Comparing Apples to Oranges: A Dataset & Analysis of LLM Humour Understanding from Traditional Puns to Topical Jokes

Anonymous ARR submission

Abstract

Humour, as complex language form, is derived from myriad aspects of life, whilst existing work on computational humour has focussed almost exclusively on short pun-based jokes. In this work, we investigate whether the ability of Large Language Models (LLMs) to explain humour depends on the particular humour form. We compare models on simple puns and more complex topical humour that requires knowledge of real-world entities and events. In doing so, we curate a dataset of 600 jokes split across 4 joke types and manually write high-quality explanations. These jokes include heterographic and homographic puns, contemporary internet humour, and topical jokes, where understanding relies on reasoning beyond "common sense", rooted instead in world knowledge regarding news events and pop culture. Using this dataset, we compare the zero-shot abilities of a range of LLMs to accurately and comprehensively explain jokes of different types, identifying key research gaps in the task of humour explanation. We find that none of the tested models (inc. reasoning models) are capable of reliably generating adequate explanations of all joke types, further highlighting the narrow focus of most works in computational humour on overly simple joke forms.[1]

1 Introduction

The perception and understanding of humour is essential to human interaction and entertainment. Consequently, the ability to correctly process jokes is paramount to achieving a human-like understanding of ambiguity and social context in language. However, most existing work on computational humour have narrowly focussed on the simpler task of humour detection (Loakman et al., 2023a; Meaney et al., 2021; Miller et al., 2017), which allows models to learn from confounding signals to detect certain joke types (Lima Inácio et al., 2023). On the other hand, humour *explanation* tasks a computational model with the more challenging objective of explicitly outlining how and why a presented text is humorous (Hessel et al., 2023).

In addition to the general focus on joke detection, existing works on all humour processing tasks (i.e., detection, explanation, and generation) have mostly focussed on short-form puns such as "*Yesterday I accidentally swallowed some food colouring. The doctor says I'm OK, but I feel like I've dyed a little inside*" (He et al., 2019), where humour arises from the interpretation of the punning word (i.e., "dyed") with a phonetically similar word with different meaning (i.e., "died") in the case of heterographic puns, or the polysemy of a word for homographic puns (Attardo, 2008). Whilst puns are commonplace, the humour is primarily self-contained, requiring only a common-sense understanding of the world (e.g., that food colouring "dyes" things) and an understanding of the semantics of dictionary words (e.g., "dye"/"die"). On the other hand, much of the humour encountered online, on TV panel shows, and in stand-up comedy routines, is based on contemporary topical knowledge of evolving pop-culture phenomena and news events, where a full appreciation of the humour relies on potentially esoteric knowledge, rather than common-sense (Highfield, 2015; Laineste, 2002), requiring highly complex reasoning and knowledge retrieval. We present examples of each joke type in Figure 1. The *homographic* pun exploits the dual meaning of "croaked" as 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.

To address the research gap, we perform an assessment of the ability of Large Language Models

[1] **CONTENT WARNING:** Some of the presented jokes may be considered offensive or distasteful to different individuals.

(LLMs) to explain a range of joke forms in a zero-shot setting, from commonly studied homographic and heterographic puns, to non-topical incongruity-based jokes and topical humour seen on Reddit. As a result, we present the first work on assessing whether state-of-the-art LLMs are equally able to explain different joke formats, from those that require an understanding of basic semantics and phonetics, to those that require esoteric knowledge of pop-culture and news events, therefore assessing the models' abilities to perform complex reasoning and knowledge retrieval.

We summarise our main contributions below.[2] In presenting these contributions, we address the primary research questions of (**RQ1**) whether different joke formats have an impact on the ability of LLMs to explain humour and (**RQ2**) by extension, whether existing works on computational humour representative of the different joke formats found in everyday life.

- We present a novel, balanced dataset of 600 jokes, categorised into traditional puns (homographic/polysemous and heterographic/phonetic), non-topical Reddit jokes (irony, stereotypes, cultural norms, etc.), and topical Reddit jokes (pop culture, news events, etc.). Each joke is paired with a high-quality, succinct, human-authored reference explanation, enabling the first empirical assessment of how joke format affects the ability of LLMs to explain humour.
- We use the aforementioned dataset to assess the accuracy and completeness of 4800 joke explanations from 8 state-of-the-art open- and closed-source LLMs of various sizes, including reasoning models (600 jokes * 8 models). We analyse these explanations using human evaluation and automatic metrics, including using an LLM judge.
- We qualitatively evaluate the models' abilities to understand complex humour through LLM-generated explanations of a topical online joke in the form of a case study.

2 Related Works

Humour Generation Early humour generation efforts relied on simple template-based systems to create puns (Valitutti et al., 2013; Ritchie, 2005), whilst contemporary works fine-tune deep learning models on extensive datasets of existing jokes (Garimella et al., 2020; Weller et al., 2020; He et al., 2019; Petrović and Matthews, 2013). More recently, the use of LLMs has become the primary paradigm (Chen et al., 2024; Mittal et al., 2022). Furthermore, Sun et al. (2022b) presented CUP, which turned the focus away from standalone joke generation to context-situated generation. Additionally, work has also been performed on humour-adjacent language forms such as tongue-twisters (Loakman et al., 2023b; Keh et al., 2023).

Humour Detection Similarly, humour detection has evolved from using lexical and syntactic feature-based methods (van den Beukel and Aroyo, 2018) to training language-model-based classifiers (Meaney et al., 2021; Wang et al., 2020; Weller and Seppi, 2019). Ao et al. (2022) additionally present work using humour to aid in the detection of political parody, whilst Meaney (2020) and Loakman et al. (2023a) investigated the impact of demographic variables on humour perception. Furthermore, some works have approached humour detection from a multimodal standpoint, using elements such as audience laughter to aid in detection (Hasan et al., 2019).

Humour Explanation Lastly, humour explanation provides an account for why a given text is funny. Consequently, this task requires models to exhibit extensive reasoning capabilities. Whilst explaining standalone puns usually only requires a model to identify the pun word and its alternative interpretations in the given context, more complex jokes require an understanding of common sense, social constructs, and esoteric knowledge to identify key incongruities. Unlike the task of detection, explanation requires models to not only be able to identify humour, but also "understand" it to provide a textual explanation. Miller et al. (2017) presented a shared task concerning humour, in which subtask 3 pertains to humour interpretation by tasking models with assigning the correct WordNet sense keys to given instances of pun words. Lima Inácio et al. (2023) tackle the humour explanation task indirectly by investigating what specific elements humour classifiers learn. Regarding the production of natural language explanations, Hessel et al. (2023) perform the task of humour explanation on multimodal data from caption contests in The New Yorker, generating explanations as to why a particular best-rated caption is funny.

Our work empirically investigates the performance gap for joke explanation on different types of humour, assessing the challenges presented by contemporary topical jokes in particular.

[2] We will release all of our code on GitHub. The central dataset has been supplied as supplementary material.

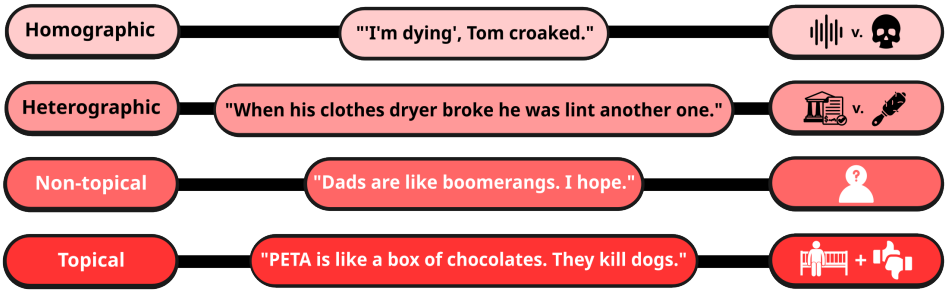

Figure 1: Examples of the 4 broad categories of joke compared in this work.

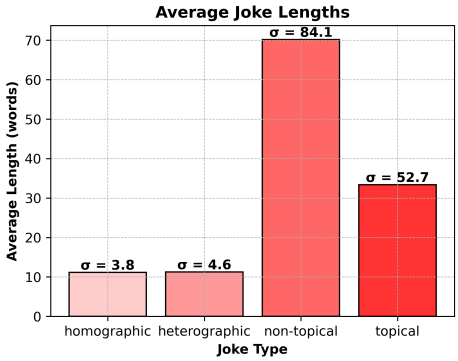

Figure 2: Average joke lengths by subset (with standard deviations presented on top of each bar).

3 Dataset Compilation

3.1 Existing Datasets

Many existing datasets focus on size, rather than annotation depth and consistency. For example, r/Jokes (Weller and Seppi, 2020) presents 550k jokes taken from Reddit with metadata such as upvote count, whilst SemEval 2017 Task 7 (Miller et al., 2017) consists of 3387 simple puns. Whilst Miller et al. (2017) add annotation of the puns' meanings via WordNet senses, there are no natural language explanations. Sun et al. (2022a) improve upon the SemEval dataset by crowdsourcing textual explanations for a subset, but remains limited to only puns. Furthermore, Hessel et al. (2023) present a dataset of multimodal humour from The New Yorker cartoon caption contest, providing written explanations and Wikipedia links to required world knowledge. Finally, whilst *ExplainTheJoke.com*[3] contains detailed joke explanations, the majority are simple puns, in addition to the explanations lacking objectivity, being highly editorialised for entertainment purposes.

[3] https://explainthejoke.com/

3.2 Our Dataset

In the following section, we outline the process of creating our own dataset with a larger variation of joke types and concise, objective explanations to form the backbone of our humour analysis. Overall, our dataset consists of 600 jokes, containing 150 of each type: homographic puns, heterographic puns, non-topical Reddit humour, and topical Reddit humour. Figure 2 presents the average joke lengths within each subset, whilst Figure 1 presents examples of each subtype (see Appendix A.6 for further examples).

3.3 Topical Online Humour

To identify jokes that require knowledge of real-world entities and events, we use a subset of *r/Jokes* (Weller and Seppi, 2020). Whilst Reddit contains potentially unsavoury material, the presence of strong opinions on topics such as politics and pop culture phenomena results in a rich array of topical humour. Whilst r/Jokes contains a plethora of different topics, not all joke instances rely on contemporary world knowledge. We therefore perform filtering based on upvotes and the presence of named entities (see Appendix A.2 for more details). Of the filtered jokes, we take the top 150 with the most upvotes. We removed jokes that did not require external knowledge, in addition to meta-level humour that relied on knowing that the joke came from Reddit (where this is not explicit).

3.4 Non-topical Online Humour

We additionally desire to compare the results of topical humour explanation with other common forms of internet humour that do not require esoteric world knowledge. These jokes often present different forms of humour than the simple puns that have been the focus of most existing works on automatic joke explanation, but require less esoteric knowledge than topical jokes. Consequently, we select another 150 top-scoring jokes from the r/Jokes dataset that do not require knowledge beyond what may be considered

"common sense" (e.g., cultural norms). Both the *topical* and *non-topical* subsets are checked by another linguist, who is not an author of this work, to ensure their correct categorisation. In effect, the 300 jokes that form the *topical* and *non-topical* categories are the top 300 highest-scoring jokes based on upvote-downvote ratio, categorised by their content.

3.5 Traditional Puns

We source an additional 300 simple pun-based jokes from SemEval 2017 Task 7 (Miller et al., 2017)), which has acted as the source data for many works on computational humour. Specifically, we take the first 150 jokes from both the homographic and heterographic training data for Subtask 3 as the basis of the *homographic* and *heterographic* joke subsets.

The jokes contained within the *topical* and *non-topical* subsets from Reddit may rely on the features of heterographic or homographic puns. However, the interpretation of these puns requires additional real-world knowledge besides the meaning of the pun words. For example, "*New Teslas don't come with a new car smell. They come with an Elon Musk.*", requires knowledge regarding Elon Musk's relationship to Tesla.

3.6 Ground Truth Joke Explanations

As a central contribution, we write ground-truth explanations for all jokes. Due to being written by an author of this work, we perform quality control using 3 native English-speaking computer science students. Each evaluator assessed 200 joke explanations (50 per type), each seeing different jokes.[4] We evaluate these explanations using the same evaluation criteria used for human evaluation in §5 (with the rubric presented in Appendix A.1), using a 0-5 Likert scale on the criteria of explanation accuracy and completeness. All evaluators scored all human-authored explanations a maximum score of 5 on both criteria or commented on required improvements, which were implemented. Minor explanation post-editing was required for less than 1% of explanations. Further details are in Appendix A.2.

3.7 Additional Dataset Features

We also present URLs to webpages containing the relevant knowledge required to understand jokes from the *topical* subset, including Wikipedia pages, news articles, and blog posts. We include phonetic transcriptions of all jokes after identifying model hallucinations stemming from a lack of phonetic knowledge (details in Appendix A.3). Neither of these fields are used in the following analyses.

[4]Participants were given a 30 GBP Amazon voucher for approximately 1.5 hours of evaluation.

4 Methodology

In the following section, we present our hypotheses relating to the research questions posed in §1, in addition to the models that we evaluate and our scoring criteria for the explanations.

4.1 Hypotheses

To answer the research questions presented in the introduction, we pose the following hypotheses:

- **H1**. Traditional puns (homographic and heterographic) will be easier for LLMs to explain than high-rated jokes from Reddit, owing to the former's frequent reliance on the semantics and phonetics of common words.
- **H2**. Heterographic puns will be more difficult for models to explain than homographic puns due to the former's reliance on phonetic similarity, which is not explicitly encoded in orthographic text.
- **H3**. Topical humour will be more difficult for models to explain than non-topical Reddit humour, due to the former's reliance on subtle references to contemporary pop culture and events, rather than common sense reasoning and general knowledge.
- **H4**. Larger model variants will perform better than smaller variants, particularly for topical humour, due to being able to store larger amounts of information in their parameters regarding specific events and individuals, whilst smaller models focus on retaining common general knowledge.

4.2 Models

To assess the ability of state-of-the-art LLMs in this task, we select the following suite of open- and closed-source models: **Llama 3.1** (8B and 70B) (Dubey et al., 2024), **Gemini 1.5** (Pro and Flash) (Georgiev et al., 2024), **GPT-4o** (standard and Mini) (Achiam et al., 2024). These models represent a range of sizes, using both the flagship large variants and their lightweight counterparts, allowing the investigation of model size on performance in this task. We additionally use variants of **DeepSeek-R1** to identify the abilities of reasoning-specific models. Specifically, we use the 8B and 70B variants based upon Llama 3.1 and 3.3, respectively.

Importantly, all of the selected models were trained on data pertaining to events up to at least 2023, with all jokes in the *non-topical* and *topical* subsets being posted to Reddit between 2008 and 2019. Therefore, we are fairly assessing the ability of these models to explain jokes that reference events contained within their training data, investigating the models' ability to correctly associate this knowledge with the abstract references made within topical jokes.

All models were tasked with producing an explanation around 100 words (based on the lengths of the gold standard explanations). Additional details, including the full prompt and specific models used, are presented in Appendix A.2.

4.3 Evaluation Criteria

To assess the quality of joke explanations generated by the models, we propose a scoring rubric with two core criteria, accuracy and completeness, each evaluated on a 6-point scale (0-5). Focusing on such criteria allows the application of the scoring rubric to all assessed joke types (such as real-world references in the *topical* subset constituting completeness, or pun word definitions in traditional puns). The full rubric can be found in Appendix A.1.

Accuracy entails whether or not the joke explanation contains correct material, which does not negatively score content where parts of the joke have not been explained but rather assesses the presence or absence of hallucinations and misunderstandings in a response. On the other hand, *Completeness* covers whether the fundamental elements of a joke have been addressed. Consequently, a joke may be fully explained (i.e., Completeness = 5), but additional information has been included which is incorrect (Accuracy $\approx$ 3). Conversely, an explanation may identify that a joke is a pun based on a specific word (i.e., Accuracy = 5), but omit explanations of the pun's meanings or other required information for understanding (i.e., Completeness $\approx$ 2).

Due to the significant amount of time required for rating 4800 explanations ($\approx$ 40 hours), 1 author (who is a native speaker of English and holds degrees in Linguistics) rated all explanations. However, a subset of 320 explanations (10 jokes * 4 joke types * 8 models) were re-annotated by 2 third-party computer science students (also native English speakers).[5] These annotations were used exclusively to verify the primary evaluator's reliability. The results showed moderate levels of agreement for both accuracy (α = .574) and completeness (α = .553) using Krippendorff's alpha (Krippendorff, 2019), as well as high positive Pearson's correlations (Pearson, 1895) for accuracy (r = .765) and completeness (r = .795), with $p < .001$, between the scores assigned by the primary human evaluator and third-party evaluators. Overall, 82% of the Accuracy scores and 84% of the Completeness scores from the primary evaluator and the 2 third-party evaluators differed by 1 or 0 points, resulting in an 80.5% agreement between the main evaluator and third-party evaluators when converted to a binary pass/fail (as in §5.1). Across the whole dataset, we additionally leverage the LLM as a Judge paradigm, using Qwen2.5-72B-Instruct, finding agreement for accuracy (α = .565) and completeness (α = .519) to also be substantial, additionally supported by Pearson's correlations of .641 and .602, both with $p < .001$, respectively. See Appendix A.7 for details on the LLM judge set up, and Appendix A.4 for a deeper analysis of the ratings from the third-party evaluators and LLM judge.

5 Human Evaluation

Average explanation lengths alongside accuracy and completeness scores can be seen in Figure 3. Firstly, the lengths of explanations remain consistent within models regardless of joke type, suggesting that the length of a joke does not necessarily impact the required explanation length. Notably, the Gemini-based models produce the shortest explanations across all joke types, whilst GPT-4o-based models consistently generate the longest explanations, opting for more verbose outputs. However, the shorter explanations produced by the Gemini models do not seem to negatively impact their performance in terms of accuracy or completeness. Despite being more concise, the Gemini models appear to be efficient, focusing on the essential elements of an explanation without adding superfluous detail.

In terms of overall performance, completeness scores are generally lower than accuracy scores across all models. This difference indicates that the models are more likely to miss key details in their explanations than to hallucinate incorrect information or misinterpret jokes entirely. This trend suggests that whilst the LLMs may recognise the core components of a joke, they often fail to fully elaborate on all relevant aspects that contribute to humour.

When comparing model performance, GPT-4o consistently outperforms all others, demonstrating the highest accuracy and completeness scores across joke types. Interestingly, GPT-4o Mini outperforms

[5] Participants were given a 60 GBP Amazon voucher for approximately 3 hours of evaluation.

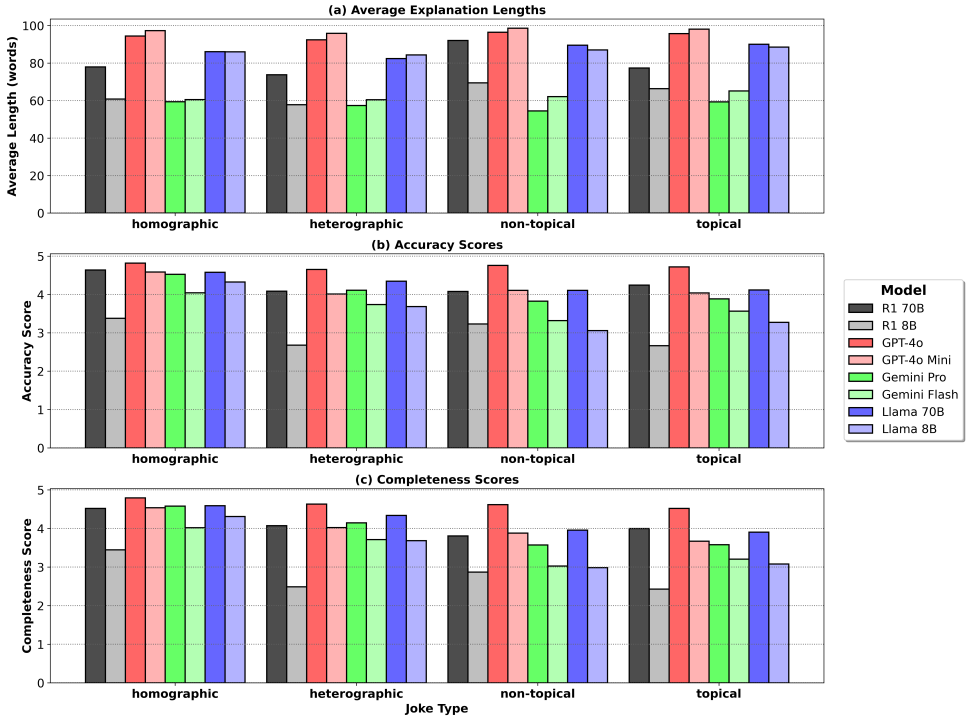

Figure 3: (a) Average joke lengths, (b) Average Accuracy score, (c) Average Completeness score, by model and joke type.

Gemini Pro in some specific cases, such as in accuracy on homographic jokes, as well as in both accuracy and completeness for non-topical and topical jokes. Additionally, open-source Llama 70B, outperforms most other models except GPT-4o, whilst the R1 models underperform.

5.1 Explanation Success Rate

For an explanation to be useful, it must be both accurate and complete, rather than scoring highly on only one criterion. As a result, in Figure 4 we present a binary categorisation of explanations, treating scores of 4 or above on both criteria as being "*good*" quality explanations (i.e., containing only minimal, non-intrusive errors, and/or omitting minor details). This facilitates a clear comparison of the challenges different joke types pose.

As illustrated in Figure 4, the success rate of explanations varies significantly across different joke types. Homographic jokes, where a word has multiple meanings but identical spelling, consistently yield the highest proportion of successful explanations across nearly all models. This suggests that it is easier to navigate the ambiguity of homographs, likely due to language models being aware of the polysemy of the different pun words. In contrast, heterographic jokes, which involve words that sound the same but are spelt differently, pose more difficulty for all models. GPT-4o is the only model that maintains a relatively high success rate in this category. The challenge likely stems from the additional complexity introduced by needing to recognise the phonetic similarity of the pun word whilst being trained on orthographic tokens. Non-topical jokes also show varying levels of performance. The non-topical subset consists of jokes from Reddit, including puns, incongruities and situational irony that cover a wide range of joke forms that have been rated highly by the Reddit userbase, causing the LLMs to struggle more severely than traditional puns. Lastly, topical jokes, which require contextual awareness of specific events, present the most notable challenges. While GPT-4o generates a relatively high proportion of "good" explanations, other models are substantially less successful. This is likely due to the subtle references to pop culture and real-world entities in the "topical" subset presenting a significant challenge to current state-of-the-art LLMs, requiring both high levels of reasoning in addition to a form of information retrieval where the information that needs retrieving is not always explicitly referred to.

We see a unique pattern of performance by the Deepseek R1 models. Whilst R1 70B performs well overall, the model's reasoning capabilities do not result in it outperforming any other tested model. On the other hand, the 8B variant is the worst-performing model. We hypothesise that this is due to the focus on "reasoning" surrounding mathematical and coding

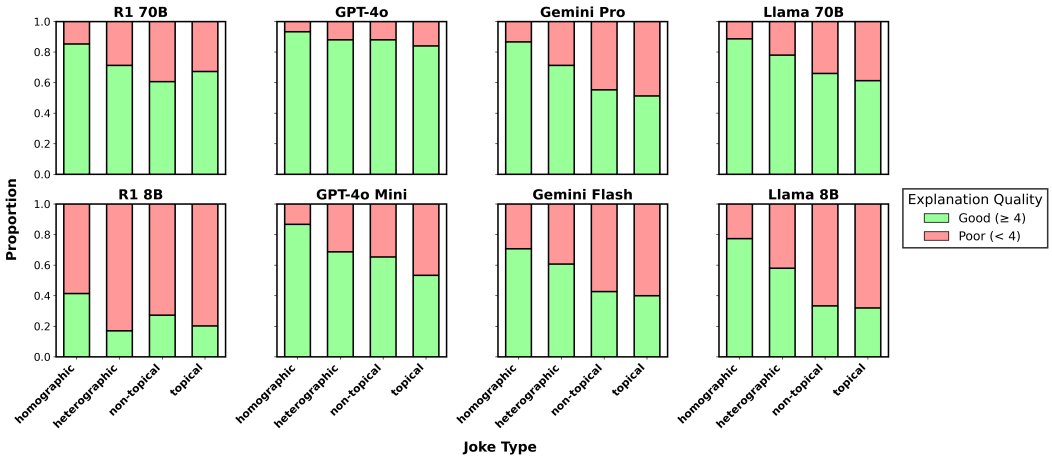

Figure 4: Proportion of "good" and "poor" joke explanations by model and joke type. A "good" explanation is defined as one that scores at least 4 on both Accuracy and Completeness.

problems, rather than "reasoning" in the general sense, such as identifying the common-sense incongruities that frequently cause humour, or "overthinking" the humour in simple puns. Furthermore, due to the reasoning steps taken, it may be the case that R1 8B begins with a moderately good answer (in line with the other models tested), but due to its uncertainty in explaining the jokes, the reasoning steps introduce many opportunities for hallucination, resulting in the smaller 8B model continuing to follow incorrect reasoning before presenting its final answer, therefore becoming progressively worse.

We perform a logistic regression analysis, which further supports our hypotheses. Model size has a strong and significant impact on explanation quality, with larger variants being far more likely to generate "good" explanations (β = 1.707, $p < 0.001$). This reinforces our finding that larger models significantly outperform smaller ones. Regarding joke difficulty, the results align with the predicted ordering. Homographic jokes are the easiest to explain (β = 0.583, $p < 0.001$), while non-topical (β = -0.511, $p < 0.001$) and topical jokes (β = -0.574, $p < 0.001$) are significantly harder. This confirms that jokes requiring external context pose a greater challenge to LLMs than those based on simple semantics and phonetics alone. Overall, the model is highly significant ($p < 0.001$) and explains a meaningful portion of the variance (pseudo $R^2 = 0.138$).

6 Automatic Evaluation

In addition to the previously presented human evaluation, we present the results of automatic metrics in Table 1. Whilst existing reference-based automatic metrics are not well designed for this task, we do confirm our originally hypothesised ordering of joke difficulties, with the explanations of homographic/heterographic puns scoring higher than the non-topical joke subset, which in turn scores higher than the topical jokes subset on most metrics.

	Het.	Hom.	Non-Topical	Topical
SacreBLEU	7.53	8.91	6.26	5.62
ROUGE-1	0.39	0.41	0.38	0.36
ROUGE-2	0.11	0.13	0.11	0.10
ROUGE-L	0.24	0.27	0.24	0.22
METEOR	0.32	0.34	0.28	0.27
BERTScore	0.88	0.89	0.87	0.87

Table 1: Automatic evaluation of LLM explanations by joke type, with human explanations as the reference. *Het* and *Hom* refer to homographic and homographic puns, respectively. Further breakdown by model and joke type can be found in Appendix A.5.

7 Case Study

Example LLM explanations of a randomly selected joke from the *topical* subset are shown in Figure 5. We omit the R1 reasoning models for space. The joke of interest refers to a 2018 internet phenomenon called the "Tide Pod Challenge," where teenagers would dare each other to consume the toxic laundry pods. The central element to the joke's explanation is the suggestion that this new "teenage snack food market" increased sales, ergo generating enough revenue for the company to afford advertising at the Super Bowl event. Interestingly, all of the full-size models inferred the reference to this challenge in their explanations. However, the smaller counterparts consistently omitted this information or presented misinterpretations in their explanations. GPT-4o Mini, for example, explains

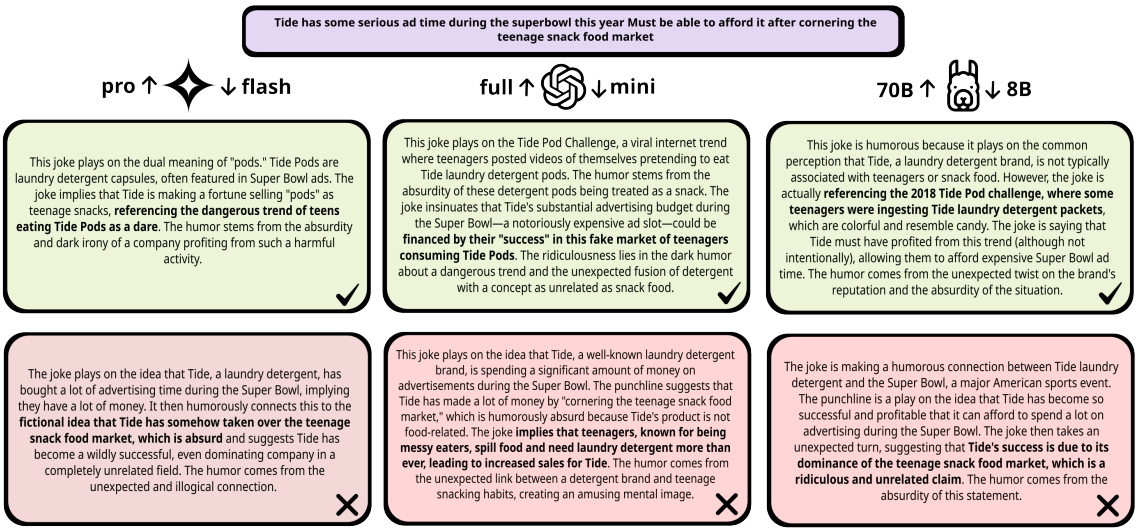

Figure 5: Case study of LLM explanations of a topical joke surrounding the "Tide Pod Challenge". From left to right: GPT-4o, Gemini 1.5, and Llama 3.1. The larger models are along the top, with the smaller variants along the bottom.

the joke by suggesting that "teenagers are known for being messy eaters", which caused increased sales. On the other hand, Llama 8B and Gemini Flash focus on the idea that the humour stems from "absurdity", without making any explicit mention of the real-world phenomenon, thereby failing to inform the user of important context. Further discussion of explanation characteristics can be found in Appendix A.6.

8 Discussion

Firstly, regarding **H1**, that traditional puns will be easier for models to explain than highly-rated jokes from Reddit, we found this to be true, with the non-topical and topical subsets significantly impacting explanation success rates.

Similarly, **H2** concerned the hypothesis that homographic puns (which rely on polysemy) are easier to explain than heterographic puns, which was found to be the case in our tests. We observed that heterographic puns more frequently lead to hallucination, owing to their reliance on phonetic characteristics, with LLMs having only implicit knowledge of how similar different words may "sound".

H3 posited that, of the online Reddit humour, the *topical* subset would present the most difficulty for LLMs. Our testing revealed a statistically significant decrease in explanation success rates for topical humour compared to all other joke formats, with many explanations falling short of correctly identifying contemporary real-world references.

Finally, **H4** stated that the larger models would outperform smaller variants. Our results showed that this was the case, with R1 70B outperforming R1 8B, GPT-4o outperforming Mini, Gemini Pro outperforming Gemini Flash, and Llama 70B outperforming Llama 8B.

As predicted, the performance discrepancy was most noticeable within the *topical* subset. Whilst these jokes were significantly shorter than the *non-topical* subset (see Figure 2), their reliance on subtle references to esoteric entities and events required the models to identify information that was not explicitly stated if they were to correctly identify the humour. This creates an environment ripe for hallucinations, as the models attempt to generate an explanation for a joke where they often cannot identify the location of the punchline without additional information. We further hypothesise that the larger models struggled less with topical jokes, relative to smaller models, due to the additional parameters allowing the storage of more specific knowledge that does not appear in the training data as frequently as common sense concepts (e.g., that hot things cause burns, or that sexual relations with a blood relative are taboo).

9 Conclusion

We presented the first empirical assessment of how well LLMs are able to explain the humour found within jokes of varying types, including simple puns and jokes that rely on contemporary topical knowledge. Our findings demonstrate that the primary focus on puns in prior work is not representative of the abilities of LLMs to accurately and comprehensively explain a more broad array of joke types. Furthermore, we also observe that no existing state-of-the-art models can consistently explain jokes of any type correctly.

Limitations

In this work, we investigated the extent to which joke format and complexity affect the ability of LLMs to generate accurate and complete explanations. However, the nature of topical humour is that it is ever-evolving, with jokes relating to developing news stories and pop culture events continuously appearing. Consequently, our work focuses only on a subset of possible jokes and does not assess the ability of models to explain highly recent jokes, of which there may be only limited knowledge contained within a given model's training data (as the volume of reporting and discussion on a given topic grows over time).

Additionally, we performed a comparatively small-scale analysis of 600 jokes (150 per type). Whilst this is necessary to enable manual human assessment of the generated explanations as well as facilitate the authoring of gold-standard reference explanations, a wider range of jokes would allow deeper insight.

Due to the jokes being from secondary sources (i.e., not novel jokes written for the purpose of this work), it is likely that the majority of the jokes have been seen before by the LLMs. However, explanations relating to these jokes are much less likely to be present. For the Reddit jokes in particular, the presence of explanations in the training data is reliant on individual Reddit users asking for an explanation within the relevant thread, and someone replying. Whilst this is possible, we sorted the r/Jokes dataset based on upvotes, and intuitively it is less likely that popular jokes require explanations, with Reddit's userbase already possessing the required knowledge in most cases (hence the jokes receiving a significant amount of upvotes). We manually verified a subset of the jokes and found readily available explanations to be rare.

Whilst a substantial portion of our evaluation is performed by 1 author, we robustly prove the reliability of the evaluation through comparisons with third-party annotators and an LLM-judge. Owing to using Likert scales with a degree of inherent subjectivity, agreement statistics in the interval between 0.4 and 0.6 are considered good, whilst our general correlations among the different methods are very high.

Finally, myriad joke formats exist, and more fine-grained categorisation of joke types is possible.

Ethics Statement

We believe in and firmly adhere to the Code of Conduct in the performance of this work and the methods involved. Whilst the dataset element of our benchmark contains potentially offensive materials (reflecting the often controversial nature of humour found on Reddit), the aim of this resource is to offer explanations of potentially esoteric jokes in order to clarify where particular parties may be perceiving humour from. In this work, we do not create novel humour, nor present the included dataset as appropriate for training humour generation models. Whilst detoxifying and debiasing language models is a growing interest in the NLP community, humans are fundamentally opinionated and biased in ways that others may take offence at. Whilst we agree language models should not perpetuate these biases, the authors of this work believe that, if queried, language models should be able to make use of knowledge pertaining to human beliefs and opinions in offering impartial, objective explanations of phenomena such as humour. We therefore do not view the creation of an *explanation* for a joke to be equatable with an *endorsement* of said joke, but rather as an ideologically neutral task. We additionally consciously write our humour explanations in an objective manner, not lending unjust credence to unverified scandals or presenting opinions as facts.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Xiao Ao, Danae Sanchez Villegas, Daniel Preotiuc-Pietro, and Nikolaos Aletras. 2022. Combining humor and sarcasm for improving political parody detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1800–1807, Seattle, United States. Association for Computational Linguistics.

Salvatore Attardo. 2008. *A primer for the linguistics of humor*, pages 101–156. De Gruyter Mouton, Berlin, New York.

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 and virtual meeting. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. 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 2046–2056, Hong Kong, China. Association for Computational Linguistics.

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).

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. *arXiv preprint*.

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.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Klaus Krippendorff. 2019. Content analysis: An introduction to its methodology.

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.

Marcio Lima Inácio, Gabriela Wick-pedro, and Hugo Goncalo Oliveira. 2023. What do humor classifiers learn? an attempt to explain humor recognition models. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 88–98, Dubrovnik, Croatia. Association for Computational Linguistics.

Tyler Loakman, Aaron Maladry, and Chenghua Lin. 2023a. The iron(ic) melting pot: Reviewing human evaluation in humour, irony and sarcasm generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6676–6689, Singapore. Association for Computational Linguistics.

Tyler Loakman, Chen Tang, and Chenghua Lin. 2023b. 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.

J. A. Meaney. 2020. Crossing the line: Where do demographic variables fit into humor detection? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 176–181, Online. Association for Computational Linguistics.

J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119, Online. Association for Computational Linguistics.

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.

Karl Pearson. 1895. Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London Series I*, 58:240–242.

Saša Petrović and David Matthews. 2013. Unsupervised joke generation from big data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–232, Sofia, Bulgaria. Association for Computational Linguistics.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. *arXiv preprint*.

Graeme Ritchie. 2005. Computational mechanisms for pun generation. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*, Aberdeen, Scotland. Association for Computational Linguistics.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, et al. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. *arXiv preprint*.

Jiao Sun, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Tagyoung Chung, Jing Huang, Yang Liu, and Nanyun Peng. 2022a. ExPUNations: Augmenting puns with keywords and explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4590–4605, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiao Sun, Anjali Narayan-Chen, Shereen Oraby, Shuyang Gao, Tagyoung Chung, Jing Huang, Yang Liu, and Nanyun Peng. 2022b. 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.

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.

Sven van den Beukel and Lora Aroyo. 2018. Homonym detection for humor recognition in short text. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 286–291, Brussels, Belgium. Association for Computational Linguistics.

Minghan Wang, Hao Yang, Ying Qin, Shiliang Sun, and Yao Deng. 2020. Unified humor detection based on sentence-pair augmentation and transfer learning. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 53–59, Lisboa, Portugal. European Association for Machine Translation.

Orion Weller, Nancy Fulda, and Kevin Seppi. 2020. Can humor prediction datasets be used for humor generation? humorous headline generation via style transfer. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 186–191, Online. Association for Computational Linguistics.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. 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 3621–3625, Hong Kong, China. Association for Computational Linguistics.

Orion Weller and Kevin Seppi. 2020. The rJokes dataset: a large scale humor collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6136–6141, Marseille, France. European Language Resources Association.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large Language Models Are Human-Level Prompt Engineers. *arXiv preprint*.

A Appendix

A.1 Evaluation Rubric

1. **Explanation Accuracy (0-5 points)**
   - **5 points**: The explanation is fully accurate, correctly identifying the joke's main humour and context without errors.
   - **4 points**: The explanation is mostly accurate with minor errors or omissions that do not significantly impact understanding.
   - **3 points**: The explanation contains mostly correct elements but includes noticeable errors that may hinder understanding.

- **2 points**: The explanation is largely inaccurate, with only a few correct points amidst significant errors.
- **1 point**: The explanation is mostly incorrect, with minimal recognition of the joke's intent.
- **0 points**: The explanation is entirely incorrect, failing to identify the humour or context correctly.

2. **Explanation Completeness (0-5 points)**
   - **5 points**: The explanation is comprehensive, covering all relevant aspects and nuances.
   - **4 points**: The explanation covers most aspects but misses minor details.
   - **3 points**: The explanation addresses the central humorous element but omits some important details that maximise understanding.
   - **2 points**: The explanation is incomplete, addressing only basic aspects while missing significant details.
   - **1 point**: The explanation is minimal, covering only a small part of the joke.
   - **0 points**: The explanation is entirely superficial or irrelevant, failing to address any key components.

A.2 Implementation Details

r/Jokes Filtering In order to select 150 jokes from the r/Jokes dataset to represent the *topical* subset. we sort the original dataset by the assigned "score" field (calculated by Weller and Seppi (2020) based on the upvote-downvote ratio) to find the most popular jokes. Following this, we performed Named Entity Recognition (NER) via SpaCy, keeping jokes if they were found to contain entities of the following types from the SpaCy NER taxonomy: ORG (organisation), FAC (facility), WORK_OF_ART, LAW, ORG (organisations), PROD (products), EVENT, and PERSON. With these entity types consisting mainly of proper nouns of people, places and events, this filtering allows us to easily identify r/Jokes entries that are likely to require potentially esoteric world knowledge to perceive the humour within, as opposed to common sense or definitions of dictionary terms. This process reduced the dataset to just over 153k samples, of which we then take the top 150 highest scoring following manual verification.

Prompt For all models, we used the following prompt: `"Explain the following joke (presented in square brackets) in approximately 100 words.\n\n[JOKE]\n\n"`. This is based on good practice for clearly marking the boundaries between the task and the joke instance, whilst the 100-word length is based on the average length of the human-authored explanations.

Models In our work, we specifically made use of the following model variants: GPT-4o (gpt-4o-2024-05-13), GPT-4o Mini (gpt-4o-mini-2024-07-18), Gemini 1.5 Pro (gemini-1.5-pro-001), Gemini 1.5 Flash (gemini-1.5-flash-001), Llama 3.1 8B (Meta-Llama-3.1-8B-Instruct), Llama 3.1 70B (Meta-Llama-3.1-70B-Instruct), Deepseek R1 8B (DeepSeek-R1-Distill-Llama-8B), and DeepSeek R1 70B (DeepSeek-R1-Distill-Llama-8B). For GPT and Gemini, these refer to the stable releases available at the time (August 2024). All models had their settings (e.g., temperature) at default. We use the Google API with all user-accessible safety limits disabled, due to the potentially offensive content of some of the jokes.

Inference OpenAI models were prompted via the OpenAI API, whilst the Google models used the Google GenerativeAI API. Llama 8B was run on a desktop PC with an RTX 4080 Super, whilst Llama 70B, both R1 models, and the LLM Judge was run on an HPC node using two A100s.

Participants We recruited 2 third-party participants (i.e., not authors of this work) to perform evaluation on a subset of 240 explanations to calculate agreement. In compensation, they received a 60 GBP Amazon voucher for approximately 3 hours of work. Furthermore, 3 participants also evaluated the gold standard reference explanations (200 each), for which they received a 30 GBP Amazon voucher. Their involvement was conducted with the full ethical approval of the institution that the authors represent, with consent forms signed and a participant information sheet provided. The third-party checks of joke categorisation did not involve monetary compensation.

A.3 Dataset Enhancement

We additionally provide phonetic transcriptions in the International Phonetic Alphabet (IPA) of all jokes in the dataset. To do so, we utilise Anthropic's Claude Sonnet 3.5 (20241022), owing to advanced handling of factors such as abbreviations over traditional grapheme-to-phoneme (G2P) models. A subset of these transcriptions were manually verified by an author of this work who is a trained phonetician. We additionally present a test phrase of "*The pikachu weighs 45 lbs and is 5ft 9". He works at the CIA but previously worked at NASA"*, in order to verify its handling

of non-dictionary terms, abbreviations, initialisms and acronyms, which it did successfully. Specifically, we request transcriptions of the General American (GenAm) accent, owing to the primary userbase of Reddit with L1 English being North Americans.

When constructing the prompt, we provided a clear task description, also to Claude Sonnet, and requested that it construct its own prompt. Prior work has shown this to be an effective strategy (Zhou et al., 2023). The resulting prompt is as follows:

You are a linguistic transcription tool. For any text samples provided:\n\n 1. Convert EVERY sample into IPA format without exception - no samples should be skipped or partially transcribed\n 2. Use General American (GenAm) pronunciation standards consistently\n 3. If uncertain about any transcription, ask for clarification rather than making assumptions\n 4. Do not proceed without transcribing ALL text samples provided\n\n Example input:\n 'Hello world'\n\n Expected output:\n həˈloʊ wɚld\n\n Please proceed with the transcription."

A.4 Additional Annotator & LLM Judge Results

Figure 6 and Figure 7 present the results of the evaluation from the 2 third-party evaluators we recruit to analyse a subset of joke explanations. Overall, the same trend that was viewed in Figure 3 and Figure 4 can be seen, with the order of joke difficulty being the same for these evaluators on the tested subset. Owing to the smaller subsets (10 jokes per type), some minor differences can be seen in the ordering of performance for some model/joke type combinations. The same is provided in Figure 8 and Figure 9 for our use of Qwen 72B Instruct as an LLM judge.

A.5 Additional Automatic Evaluation

Full results of automatic evaluation, broken down by both model and joke type, can be seen in Table 2.

A.6 Additional Qualitative Analysis

Reddit Joke Types Table 3 and Table 4 present examples of non-topical and topical jokes from the Reddit subsets, respectively. Regarding the non-topical jokes, the first example plays on common knowledge regarding religious practices and the focus of churches on the bible. Joke 2 is a form of deadpan humour that refers to the higher mortality rates of children who are not vaccinated against preventable diseases. The third joke presents phonetic wordplay, requiring the reader to interpret "12345678" as "one 2, 3 fours..." and so on. Finally, joke 4 plays on the mathematical concept of imaginary numbers. On the other hand, regarding the topical jokes, joke 1 makes reference to the media franchise of Alien v. Predator, whilst additionally referring to the 2019 news phenomena where a Facebook event was created to convince people to storm Area 51 to find aliens. The second joke refers to the Marvel film franchise and their penchant for extended scenes following the credits, in addition to Stan Lee's real-life death. On the other hand, joke 3 makes light of a celebrity scandal involving R Kelly and sexual assault, performing wordplay that requires the letters "art" to be removed from "rap artist" to identify the particular scandal involving R Kelly. Finally, the fourth joke refers to Barack Obama losing his presidency to Donald Trump, and therefore being "unemployed" and "recently evicted" (from the White House).

Explanation Characteristics During the process of evaluating LLM responses, several unique response characteristics were observed. Firstly, the Google API refused to allow the Gemini models to produce explanations for some jokes, likely due to triggering content safeguards (5 explanations across 3 jokes, one of which the request was only blocked by Flash). Such instances were not notably any more/less offensive than other humour found within the Reddit joke subsets. In some instances, models provided textual responses regarding their refusal to address a particular joke (however, both cases only account for ≈1% of all jokes). Furthermore, in some instances, we observed that explanations contained subjective judgments as to the appropriateness of a particular joke, with the humour being deemed "unacceptable", without acknowledging the nuance and subjectivity of such a claim. Importantly, potentially offensive content was not exclusively found in the non-topical and topical Reddit joke subsets, with many of the traditional puns containing sexual innuendo and derogatory terms. Furthermore, whilst all of the models tested are capable of producing high-quality coherent text, we observed some instances of the Llama models resorting to degenerate patterns, such as being stuck in repetitive loops. Again, these account for a very small number of jokes (<1%). Regarding the heterographic pun subset of jokes, we noticed many instances of hallucination whereby the LLMs would explain that a punning word is phonetically similar to a word that shares little phonetic similarity. This is demonstrative of the challenge that heterographic puns present to LLMs, relying on phonetic characteristics of words, whilst such models are trained on orthographic text and therefore have limited phonetic knowledge.

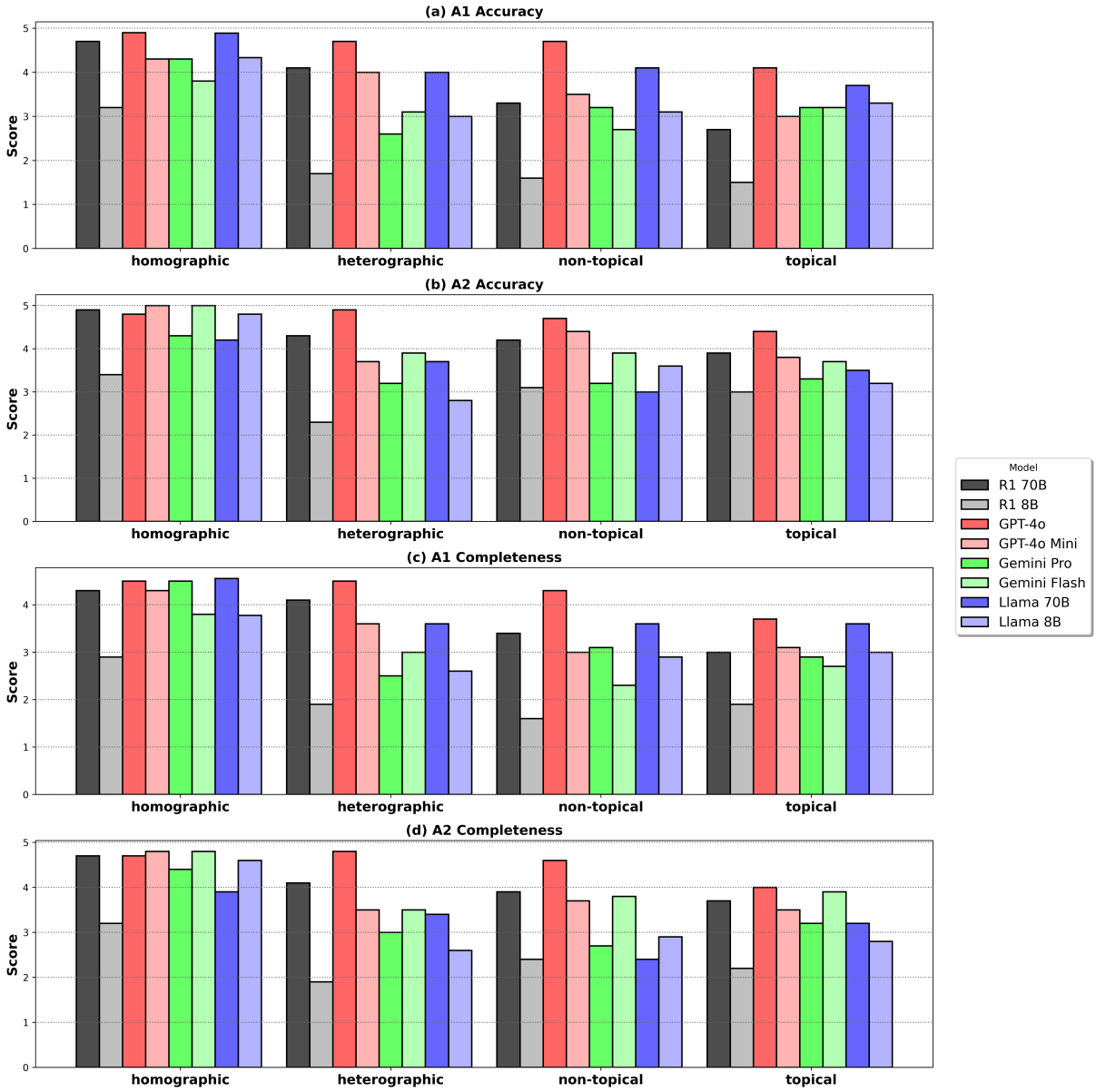

Figure 6: Human evaluation results on the metrics of accuracy and completeness on a subset of 320 explanations (10 jokes * 4 types * 8 models). A1 and A2 refer to two different annotators.

A.7 LLM as a Judge Models and Prompts

When selecting our model judge for automatic evaluation, we required that candidates not be from the same model family as any of those under evaluation or from the same family as each other. Our first model is the Mistral Mixture-of-Experts (Jiang et al., 2024) derivative, Prometheus 2 8x7B (Kim et al., 2024), which was specifically developed for LLM-based evaluation. The other two models are Qwen 72B Instruct (Qwen et al., 2025) and Gemma 2 9B Instruct (Riviere et al., 2024). These were selected as the highest performing models in the Hugging Face Judge Arena[6] (with preliminary results disabled) that met our previously outlined criteria. At the time of selection (04/02/2025). Qwen was ranked 6th with 1203 ELO and Gemma 13th with 1157. The weights for the models were obtained from their respective Hugging Face model repositories. Overall, we found Qwen alone to demonstrate the higher alignment with our human evaluation (versus a jury ensemble or any other individual judge).

Judge hyperparameters were all set to the model default except for temperature which was fixed to 0.1 for all models rather than zero based on prior work (Sanh et al., 2022).

To obtain scores from each judge, we used the following prompt. The scoring criteria was the same

[6] https://huggingface.co/spaces/AtlaAI/judge-arena

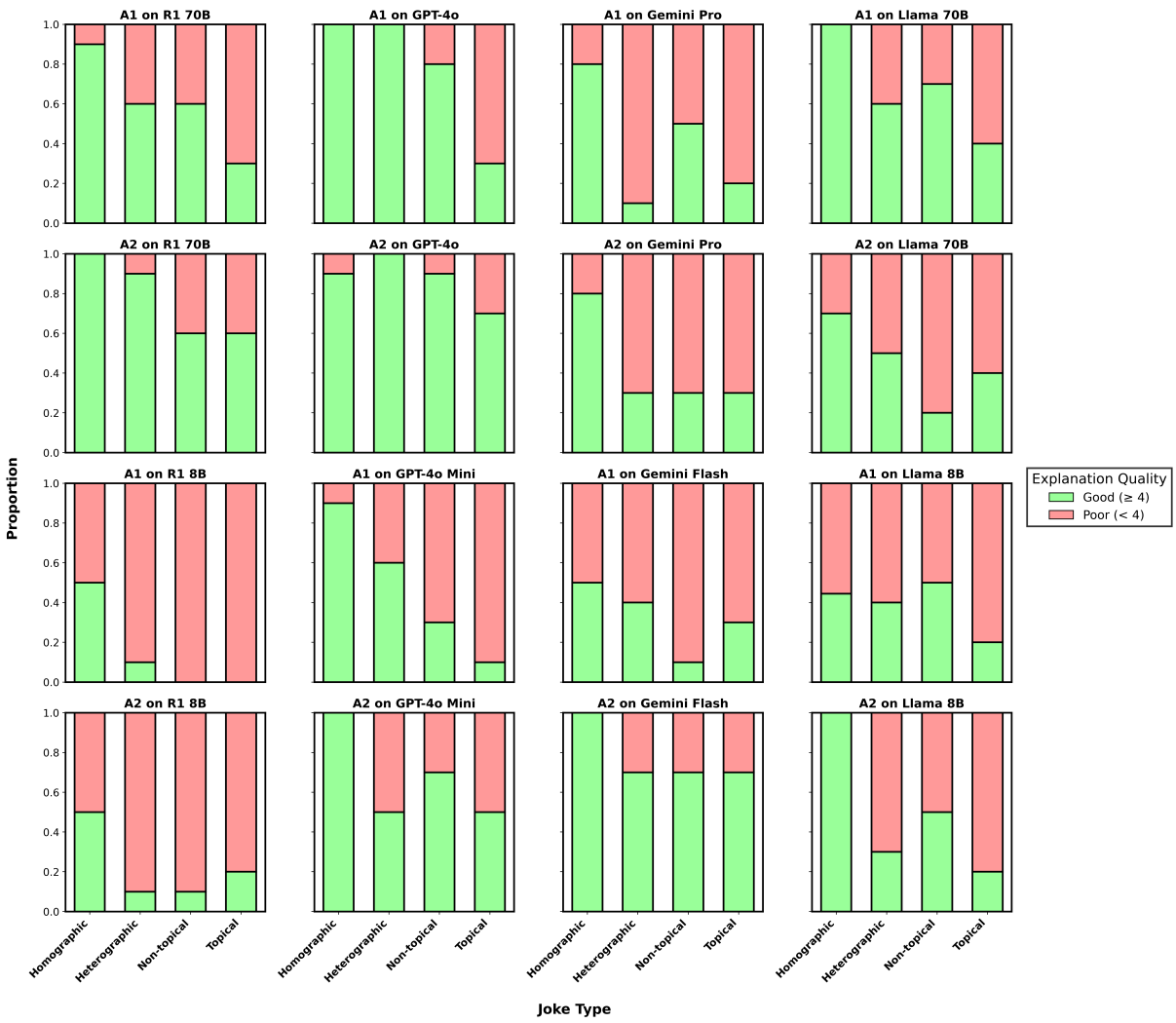

Figure 7: Binary evaluation results on a subset of 320 explanations (10 jokes * 4 types * 8 models). A1 and A2 refer to two different annotators.

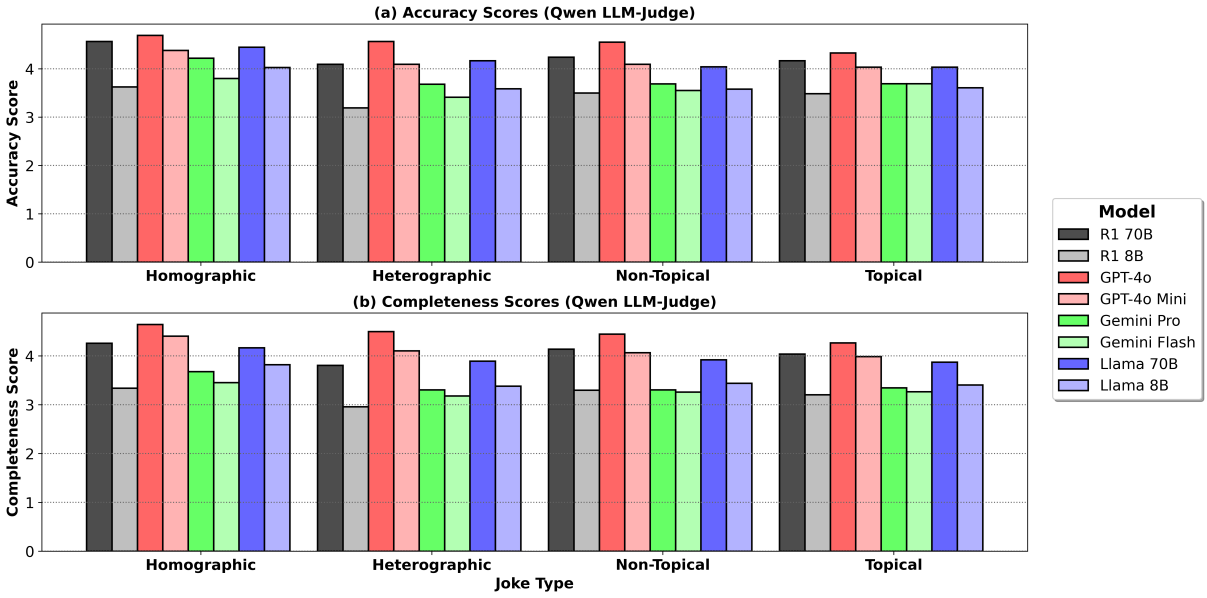

Figure 8: LLM-judge results from Qwen 72B on the metrics of Accuracy and Completeness across model and joke type.

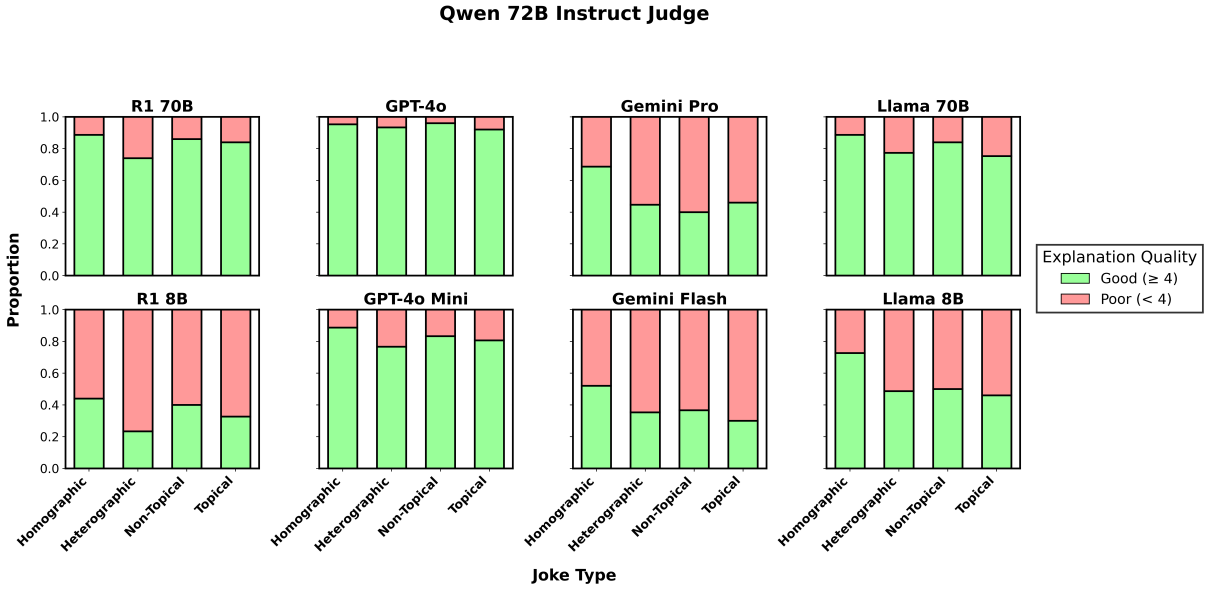

Figure 9: Binary explanation success results from the LLM-judge of Qwen 72B across model and joke type.

as given to the human annotators with the exception of score categories which were listed in ascending order rather than descending. This was done to better align with the training prompt of Prometheus.

You are an expert judge evaluating explanations of jokes. You will be given a joke, a reference explanation (human-authored), and a model-generated explanation to evaluate.\n\n You must evaluate the model's explanation on the matter of {criteria} using only the following scoring criteria:\n\n {scoring_criteria}\n\n The joke:\n {joke}\n\n Reference explanation:\n {reference_explanation}\n\n Model explanation to evaluate:\n {model_explanation}\n\n Provide a score for this joke using only the scoring criteria provided. You must only respond with a single number between 0 and 5. You must produce no other output.

Model	Metric	Hom.	Het.	Non-Topical	Topical
GPT-4o	SacreBLEU	8.51	10.15	7.85	7.09
	ROUGE-1	0.41	0.43	0.41	0.40
	ROUGE-2	0.12	0.15	0.13	0.12
	ROUGE-L	0.25	0.28	0.25	0.23
	METEOR	0.37	0.39	0.32	0.32
	BERTScore	0.88	0.89	0.88	0.87
GPT-4o Mini	SacreBLEU	7.06	9.27	6.12	5.30
	ROUGE-1	0.38	0.42	0.38	0.37
	ROUGE-2	0.10	0.14	0.11	0.10
	ROUGE-L	0.23	0.27	0.23	0.22
	METEOR	0.33	0.37	0.30	0.30
	BERTScore	0.88	0.89	0.87	0.87
Gemini Pro	SacreBLEU	7.02	8.15	4.71	4.54
	ROUGE-1	0.40	0.41	0.33	0.35
	ROUGE-2	0.10	0.13	0.09	0.09
	ROUGE-L	0.25	0.28	0.22	0.21
	METEOR	0.30	0.32	0.22	0.24
	BERTScore	0.88	0.88	0.87	0.87
Llama 70B	SacreBLEU	9.13	10.60	7.95	7.27
	ROUGE-1	0.42	0.43	0.41	0.39
	ROUGE-2	0.13	0.14	0.14	0.12
	ROUGE-L	0.25	0.27	0.26	0.23
	METEOR	0.36	0.37	0.31	0.30
	BERTScore	0.88	0.89	0.88	0.87
Llama 8B	SacreBLEU	8.41	9.76	7.16	5.83
	ROUGE-1	0.39	0.42	0.40	0.36
	ROUGE-2	0.12	0.14	0.12	0.10
	ROUGE-L	0.24	0.27	0.24	0.22
	METEOR	0.34	0.36	0.30	0.28
	BERTScore	0.88	0.89	0.87	0.87
R1 70B	SacreBLEU	7.73	8.47	6.27	5.89
	ROUGE-1	0.39	0.41	0.37	0.36
	ROUGE-2	0.10	0.12	0.11	0.10
	ROUGE-L	0.22	0.24	0.23	0.21
	METEOR	0.33	0.34	0.28	0.27
	BERTScore	0.88	0.88	0.87	0.87
R1 8B	SacreBLEU	5.25	6.48	4.85	4.12
	ROUGE-1	0.35	0.37	0.36	0.33
	ROUGE-2	0.08	0.10	0.09	0.08
	ROUGE-L	0.21	0.23	0.22	0.20
	METEOR	0.25	0.28	0.24	0.23
	BERTScore	0.87	0.88	0.87	0.86

Table 2: Automatic evaluation results for each tested model, broken down by joke type. We use human-authored explanations as the ground truth reference.

	Non-topical Joke Examples
1	"What do you call a book club that's been stuck on one book for years? Church."
2	"TIL unvaccinated children are less likely to be autistic. Because they are more likely to be dead."
3	"Set your wifi password to 2444666668888888. So when someone asks tell them it's 12345678."
4	"My girlfriend is like the square root of -100. A solid 10, but also imaginary."

Table 3: Examples of jokes from the non-topical subset with different types of common sense knowledge.

	Topical Joke Examples
1	"If America is storming Area 51 then the Europeans can storm the Vatican. We'll take the aliens, you get the predators."
2	"For anyone attending Stan Lee's funeral... Make sure you stay after the ceremony is finished."
3	"R Kelly is really changing the rap game. He takes the art out of rap artist."
4	"Obama smoked weed growing up, and now look where he is today. Unemployed with two kids and recently evicted."

Table 4: Examples of jokes from the topical subset with different types of assumed world knowledge.