokes that rely on esoteric topical knowledge present even more difficulty to LLMs. Whilst a zero-shot non-chain-of-thought approach is used, the authors find that the Llama models distilled from Deepseek R1, are among the worst-performing. They hypothesise that this is a result of the ambiguity in real-world references within topical humour, leading to early misunderstandings being propagated through the reasoning process. A further, smaller-scale evaluation of ChatGPT's humour explanation ability is presented by Jentzsch and Kersting (2023). Similarly, Wang et al. (2025a) investigate the ability of LLMs to explain "drivelology", a linguistic form characterised as "nonsense with depth", incorporating humour alongside other elements such as sarcasm, irony, and tautologies. Whilst they investigate the ability for LLMs to correctly categorise the type of "drivel" being used, they additionally perform zero-shot explanation generation and likewise find that models struggle significantly with creating high-quality explanations.

Identifying the common failure of models to understand jokes where phonetic characteristics play a vital role (e.g., heterographic puns), Baluja (2025) investigate whether access to speech audio of the joke being read aloud leads to improved performance. They assessed the ability of Gemini 1.5 (Team, 2024) to explain jokes with and without access to text-to-speech readings, demonstrating approximately a 2.5% to 4% improvement in explanation performance. Whilst moderate, such findings indicate the affordances provided by multimodal prompting in

language forms that rely heavily on modalities that are absent from text (i.e., pronunciations), suggesting room for further performance gains alongside improvements in the fusion of modalities.

Multimodal Humour and Meme Explanation In the realm of multimodal humour, Hessel et al. (2023) present work on the explanation of visual jokes in the form of humorous captions from the New Yorker Cartoon Caption Contest. In such a task, models must understand the visual cartoon in order to disambiguate pun words or correctly establish the key incongruity, owing to the text alone being ambiguous. For instance, one cartoon presents a barbershop with a hole in the roof and a spring coming out of a barbershop chair, with the caption reading "He'll be back". This phrase typically means that someone will return to an establishment even if they were dissatisfied (e.g., they provide an essential service or the individual's criticism was unreasonable). However, in this instance, the visual cues reveal that "He'll be back" is a literal statement, referring to the effects of gravity on the man who was ejected from his chair by a spring. To evaluate this, Hessel et al. (2023) finetune GPT-3 on human explanations, as well as perform 5-shot prompting with GPT-4. The results showed that access to visual information from the cartoons resulted in a better explanation in 84.7% of cases (via human preference judgements). However, whilst in-context learning with 5-shot GPT-4 outperforms finetuned GPT-3, it was shown that human-authored explanations are still preferred in 68% of instances.

On the other hand, the lion's share of work in multimodal humour explanation specifically investigates memes, a type of media (typically an image) that is copied and propagated rapidly across the internet, often comprising text captions and related imagery (see Figure 2 for an example). Hwang and Shwartz (2023) present MEMECAP, a dataset tailored to meme understanding, consisting of 6.3K memes from the *r/Memes* subreddit alongside crowdsourced captions that explain the semantics that the meme is trying to convey. Additionally, Khan et al. (2024) present a dataset of 13K+ memes, including those with audio. In this instance, an LLM pipeline was utilised to generate explanations of the memes, which were then refined by human annotators. Interestingly, the authors aim to generate explanations that achieve the aim of entertaining and amusing readers (therefore being somewhat humorous themselves), rather than being strictly objective accounts of the humour.

Figure 2: An example of a meme, a form of multimodal humour that incorporates visual elements, often alongside text. In this instance, the humour arises from the absurdity of the belief that asking about cheese is political, whilst being enhanced by the confused T-pose Kermit the Frog render. The original meme has been edited to remove expletives.

Park et al. (2024) acknowledge that author's intent contributes significantly to meaning, ergo perception. They define the task of *intent description generation*, accompanied by a dataset of 950 samples that are annotated with perceived intention and the necessary context. We believe that intent-aware systems are a compelling direction for future study, which we explore further in §5.4.

Furthermore, Agarwal et al. (2024) present MEMEMQA, a multimodal Q&A dataset for asking questions regarding the content of memes in order to better understand them. From this, they develop ARSENAL, a multimodal meme understanding pipeline that uses LLMs to reason about a given meme in relation to a specific question.

The site *Know Your Meme*[11] tracks the provenance of memes as they develop, making it a wealth of knowledge regarding the origin and growth of memes over time. From this, Tommasini et al. (2023) built the Internet Meme Knowledge Graph, comprising 2 million edges to represent the semantics of multimodal memes.

In a more directly application-based setting, Jha et al. (2024) explore the explanation of memes being used explicitly for purposes of cyberbullying. Using the MultiBully dataset (Maity et al., 2022), they add annotations to highlight pertinent visual and linguistic aspects of a meme that highlight its intent as cyberbullying, allowing the training of enhanced bullying detection models.

[11] https://knowyourmeme.com/

5 Discussion & Future Directions

Generative tasks in the area of computational humour present a range of challenges owing to aspects such as the subjective and potentially offensive nature of humour and the ethics of generating any creative language form. In this section, we present a range of promising research directions for computational humour generation and explanation, relating each proposed direction to the practical and ethical challenges that they help ameliorate.

5.1 Demographic Aware Humour Generation

In following the recent trends established by works such as Sun et al. (2022) and Garimella et al. (2020), an essential primary focus of future research should be audience-tailored, demographic-aware, and contextually nuanced humour generation that is able to better address the preferences of particular end users. Owing to the wide gamut of topics that humour may be found in, such approaches would aid in decreasing the risk of generating material that is considered offensive to a given end-user. Consequently, a promising direction is to explicitly incorporate human preference alignment techniques such as RLHF (Ouyang et al., 2022; Stiennon et al., 2020; Christiano et al., 2017), DPO (Rafailov et al., 2023), and PPO (Schulman et al., 2017) to the task of humour generation.[12]

Additionally, the developing area of perspectivist approaches in NLP (Fleisig et al., 2024; Valette, 2024; Abercrombie et al., 2022) provide a route through which to consider individual preferences for subjective domains such as humour. Progress to this end is demonstrated by Casola et al. (2024) and Frenda et al. (2023), who present perspective-aware datasets for irony processing (a phenomenon highly related to humour). The types of humour and jokes that are most often required to be explained to someone are those that are ambiguous, nuanced, and frequently focused on sensitive topics. Whilst the potential sensitivity of a joke topic can have an intensifying impact on the level of humour perceived by the *intended* audience, it acts as part of a risk-to-reward tradeoff, likewise increasing the risk of offence if presented to the wrong audience (McGraw and Warren, 2010).

5.2 Human-in-the-Loop Humour Generation

Humour is a quintessential demonstration of human intelligence and creativity. As such, the development of computational models for the generation of such content poses the risk of increasing the anthropomorphism of machine learning models - something that is at times desirable, but also disproportionately impacts vulnerable users of these technologies (Shevlin, 2024). In addition to this, the creation of (intentional) humour is a creative endeavour. The generation of "*creative*" language forms with computational models has the potential to have severe negative impacts on the arts and creative industries as a whole, and reduce the outlets available to people to realise financial gains from their own creativity. Whilst systems such as Witscript (Toplyn, 2023) were developed by real-world comedy writers, such examples are the exception to the rule. As a result, future work should focus increasingly on human-in-the-loop approaches (Wu et al., 2022) to humour generation, requiring meaningful contributions from the end-user, or combining elements of humour generation and explanation to develop systems for joke workshopping, offering valuable feedback and direction on human-authored humour, rather than generating jokes wholesale.

5.3 Complex Humour Explanation

As highlighted in this survey paper, humour explanation is an interesting, challenging, and important task for assessing the verbal and commonsense reasoning abilities of models (particularly LLMs), yet is a scarcely researched domain. If the end goal is to imbue systems with human-level reasoning abilities, being able to explain humour is fundamental. Particular focus should be given to context-sensitive, esoteric humour such as topical jokes. Such jokes present a rich area for aiding in the development of advanced information retrieval systems that are able to work with ambiguous topics such as the references to world events and pop culture phenomena found in complex jokes.

5.4 Intent-Aware Humour Explanation

Whilst humour explanation is a challenging task with a real-world benefit of reducing communication barriers between people from different backgrounds, it is also an area with considerable ethical considerations. Another potential future direction for research in the area of humour explanation is the development of intent-aware models (Ma et al., 2025; Park et al., 2024). Such approaches would attempt to model the characteristics and intent of the author based on available information (such as prior language use and known/inferred demographic variables) prior to attempting to explain potentially humorous language use. Not only could this information aid in inferring references to ambiguous content (aiding reasoning

[12] See Jiang et al. (2024) for an overview of approaches to learning from human preference data.

and relevant document retrieval), it would also aid in overcoming the undesirable effect of legitimising the instances where "*it's just a joke!*" is used as a guise for spreading hatred through offensive material (Brommage, 2015; Hodson et al., 2010; Ford et al., 2015). Such approaches have extensive applications in online communication venues such as social media and in handling interpersonal conflict. Whilst individual preferences and beliefs as to what topics are possible to derive humour from remain highly subjective (Smuts, 2010; Gaut, 1998), improved modelling of the underlying intent would help identify instances where offence was likely caused accidentally.[13]

6 Conclusion

In this paper, we have explored the current landscape in the domain of computational humour generation and explanation, outlining the approaches taken in existing work and outlining promising future directions. We propose that computational humour processing remains an unsolved task with a wide range of real-world applications, as well as being one of the most promising yet overlooked domains in which to evaluate the verbal reasoning abilities of modern LLMs. Furthermore, we identified and outlined a range of future approaches to be taken in generative computational humour research and made the case that future work should focus both on an increased breadth of humour formats, as well as giving explicit consideration to the ethics of computational humour.

Limitations

We position this work as an overview of computational humour generation and explanation. As a result, we focus on the breadth of research available, rather than presenting in-depth accounts of the exact technical novelty of existing works. Additionally, whilst we aim to be comprehensive with our overview, some instances of relevant research may have been missed. Furthermore, we have not extensively explored other areas of computational humour processing, such as automatic metrics and human evaluation paradigms, available datasets, or approaches to humour detection.

Ethics Statement

Relating to the ethical considerations explored in this paper, whilst we believe that humour generation and explanation are worthwhile pursuits, we acknowledge and appreciate the stances taken by other individuals. This relates specifically to ethical considerations surrounding the generation of creative language in any form, the generation of humour that is potentially offensive, and whether or not providing an explanation for potentially offensive humour equates to an endorsement of the content of such humour. Such decisions should continue to be a source of discussion in the NLP community as a whole, and we encourage individual researchers working in these domains to explicitly state their stance in published works.

[13]We of course do not advocate for the extreme version of such systems in practice, whereby individuals are effectively being accused of thought-crime due to an intent assigned to an author via a computational model.

Acknowledgments

Tyler Loakman is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

References

Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma, editors. 2022. *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*. European Language Resources Association, Marseille, France.

Siddhant Agarwal, Shivam Sharma, Preslav Nakov, and Tanmoy Chakraborty. 2024. MemeMQA: Multimodal question answering for memes via rationale-based inferencing. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5042–5078, Bangkok, Thailand. Association for Computational Linguistics.

D Akila and C Jayakumar. 2014. Semantic similarity- a review of approaches and metrics. *International Journal of Applied Engineering Research*, 9:27581–27600.

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

Aristotle. 1902. *Poetics*. Macmillan and Co., London. First composed c. 335 BCE.

Ashwin Baluja. 2025. Text is not all you need: Multimodal prompting helps LLMs understand humor. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 9–17, Online. Association for Computational Linguistics.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. 2025. τ^2-bench: Evaluating conversational agents in a dual-control environment.

Antonia Bartulović and Dóra Paula Váradi. 2024. University of split and university of malta (team ab&dpv) at the clef 2024 joker track: From 'lol' to 'mdr' using artificial intelligence models to retrieve and translate puns. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, volume 3740 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Nancy Bell and Salvatore Attardo. 2010. Failed humor: Issues in non-native speakers' appreciation and understanding of humor. *Intercultural Pragmatics*, 7(3):423–447.

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

Kim Binsted and Graeme Ritchie. 1994. An implemented model of punning riddles. In *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*, AAAI'94, page 633–638. AAAI Press.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Thomas Brommage. 2015. Just kidding, folks!: An expressivist analysis of offensive humor. *Florida Philosophical Review*, 15(1).

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

Chen Chen, Xinlong Hao, Weiwen Liu, Xu Huang, Xingshan Zeng, Shuai Yu, Dexun Li, Shuai Wang, Weinan Gan, Yuefeng Huang, Wulong Liu, Xinzhi Wang, Defu Lian, Baoqun Yin, Yasheng Wang, and Wu Liu. 2025. Acebench: Who wins the match point in tool usage?

Yang Chen, Chong Yang, Tu Hu, Xinhao Chen, Man Lan, Li Cai, Xinlin Zhuang, Xuan Lin, Xin Lu, and Aimin Zhou. 2024. Are U a joke master? pun generation via multi-stage curriculum learning towards a humor LLM. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 878–890, Bangkok, Thailand. Association for Computational Linguistics.

Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. 2025. Arc-agi-2: A new challenge for frontier ai reasoning systems.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4302–4310, Red Hook, NY, USA. Curran Associates Inc.

Jiwan Chung, Seungwon Lim, Jaehyun Jeon, Seungbeen Lee, and Youngjae Yu. 2024. Can visual language models resolve textual ambiguity with visual cues? let visual puns tell you! In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2452–2469, Miami, Florida, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Marta Dynel. 2009. Beyond a joke: Types of conversational humour. *Language and Linguistics Compass*, 3(5):1284–1299.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.

Thomas E. Ford, Kyle Richardson, and Whitney E. Petit. 2015. Disparagement humor and prejudice: Contemporary theory and research. *HUMOR*, 28(2):171–186.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella,

Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.

Sigmund Freud. 1963. *Jokes and their relation to the unconscious.* Jokes and their relation to the unconscious. W. W. Norton, Oxford, England.

Vaishali Ganganwar, Manvainder, Mohit Singh, Priyank Patil, and Saurabh Joshi. 2024. Sarcasm and humor detection in code-mixed hindi data: A survey. In *Computing and Machine Learning*, pages 453–469, Singapore. Springer Nature Singapore.

Aparna Garimella, Carmen Banea, Nabil Hossain, and Rada Mihalcea. 2020. "judge me by my size (noun), do you?" YodaLib: A demographic-aware humor generation framework. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2814–2825, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Berys Nigel Gaut. 1998. Just joking: The ethics and aesthetics of humor. *Philosophy and Literature*, 22(1):51–68.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. 2025. Scaling up Test-Time Compute with Latent Reasoning: A Recurrent Depth Approach.

Mayank Goel, Parameswari Krishnamurthy, and Radhika Mamidi. 2024. Automating humor: A novel approach to joke generation using template extraction and infilling. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 442–448, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks.

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

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Tim Highfield. 2015. Tweeted joke lifespans and appropriated punch lines: Practices around topical humor on social media. *International Journal of Communication*, 9(0).

Thomas Hobbes. 1660. *The Leviathan*.

Gordon Hodson, Jonathan Rush, and Cara C. Macinnis. 2010. A joke is just a joke (except when it isn't): cavalier humor beliefs facilitate the expression of group dominance motives. *Journal of Personality and Social Psychology*, 99(4):660–682.

Zachary Horvitz, Nam Do, and Michael L. Littman. 2020. Context-driven satirical news generation. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 40–50, Online. Association for Computational Linguistics.

Lluís-F. Hurtado, Encarna Segarra, Ferran Pla, Pascual Carrasco, and José-Ángel González. 2017. ELiRF-UPV at SemEval-2017 task 7: Pun detection and interpretation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 440–443, Vancouver, Canada. Association for Computational Linguistics.

EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A dataset for captioning and interpreting memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445, Singapore. Association for Computational Linguistics.

Oana Ignat, Zhijing Jin, Artem Abzaliev, Laura Biester, Santiago Castro, Naihao Deng, Xinyi Gao, Aylin Ece Gunal, Jacky He, Ashkan Kazemi, Muhammad Khalifa, Namho Koh, Andrew Lee, Siyang Liu, Do June Min, Shinka Mori, Joan C. Nwatu, Veronica Perez-Rosas, Siqi Shen, Zekun Wang, Winston Wu, and Rada Mihalcea. 2024. Has it all been solved? open NLP research questions not solved by large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8050–8094, Torino, Italia. ELRA and ICCL.

Vijayasaradhi Indurthi and Subba Reddy Oota. 2017. Fermi at SemEval-2017 task 7: Detection and interpretation of homographic puns in English language. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 457–460, Vancouver, Canada. Association for Computational Linguistics.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code.

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

Prince Jha, Krishanu Maity, Raghav Jain, Apoorv Verma, Sriparna Saha, and Pushpak Bhattacharyya. 2024. Meme-ingful analysis: Enhanced understanding of cyberbullying in memes through multimodal explanations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 930–943, St. Julian's, Malta. Association for Computational Linguistics.

Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. 2024. A survey on human preference learning for large language models.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. Swe-bench: Can language models resolve real-world github issues?

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

Immanuel Kant. 1790. *Critique of Judgment*. Barnes & Noble.

Justine T. Kao, Roger Levy, and Noah D. Goodman. 2016. A computational model of linguistic humor in puns. *Cognitive Science*, 40(5):1270–1285.

Sedrick Scott Keh, Steven Y. Feng, Varun Gangal, Malihe Alikhani, and Eduard Hovy. 2023. PANCETTA: Phoneme aware neural completion to elicit tongue twisters automatically. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 491–504, Dubrovnik, Croatia. Association for Computational Linguistics.

Mary Ogbuka Kenneth, Foaad Khosmood, and Abbas Edalat. 2024. Systematic literature review: Computational approaches for humour style classification.

Anas Anwarul Haq Khan, Tanik Saikh, Arpan Phukan, and Asif Ekbal. 2024. Hope 'the paragraph guy' explains the rest : Introducing MeSum, the meme summarizer. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6654–6668, Miami, Florida, USA. Association for Computational Linguistics.

Liisi Laineste. 2002. Take it with a grain of salt: The kernel of truth in topical jokes. *Folklore: Electronic Journal of Folklore*, 21:7–25.

G. Lessard. 1992. Computational modelling of linguistic humour : Tom swifties. *Selected Papers from the 1992 Association for Literary and Linguistic Computing (ALLC) and the Association for Computers and the Humanities (ACH) Joint Annual Conference*, pages 175–178.

Marcio Lima Inácio and Hugo Gonçalo Oliveira. 2023. Towards generation and recognition of humorous texts in Portuguese. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 26–36, Dubrovnik, Croatia. Association for Computational Linguistics.

Tyler Loakman, Chen Tang, and Chenghua Lin. 2023. TwistList: Resources and baselines for tongue twister generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–589, Toronto, Canada. Association for Computational Linguistics.

Tyler Loakman, Chen Tang, and Chenghua Lin. 2025a. Train and constrain: Phonologically informed tongue twister generation from topics and paraphrases. *Computational Linguistics*, 51(2):415–466.

Tyler Loakman, William Thorne, and Chenghua Lin. 2025b. Comparing apples to oranges: A dataset analysis of llm humour understanding from traditional puns to topical jokes.

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

Jiayuan Ma, Hongbin Na, Zimu Wang, Yining Hua, Yue Liu, Wei Wang, and Ling Chen. 2025. Detecting conversational mental manipulation with intent-aware prompting. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9176–9183, Abu Dhabi, UAE. Association for Computational Linguistics.

Krishanu Maity, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1739–1749, New York, NY, USA. Association for Computing Machinery.

Wingyun Mak and Brian D. Carpenter. 2007. Humor comprehension in older adults. *Journal of the International Neuropsychological Society*, 13(4):606–614.

A. Peter McGraw and Caleb Warren. 2010. Benign violations: Making immoral behavior funny. *Psychological Science*, 21(8):1141–1149.

A. Peter McGraw, Caleb Warren, Lawrence E. Williams, and Bridget Leonard. 2012. Too close for comfort, or too far to care? finding humor in distant tragedies and close mishaps. *Psychological Science*, 23(10):1215–1223. PMID: 22941877.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Tristan Miller, Christian Hempelmann, and Iryna Gurevych. 2017. SemEval-2017 task 7: Detection and interpretation of English puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 58–68, Vancouver, Canada. Association for Computational Linguistics.

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

John Morreall. 2024. Philosophy of Humor. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2024 edition. Metaphysics Research Lab, Stanford University.

Sarvesh Narayanan, Jayasimman J, and Shiva Ganesh V. 2024. Clef 2024 joker task 2: Using roberta and bert-uncased for humour classification according to genre and technique. In *CEUR Workshop Proceedings*, volume 3740 of *GEUR Workshop Proceedings*. Notebook for the Joker Lab at CLEF 2024.

Khoi P. N. Nguyen and Vincent Ng. 2024. Computational meme understanding: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21251–21267, Miami, Florida, USA. Association for Computational Linguistics.

Anton Nijholt, Andreea Niculescu, Alessandro Valitutti, and Rafael Enrique Banchs. 2017. Humor in human-computer interaction: A short survey. In *IFIP TC13 International Conference on Human-Computer Interaction*.

Dieke Oele and Kilian Evang. 2017. BuzzSaw at SemEval-2017 task 7: Global vs. local context for interpreting and locating homographic English puns with sense embeddings. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 444–448, Vancouver, Canada. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Victor Manuel Palma Preciado, Grigori Sidorov, Liana Ermakova, Anne-Gwenn Bosser, Tristan Miller, and Adam Jatowt. 2024. Overview of the clef 2024 joker task 2: Humour classification according to genre and technique. In *CEUR Workshop Proceedings*, volume 3740 of *GEUR Workshop Proceedings*.

Jeongsik Park, Khoi P. N. Nguyen, Terrence Li, Suyesh Shrestha, Megan Kim Vu, Jerry Yining Wang, and Vincent Ng. 2024. MemeIntent: Benchmarking intent description generation for memes. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 631–643, Kyoto, Japan. Association for Computational Linguistics.

Ivo Petrov, Jasper Dekoninck, Lyuben Baltadzhiev, Maria Drencheva, Kristian Minchev, Mislav Balunović, Nikola Jovanović, and Martin Vechev. 2025. Proof or bluff? evaluating llms on 2025 usa math olympiad.

Plato. 1892. *Philebus*. Clarendon Press, Oxford. First composed c. 360-347 BCE.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

C. Ramakristanaiah, P. Namratha, Rajendra Kumar Ganiya, and Midde Ranjit Reddy. 2021. A survey on humor detection methods in communications. In *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pages 668–674.

Walter Redfern. 1987. Puns. *Journal of English Linguistics*, 20(1):158–158.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Graeme Ritchie. 2001. Current Directions in Computational Humour. *Artificial Intelligence Review*, 16(2):119–135.

Graeme D Ritchie, Ruli Manurung, Helen Pain, Annalu Waller, Rolf Black, and Dave O'Mara. 2007. A practical application of computational humour. In *Proceedings of the Fourth International Joint Conference on Computational Creativity (Goldsmith's, London)*, pages 91–98.

Hannes Ritschel and Elisabeth André. 2018. Shaping a social robot's humor with natural language generation and socially-aware reinforcement learning. In *Proceedings of the Workshop on NLG for Human–Robot Interaction*, pages 12–16, Tilburg, The Netherlands. Association for Computational Linguistics.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. *Automatic Keyword Extraction from Individual Documents*, chapter 1. John Wiley Sons, Ltd.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Sergio Servantez, Joe Barrow, Kristian Hammond, and Rajiv Jain. 2024. Chain of logic: Rule-based reasoning with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2721–2733, Bangkok, Thailand. Association for Computational Linguistics.

Henry Shevlin. 2024. All too human? identifying and mitigating ethical risks of social ai. *Law, Ethics Technology*, 1(2).

Aaron Smuts. 2010. The ethics of humor: Can your sense of humor be wrong? *Ethical theory and moral practice*, 13(3):333–347.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters.

Herbert Spencer. 1875. The physiology of laughter. In *Illustrations of universal progress: A series of discussions.*, pages 194–209. D Appleton & Company.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jiao Sun, Anjali Narayan-Chen, Shereen Oraby, Shuyang Gao, Tagyoung Chung, Jing Huang, Yang Liu, and Nanyun Peng. 2022. Context-situated pun generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4635–4648, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Yufei Tian, Divyanshu Sheth, and Nanyun Peng. 2022. A unified framework for pun generation with humor principles. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3253–3261, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yufei Tian, Arvind krishna Sridhar, and Nanyun Peng. 2021. HypoGen: Hyperbole generation with commonsense and counterfactual knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1583–1593, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Riccardo Tommasini, Filip Ilievski, and Thilini Wijesiriwardene. 2023. IMKG: The Internet Meme Knowledge Graph. In Catia Pesquita, Ernesto Jimenez-Ruiz, Jamie McCusker, Daniel Faria, Mauro Dragoni, Anastasia Dimou, Raphael Troncy, and Sven Hertling, editors, *The Semantic Web*, volume 13870, pages 354–371. Springer Nature Switzerland, Cham.

Joe Toplyn. 2023. Witscript 3: A hybrid ai system for improvising jokes in a conversation.

Paul Trichelair, Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2019. How reasonable are common-sense reasoning tasks: A case-study on the Winograd schema challenge and SWAG. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3382–3387, Hong Kong, China. Association for Computational Linguistics.

S. Tu, X. Cao, X. Yun, K. Wang, G. Zhao, and J. Qiu. 2014. A new association evaluation stage in cartoon apprehension: Evidence from an erp study. *Journal of Behavioral and Brain Science*, 4:75–83.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Mathieu Valette. 2024. What does perspectivism mean? an ethical and methodological countercriticism. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 111–115, Torino, Italia. ELRA and ICCL.

Alessandro Valitutti, Hannu Toivonen, Antoine Doucet, and Jukka M. Toivanen. 2013. "let everything turn well in your wife": Generation of adult humor using lexical constraints. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 243–248, Sofia, Bulgaria. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024a. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand. Association for Computational Linguistics.

Yang Wang, Chenghao Xiao, Chia-Yi Hsiao, Zi Yan Chang, Chi-Li Chen, Tyler Loakman, and Chenghua Lin. 2025a. Drivel-ology: Challenging llms with interpreting nonsense with depth.

Zhexu Wang, Yiping Liu, Yejie Wang, Wenyang He, Bofei Gao, Muxi Diao, Yanxu Chen, Kelin Fu, Flood Sung, Zhilin Yang, Tianyu Liu, and Weiran Xu. 2025b. OJBench: A Competition Level Code Benchmark For Large Language Models.

Zihan Wang, Yunxuan Li, Yuexin Wu, Liangchen Luo, Le Hou, Hongkun Yu, and Jingbo Shang. 2024b. Multi-step problem solving through a verifier: An empirical analysis on model-induced process supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7309–7319, Miami, Florida, USA. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

Michael Wierzbicki and Richard David Young. 1978. The relation of intelligence and task difficulty to appreciation of humor. *The Journal of General Psychology*, 99(1):25–32.

Shih-Hung Wu, Yu-Feng Huang, and Tsz-Yeung Lau. 2024a. Humour classification by fine-tuning llms: Cyut at clef 2024 joker lab subtask humour classification according to genre and technique. In *CEUR Workshop Proceedings*, volume 3740 of *GEUR Workshop Proceedings*. Notebook for the CYUT Lab at CLEF 2024.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024b. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862, Mexico City, Mexico. Association for Computational Linguistics.

Zhangchen Xu, Yang Liu, Yueqin Yin, Mingyuan Zhou, and Radha Poovendran. 2025. Kodcode: A diverse, challenging, and verifiable synthetic dataset for coding.

Zhijun Xu, Siyu Yuan, Lingjie Chen, and Deqing Yang. 2024. "a good pun is its own reword": Can large language models understand puns? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11766–11782, Miami, Florida, USA. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Zhiyuan Zeng, Xiaonan Li, Junqi Dai, Qinyuan Cheng, Xuanjing Huang, and Xipeng Qiu. 2024. Reasoning in flux: Enhancing large language models reasoning through uncertainty-aware adaptive guidance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2401–2416, Bangkok, Thailand. Association for Computational Linguistics.

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

Zhiwei Yu, Hongyu Zang, and Xiaojun Wan. 2020. Homophonic pun generation with lexically constrained rewriting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2870–2876, Online. Association for Computational Linguistics.

Francisco Yus. 2013. *An inference-centered analysis of jokes: The Intersecting Circles Model of humorous communication*, pages 59–82.