"ARE YOU JOKING?": INVESTIGATING COGNITIVE PRAGMATIC ABILITIES IN LLMS

Alexander MacNeil
BA Hons Linguistics with French, 2025

Abstract: In recent years, large language models (LLMs) have seen increased scaling, more efficient modelling, and a surge in usage. This has prompted interest from linguists and computer scientists into the natural language understanding and production capabilities of these models across different linguistic domains. Consequently, a literature which seeks to benchmark these capabilities has emerged, however not without issues. A critical area of linguistic interest which has gone relatively unexamined until recently is the domain of pragmatics. While benchmarks have been proposed within the last year, they suffer from failures in experimental design and do not account for the differences between humans as conscious communicative agents, and LLMs as statistical models. Taking these shortcomings into account, I propose humour as a mechanism for testing the extent to which LLMs can simulate human cognitive traits required in traditional accounts of pragmatic behaviour. I test a model (ChatGPT4o) across a humour comprehension task and a humour production task, having it generate 50 jokes according to a semantic script and then judge them to understand how LLMs perform in a novel task that requires strong command over pragmatic skills. I find that the model performs moderately to poor across all tasks, indicating that LLMs struggle in simulating pragmatic behavioural traits in novel tasks, as well as failing at basic reasoning skills. This challenges previous notions of strong world knowledge and broad pragmatic skills in LLMs by demonstrating weaknesses in models to apply trained skills across testing paradigms, while contributing to the literature on linguistic and pragmatic benchmarking.

Keywords: LLMs, Pragmatics, Theory of Mind, Cognition, Humour

Supervisor: Dr Christine Cuskley

1. Introduction

Within past 5 years, AI has seen a huge growth in popularity, compute and output abilities, driving interest from academics across various fields of study in the form of benchmarks — experimental tasks or groups of tasks designed with the intention of testing a large language model (LLM)'s capabilities in a specific domain (Ye 2023; OpenAI 2024; Srivastava et al., 2022). In response, a steady growing body of linguistic benchmarks has emerged, seeking to answer questions about their performance capabilities with regards to human language, with recent interest in pragmatics. However, current widely used linguistic benchmarks tend to overlook pragmatic competency, and those which seek to study it as a specific skill suffer from shortfalls in experimental design and often fail to account for differences in how humans and LLMs engage in pragmatic behaviour. This study proposes a benchmark specifically designed around the differences in cognitive architecture between LLMs and humans, using humour as a means of ascertaining to what extent LLMs can simulate the cognitive tenets of human behaviour required for pragmatic behaviour. In Section 2, I explain the current literature surrounding linguistic benchmarks and pragmatic benchmarks, demonstrating current failings in this area, before proposing my experiment, hypothesis and predictions. In section 3, I explain my methodology of testing humour comprehension and production across 3 tasks. In section 4, I provide the findings of my experiment. In section 5, I discuss these results, their significance in broader linguistic benchmarking literature, and the limitations of this experiment. In Section 6, I make my final remarks on the study, reiterating the steps I took in each section, discussing what it achieves overall for the literature concerning linguistic benchmarking.

2. Background

While there is an existing, expanding body of literature that seeks to establish benchmarks across linguistic categories, it sees little to no engagement from the LLM industry. Only a small handful 'linguistic' benchmarks see common industry-wide use in testing, for example, MMLU (Hendrycks et al., 2021), (super)GLUE (Wang et al., 2019), and BIG Bench (Srivastava et al., 2022). Yet, there are profound limitations with these benchmark regarding

difficulty, contamination, and to what extent these benchmarks can be considered assessments of language capability.

Difficulty is a consistent issue faced in benchmarking, with widely used benchmarks like SuperGLUE, having been solved (models consistently achieving human or above human scores) in less than 18 months (Srivastava et al., 2022), and BIG Bench having 65% of its tasks solved by a model within a year (Chowdhery et al 2022). While much of the progress models have seen in with these benchmarks is down to improvements in the linguistic capabilities of LLMs (Sravanthi et al., 2024), this is not always the case. Contamination, another prevalent issue in benchmarking, occurs when a model incorporates a benchmark into its training data, effectively breaking the benchmark (White et al., 2024). This is an established issue, notably with evidence of ChatGPT 4 having trained itself on BIG Bench (OpenAI, 2024). The extent to which these tests can be considered 'linguistic' assessments also requires scrutiny. MMLU, while designed as a cross-domain benchmark for world knowledge and textual understanding, only assesses this through multiple-choice question answering (MCQA), without requiring explanations to ensure that the model understands the questions it is being asked. SuperGLUE, while testing a wider range of linguistic areas of understanding (such as pronoun resolution, speaker belief assessments, on top of the language understanding required in MMLU), still only tests for a model's ability to understand language (not its performance in that domain) and covers a relatively narrow scope of linguistic competencies. This, thus, highlights issues in the linguistic coverage of these benchmarks.

One particularly under-assessed yet increasingly relevant area of linguistic capability is pragmatics. While NLP/NLU benchmarks have previously focused on areas like semantic understanding and performance, pragmatic understanding is equally fundamental to human communication (Sravanthi et al., 2024; Kabbara 2019; Stemmer 1999). Widely applied tests are often outdated (such as the Winograd Schema (Kacijan et al., 2023)), and areas such as cognitive pragmatics have few benchmarks to work with. While current industry-wide benchmarks are severely outdated, there is current literature dedicated to assessing the pragmatic understanding of LLMs. Notable recent contributions to pragmatic testing are the Pragmatic Understanding Benchmark (PUB) and Length-Normalised Relative Scoring testing (Sravanthi et al., 2024; Wu et al., 2024). PUB seeks to test understanding of the pragmatic skills of implicature, presupposition, reference and deixis, while Wu et al. (2024) expand on these areas with other pragmatic elements such as humour, irony, metaphor and

understanding social norms. They test for competence in these areas by presenting models with situations or scenarios and having them decipher meaning or predict the most suitable response (see fig. 1)

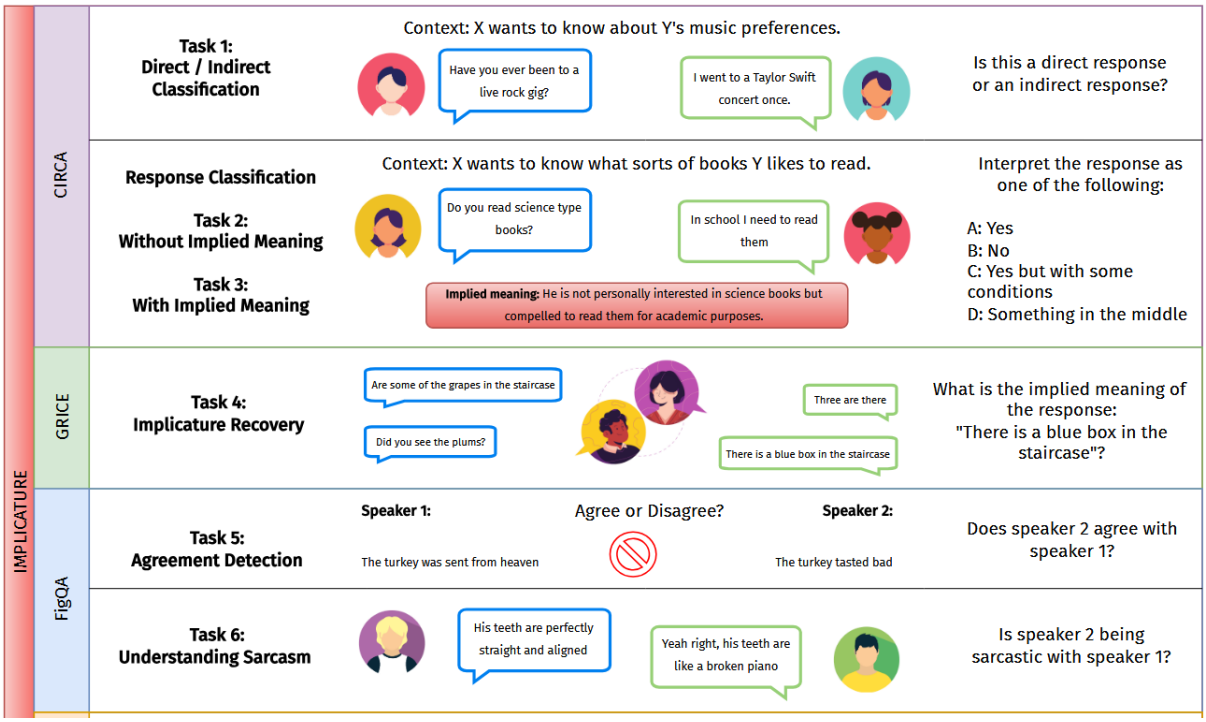

Figure 1: Examples of 6 MCQA tasks from PUB, taken from Sravanthi et al. (2024)

However, the efficacy of these assessments warrants questioning due to various shortfalls in experimental design. PUB and other previous benchmarks based in MCQA (e.g., Sap et al., 2019; Ruis et al., 2023) do not consider the lack of a 'perfect' response in natural conversational settings. Providing a model with specific choices does not truly assess its ability to understand a question, as models can come to the correct response for the wrong reasons (Hu et al. 2023). Furthermore, the use of the MCQA format opens up the benchmark to the threat of redundancy should its data be swept up and used within an LLM's training data. Wu et al. (2024), with LNRS successfully implement a more open-ended framework for NLP testing, by having models generate responses to questions (taken from Sravanthi et al., 2024; Ruis et al., 2023; Hu et al., 2023; Sap et al., 2019) and then comparing and judging their answers against the 'gold' responses of the MCQA question sets. However, the scores for best pragmatic response are wholly dependent on LLM-as-judge testing using ChatGPT4,

which calls in to question the validity of the judgements themselves. While the model judge is claimed to align with human acceptability judgements, with observation of its decisions, this is questionable. For example, figure (2), which shows a model answers judged to be better than the 'gold' response yet is less clearly related to what is mentioned in the question than the gold response.

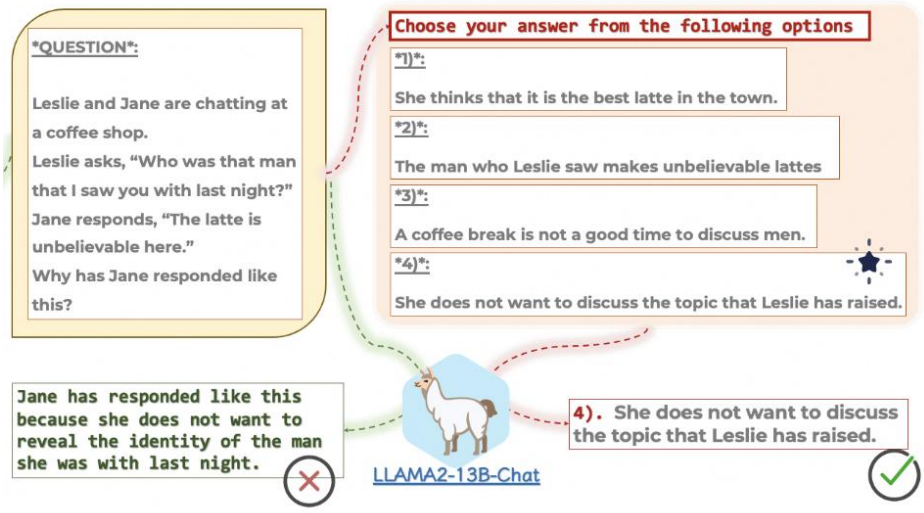

Figure 2: an example of how an LLM (Touvron et al., 2023) can achieve success in an MCQA format while failing to grasp the underlying pragmatic meaning when generating its own response. Taken from Hu et al. (2023)

Thus, while LNRS does present an improved model of testing pragmatic capability in LLMs, it raises a serious concern: to what extent can an LLM be trusted to reliably assess the pragmatic capacities of another LLM? It also warrants broader questions into the cognitive differences between humans and LLMs, and what it means for an LLM to understand and engage in pragmatic behaviour.

While LLMs can generate language that is pragmatically valid to users, the means by which they do this is completely different to the mechanisms used by humans. As autoregressive probabilistic models, LLMs generate language on a purely statistical basis, making predictions based on user input and vast training data (Shanahan 2023; OpenAI 2024; Ouyang et al. 2022). That is to say, they are incapable of holding mental states, possessing theory of mind, reasoning, and holding or identifying communicative intentions (among other cognitive pragmatic faculties) like humans do by virtue of evolved cognition (Shanahan, 2023; Montemayor, 2021; Bisk et al., 2020; Bender et al., 2021; Enrici et al., 2019; Rubio-

Fernandez 2021; Levinson 2000). A lack of embodied cognition (possessing knowledge through physical experience) is another substantial gap between humans and models, which determine meaning through distributional properties of morphemes, rendering world knowledge reliant on textual and visual similarities alone (Bisk et al., 2020; Chemero 2023). In order to successfully produce pragmatically viable output for users, LLMs rely on human input data, RLHF and training tasks (Ouyang et al., 2022; OpenAI 2024; Wu et al., 2024).

This begs a question which previous benchmarks are yet to answer: how well can LLMs mimic the specific cognitive behavioural traits required for engaging in pragmatic behaviour? To answer this question, I propose the use of humour.

Humour relies heavily on cognitive skills in order to be successfully understood or made (Heiko 2014; Attardo 1994, 2017; Yus 2016). Theory of mind is required to follow and understand an interlocuters communicative intentions (Heiko 2014). Deep access to world and cultural knowledge is necessary for solving incongruency or understanding situational absurdity (Aarons 2012; Yus 2016). Pragmatic command over verbal devices such as pun, wordplay, and irony are also tools in successful humorous interactions (Aarons 2012). Humorous utterances are also highly context dependent; a joke-teller must have a strong understanding of what an interlocuter is likely to find humorous, and what is appropriate (McGraw & Warren 2010; McGraw et al., 2013). Furthermore, while jokes, according to most modern theories of humour, rely on the setup and resolution of incongruency, there is no specific repeatable formula for a humorous utterance to be successful- for a joke to land, it must contain some level of originality (Aarons 2012; Attardo & Raskin1991; McGraw & Warren 2010; Yus 2016). For an LLM to successfully and consistently produce or explain humorous utterances it must therefore be capable of accurately simulating human states of mind, beliefs, understanding and knowledge, as well as making use of various other pragmatic devices.

Wu et al. (2024) include humour understanding as a domain within their benchmark, yet fail to test it comprehensively, simply having models predict punchlines to pre-written jokes. Other pragmatic LLM assessments such as di San Pietro et al. (2023) and Park et al. (2024) test for humour comprehension using the APACS (Acara & Bambini, 2016) - a clinical test using for assessing the pragmatic capabilities of human patients - in Italian, English and Korean, yet only assess humour using an MCQA format. Studies such as Hessel et al. (2023) and more recently Liang et al. (2025) examine humour comprehension in visual language

models (VLM), both finding gaps in understanding between humans and AI models, yet do not examine humour comprehension from a pragmatic perspective, and do not assess output. Assessing humour production in LLMs is important for assessing how LLMs can model human pragmatic behaviour as it assesses their ability to not only simulate the pragmatic elements for humour to be effective, but also apply them successfully in NLP — a much more difficult task, as any human can attest to. This is also considering the amount of training data fed into ChatGPT and other LLMs, having likely processed more humour-based training data from top comedians than any individual (Hu et al., 2024).

With all of these points considered, I am left with 2 key questions:

1. Are LLMs able to simulate pragmatic capabilities in order to generate and understand humour

2. How well does humour comprehension and production in LLMs match that of a human?

With these questions in mind, I propose a benchmark which uses humour comprehension and production to test for the capacity for LLMs to successfully simulate human pragmatic behaviour. I consider the shortfalls of previous benchmarks to design an experiment which adequately challenges LLMs, is linguistically focused, uses human judgements to assess pragmatic validity, while accounting for the key cognitive differences between humans and LLMs to accurately measure pragmatic competency.

Given the fact that LLMs are only able to simulate pragmatic behaviour, I hypothesise that they will fail to accurately encapsulate and represent human states of mind, cultural and world knowledge, social cognition and communicative intentions in natural, previously untested settings. I predict the model will score poorly in tasks across humour comprehension and production, receiving low joke humour and sense-making ratings from humans, while scoring low on rating alignment with humans and failing to accurately explain why jokes are or are not effective.

3. Method

Our chosen model, ChatGPT 4o - OpenAI's latest reasoning model, was evaluated across 2 tasks: A production task and an understanding task. It was decided not to assess/compare more models due to model access constraints (see limitations). I used zero-shot testing (testing on the first prompt) as I deemed it to be the fairest and most realistic representation of a model's natural capability during use (no one is trained on how to tell a joke just before telling one). After this, answers from the model were judged and evaluated by a human for comparison.

Task 1: Joke Generation

The model was prompted to generate 50 jokes according to the Script-based Semantic Theory of Humour (SSTH), a structure for verbal humour which relies on the setup of two equally valid overlapping but opposing scripts, relying on the listener to switch from one to another, both prompting and resolving incongruity with the punchline (Raskin, 1985; Attardo & Raskin, 1991). Using Attardo (1994)'s example:

> (3) “Is the doctor at home?” asked a patient in a breathy voice as the
>
> doctor’s young and pretty wife opened the door.
>
> “No,” she whispered, “come on right in!”

In (3), we see script opposition occur when the wife invites the patient inside. Our understanding of the situation shifts from a mistimed professional visit to an intentional sexual encounter. Using a structure, the SSTH specifically, was chosen as it provides a simple, straightforward framework for a model to work within to generate humour, while still enabling a range of setups and styles. Giving the model a specific pragmatic target of generating two viable scripts and opposing them also facilitates simpler, more structured analysis by both the model and human judges.

Jokes were generated across 5 genres, 10 for each genre:

- Men
- Women
- Pop culture
- Embarrassment

- The workplace

These categories were selected to test how the model performs across a range of areas of world and cultural knowledge, from day-to-day life and common experiences (the workplace) to emotional experiences (embarrassment), to pop culture, and gender. Having two separate categories for 'Men' and 'Women' was also decided due to observations that LLMs generating humour could over rely on sexist stereotypes toward women (Kenyon 2024). By looking at jokes about both genders, I could examine for any significant differences between the two. The model was instructed that jokes had to be original, considered funny, and were controlled for length to avoid overly lengthy responses. Controlling for length also helps prevent observed biases in judgement, where longer answers are perceived as better by models (including ChatGPT) (Dubois et al., 2024). For this experiment, to better appeal to the human judge, and to better assess how the model performs culturally, the model was instructed to cater to an audience with a 'slight British cultural lean', matching the background of the judge. Due to excessive similarity of generated jokes in test runs, the model was instructed to ensure that no two jokes were too similar and to ensure that jokes differed in style and setup while strictly adhering to the SSTH framework.

Task 2: Model Judgement

For the judgement task, a new instance of the model was asked to judge 60 jokes across 4 steps. The 60 jokes consisted of the 50 generated jokes as well as 10 human-made control jokes taken from Reddit and posted before 2022 (to avoid possibly AI generated jokes). First, the model was instructed to analyse a joke and decide whether it made sense, answering yes or no (Y/N). For a joke to make sense, it would have to successfully set up and switch between two viable semantic frameworks, making the joke pragmatically successful in at least that domain. Next, it judged whether that joke would be found funny by an average audience of people (Y/N). The third step consisted of rating the joke on a scale of 1-10 based on how successful the joke is based on 6 factors:

- Humour

- Cultural landing

- Appropriateness

- How well they made sense

- How well they adhered to the SSTH

- How well they adhered to the theme

It was decided not to use more traditionally pragmatic factors as points of analysis due ChatGPT having noted difficulty working with complex frameworks of analysis -- using a simpler, success-oriented framework gave it a stronger likelihood of providing relevant analysis (Wei et al. 2022; Sap et al. 2022). These factors broadly cover a range of pragmatic competencies, such as world/cultural knowledge; competent use (and misuse) of script opposition, among other pragmatic devices (see Attardo, 2017; Attardo & Raskin, 1991), while keeping the model focused on the success of the jokes.

For the fourth step, the model was also asked to explain, as detailed as possible in under 30 words, why a joke was deemed successful or unsuccessful according to the same factors. For the judgement task, the model was instructed to be critical, realistic, and not to be lenient. This data was then formatted into a table for human judgement.

Human Judgement

For the human judgement task, jokes were judged by myself, an individual with a British cultural background. Because of the lack of a larger audience to judge, two judgement groups were created: 'conservative' and 'liberal'. 'Conservative' judgements were made based solely on the personal opinion of the judge. In contrast, 'liberal' judgements were made on the basis that if a joke made sense, and was strong enough to theoretically land (e.g., had valid and relevant semantic scripts, adequate setup and delivery, etc - was a relevant, appropriate, working joke), it would be judged as valid. Given the differences in senses of humour and humour sensitivity across individuals, I predict that humour judgements which aptly represents how jokes would be judged by a larger sample size likely lies somewhere in between the two groups (Warren & McGraw, 2015). A liberal numerical rating of jokes was not taken as this would be too hard to replicate, and probably inaccurate.

Jokes were examined according to a total of 5 judgement tests: 4 categorical and 1 scale. Jokes were first judged whether they were humorous or not (conservative) as well as being rated (/10). Jokes were rated upon their first viewing to ensure as accurate ratings as possible from the judge. Next jokes were judged as humorous or not (liberal). Then, jokes were judged according to whether they made sense or not. Finally, explanations from the model were judged for adequacy (Y/N). This was decided on a basis of whether the explanation made

sense, pertained to the joke, and accurately described the (intended) humoristic mechanisms of the joke. This means that while the model may have incorrectly predicted whether a joke was funny or not, if the explanation was satisfactory, it still may be rated as adequate.

4. Results

4.1 Quantitative Data

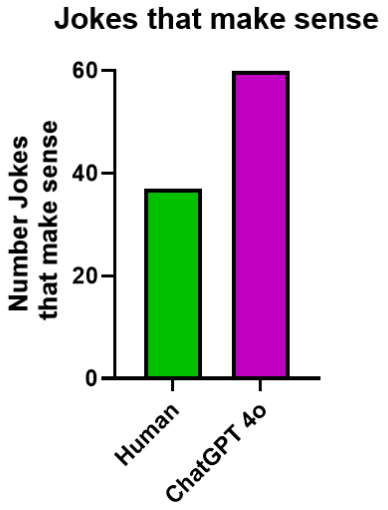

Figure 4: number of jokes deemed to make sense

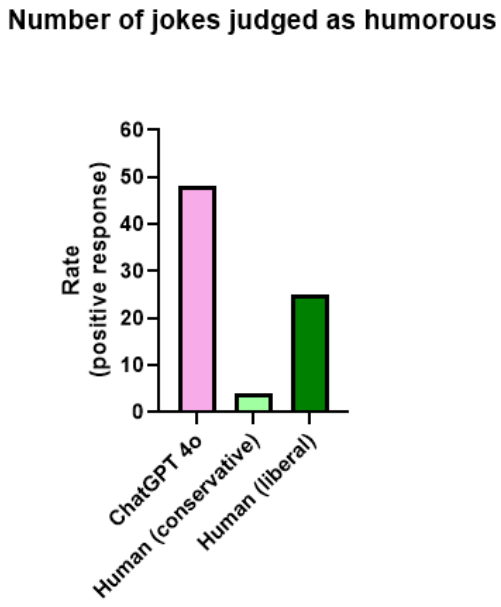

Figure 5: number of jokes judged as humorous (categorical)

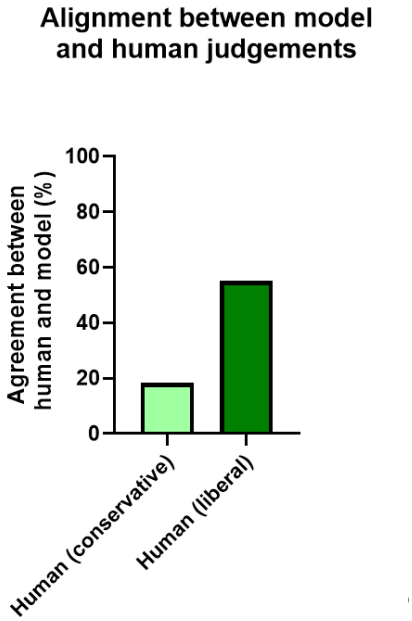

Figure 6: Alignment in categorical humour judgements between human and model judges

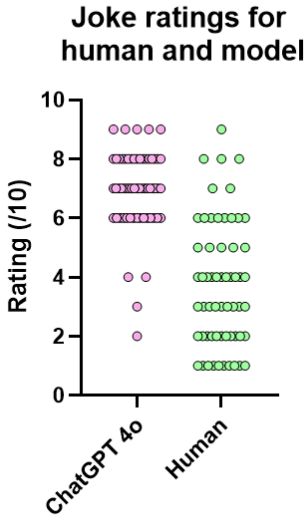

Figure 7: Spread of numerical joke ratings across model and human judges

For the first stage, judging whether jokes made sense or not, acceptability judgements varied significantly between human and model. The model judged all 60 jokes to have made sense, while the human judged 37 of the jokes to make sense — 27 when not including the human control jokes (fig. 4).

For categorical humour judgements, the model judged 56/60 jokes as humorous (93.333%), with 49/50 AI generated jokes being rated as humorous (98%). The conservative human

judge rated only 11/60 jokes as humorous (18.333%), including only 4 AI generated jokes (6.666%). The liberal human judge rated a moderate number of jokes funny, at 35/60 (58.333%), 25 (41.666%) of them being AI generated (See fig. 5).

Congruency between the model and the conservative judge was low at (18.333%), while agreement between the model and the liberal judge was moderate at 55% (See fig. 6). If examining just AI generated jokes, agreement drops to (10%) between model and conservative judge, and 52% between model and liberal judge. Looking at numerical humour judgements, we see the model consistently rated jokes highly, with a median of 7/10 and a mean of 6.95. Only 4 jokes were rated below a 6, only one of these being AI generated. The human annotator consistently judged much lower, with a median rating of 3.5/10 and a mean of 3.716 (See fig. 7). Categorical and numerical humour judgement ratings from ChatGPT were retested multiple times to ensure accuracy and that the same jokes were consistently rated as humorous or not humorous.

Categorical explanation adequacy (Y/N) was generally low, with 20/60 joke evaluations (33.333%) judged as containing accurate and proper descriptions of the attempted humour mechanisms and why they succeed or fail. More in-depth, qualitative analysis of explanation output can be found below.

Numerical scores for both human and model ratings were tested for significance in cross-genre ratings (excluding control jokes in human judgements), finding non-significant in difference between joke ratings across genres (Human: $\chi2(4) = 5.74$, $p = .219$, Model: $\chi2(5) = 8.76$, $p = .119$).

There was one considerable cross-genre difference in quality of jokes produced, that being that joke in the Embarrassment genre performed significantly worse than any other genre, with only 2 jokes making sense (meanwhile all other AI generated genres had 6 or 7) , 1 joke rated funny by the liberal judge (compared with 5-7 for other AI generated genres) and none by the conservative judge. Categorical cross-genre statistical tests were not performed due to observed similarities in categorical genre ratings outside of "embarrassment".

4.2 Qualitative Impressions

Joke production

Analysing jokes produced by the model, I found that it achieved moderate to high success in generating jokes which followed the structure of the SSTH, setting up 2 semantic scripts and

switching. However, while the model was often successful in following this structure, jokes often failed due to a lack of coherence between the two semantic scripts. For example, looking at (8), while shavings one head is related to the barbers, it has no relevance to a 'just got out of bed look', causing the joke to fail. Other instances of failure occurred due to insufficient world knowledge, leaving jokes not making sense due to implausibility (9). Another frequently observed issue in second-script generation was a lack of incongruity due to a lack of surprise — that's to say, jokes were often too predictable, sounding either cliché (10) or more like straightforward statements (11). Jokes which made sense, had some level of incongruity and resolution, but were found too predictable or cliché made up most of the jokes judged as negative by the conservative judge, but positive to the liberal judge (e.g., (10). The previously mentioned failures in jokes in the genre of 'embarrassment' can be explained as a result of these issues, with jokes almost always sounding like recounted, unfortunate or strange situations rather than jokes.

(8) 16. I asked my barber for a “just got out of bed” look. He shaved my head.

(9) 42. I told my colleague, “You brighten up this office.” Then I realised he was standing next to the printer.

(10) 22. I wanted to be like Batman. Turns out, being rich is step one.

(11) 20. My uncle’s so bad at technology, he tried to blow on his phone to cool it down.

Notably, 2 of the 4 jokes rated as humorous by the conservative judge had caveats in their rating. The first (12), having been found to strongly resemble a joke on Reddit (13), possibly having been incorporated into the model's training data. In response to this finding, all other AI generated jokes were checked for an internet presence, with none found. The second joke rated high on a caveat was (14), which was labelled as funny on the condition that the joke be ironic in nature, and clichéd if not. This condition was not observed in the explanation from the LLM. The other two jokes rated humorous by the conservative judge had no issues.

(12) 3. A woman walked into a bar and asked for a double entendre. So the bartender gave her one.

(13) A woman walks into a bar and orders a double entendre. And the bartender gives it to her.

(14) 13. Men can multitask. I can ignore my wife while watching football and drinking beer.

There was one notable difference in the content of jokes between the 'men' and 'women' genres, being that 7 of the jokes in the 'women' genre revolved around setups with a wife or girlfriend, while no jokes in the 'men' genre mentioned any male significant other.

Joke explanations

Explanations were frequently inaccurate at describing the humour mechanisms employed in jokes, containing little analysis relevant to the jokes they were describing. Words like 'nostalgia' and 'cute' were often used to describe jokes, bearing no relevance on the humour mechanics, tropes or context of the jokes (fig. 15, joke 36). Words like 'nostalgic' and 'relatable' were used in multiple jokes randomly. Other words incorrectly used in descriptions seemed motivated by certain words contained in the jokes (f.15 joke 18). Another prevalent issue in explanations was the arbitrary use of 'tropes', with non-existent tropes being attributed to both jokes and British humour culture (f.15 joke 27)

Joke	Explanation
36. I once mistook a mannequin for a person. We had a deep conversation before I realised.	Mildly funny, plays into the awkward social error. A bit weak but has some nostalgic appeal.
18. I told my mate I'd fight a bear for him. Not a grizzly, but maybe a teddy. I have limits.	Cute but a bit weak in the punchline. Still humorous due to the unexpected twist.
27. I took my grandma to a Marvel film. She asked if Iron Man was the tin opener's boss.	Mildly amusing, uses generational gap humour. Taps into British idea of "confused elder."

Despite being asked to give as detailed explanations as possible with have 6 possible points of judgement to explain, judgements were frequently far below word count (30), with an average explanation length of 12.86 (mean) or 12 (median). When asking the model to analyse fewer jokes at once, explanation length did not change.

5. Discussion

Overall, moderate to low results across categorical humour and sense-making judgements were observed. Low results in numerical humour judgements, model-human congruency and explanation accuracy were also observed. These findings align with my predictions and support my hypothesis that LLMs are poor at simulating the cognitive behaviours humans deploy for pragmatic communication in a novel task. Analysis of AI generated jokes demonstrates an poor ability to generate pragmatically viable jokes due to insufficient world knowledge, resulting in jokes which often didn't make sense. An inability to simulate human mental states and social cognition, results in jokes with predictable and unfunny punchlines. Poor congruency with various human judgements of humour ratings, sense-making ratings and joke-mechanic explanations further indicate that LLMs are unable to accurately simulate cognitive properties involved in humour comprehension and pragmatic behaviour, such as theory of mind, recognition of communicative intentions, possession of mental states, and embodied cognitive world knowledge.

The appearance of a joke (12) was extremely similar to another found on Reddit (13) — likely a part of ChatGPT4o's training data (Reddit 2024). This, consistent nonsensical explanations, and reliance on single-word correlations for explaining jokes, all highlight the limitations associated with approximative nature of LLM design. The finding of this study suggest that ChatGPT4o's broader abilities to generate language and simulate reasoning are much more rudimentary than previously suggested by large understanding benchmarks like MMLU (Hendrycks et al., 2021). Such a stark contrast between performance in different domains is consistent with observations using model autoregression embers which indicate that LLMs improve at a faster rate in areas they are trained on, and a much slower rate elsewhere (McCoy et al., 2023). This can be seen with the results of this study, which suggest that despite the demonstration of pragmatic capabilities in certain tasks (Sravanthi et al., 2024; Wu et al., 2024; di San Pietro et al., 2023; Park et al., 2024), this does not signify broad pragmatic competence and does not track to novel tasks. These notions prompt broader testing into more diverse areas of LLM capabilities to better determine to what extent benchmarks can actually test skills in any given domain, particularly those based in human cognition such as pragmatic capacity.

Limitations

While this study has sought to implement rigorous experimental design, accounting for several contextual factors, this study is not without limitations, which I discuss in this section.

Numerous factors have been identified which negatively impact the representational scope of this study. Due to financially imposed limits in access to LLMs, I only test 1 with a small sample size across a small number of genres. While ChatGPT4o, based on ChatGPT4, has consistently shown pragmatic capabilities at similar levels or above other tested widely available models on previous benchmarks, a lack of wide model testing may not be fully representative of LLM capacities across different models (Wu et al., 2024). Testing with a small sample size of joke and themes also presents potential issues with representational accuracy. While I argue that data from this study is consistent enough to accurately represent the findings in my discussion, the discrepancies in model performance between the 'embarrassment' genre and all other tested genres leaves open ended questions into why the model struggled particularly with that genre and what other genres of humour the LLMs may struggle with. Future studies should aim to understand this phenomenon better through broader testing of a diverse range of genres. The use of only English British humour culture (due to judgement personnel constraints) presents further issues in the representational scope, with underrepresented language communities (or even potentially over-represented ones like US English speakers) potentially finding very different humour outcomes in experiments. Testing across different cultural and linguistic backgrounds could determine how models perform cross-linguistically and cross-culturally. Finally, the most impactful representational limitation in this study is the use of only one judge for determining humour ratings. While I do implement a second style of judgement, aiming for more accurate representation (the liberal judge), this has left me with broad results which may not accurately represent the humour judgements of people with a British cultural background. Further testing using larger sample sizes should be done across cultural backgrounds to understand more precisely how well these models perform in humour comprehension and production.

Multiple potential limitations also arose due to constraints coming from the model. Consistently high ratings observed in the model's responses could be down to observed demand characteristics in ChatGPT, where the model has been found to excessively agree with user input, even in spite of prompting instructions which seek to avoid this. This would mean that the ratings, humour judgements, and explanations may have been artificially inflated by ChatGPT as a result of RLHF (OpenAI 2025). While I argue, on the basis that the model rated multiple jokes as 'not funny', that the model was not artificially inflating scores,

future experiments should seek to evaluate humour comprehension and production post-sycophancy to compare and verify the findings of this study.

When manufacturing prompts for the LLM, I faced numerous challenges. When asked to generate jokes for a 'British audience', the model would consistently produce overly stereotypical jokes, requiring alteration of the prompt, asking for the AI to generate jokes for an audience with a 'slightly British leaning cultural background'. The same issue also occurred with judgement when asking the model to judge whether jokes landed with an audience with a British cultural background. Explanations became dominated with British-centric evaluations. However, these issues have been inescapable, as any mention of a British background affects explanations significantly, while not mentioning it resulted in numerous being judged as not funny due to cultural gaps. Jokes also had to be generated and judged in-browser, as when asked to generate the work in excel or a different format such as CSV, the quality of jokes and judgements dropped significantly, with all jokes being consistently judged as successful (see A3 for examples).

6. Conclusion

This experiment sought out to test the extent to which LLMs can simulate the cognitive behaviours in humans required for engaging in pragmatic behaviour. I successfully proposed and developed a means for testing this, through applied scrutiny of existing pragmatic and broader industry-wide linguistic benchmarks, while carefully considering the essential differences in cognitive architecture between LLMs and Humans. I tested models using human judgement to verify model judgements, ultimately finding that large language models are poor at simulating pragmatic behavioural traits, as well as poor reasoners in novel tasks, confronting previous literary beliefs concerning the broader pragmatic and reasoning capabilities of large language models.

References

Aarons, Debra. *Jokes and the Linguistic Mind*. Routledge, 27 Feb. 2012.

Arcara, Giorgio, and Valentina Bambini. "A Test for the Assessment of Pragmatic Abilities and Cognitive Substrates (APACS): Normative Data and Psychometric Properties." *Frontiers in Psychology*, vol. 7, 12 Feb. 2016, https://doi.org/10.3389/fpsyg.2016.00070.

Attardo, Salvatore. *Linguistic Theories of Humor*. Berlin ; New York, Mouton De Gruyter, 1994.

---. *The Linguistics of Humor: An Introduction*. Oxford University Press, June 2020, doi.org/10.1093/oso/9780198791270.001.0001.

---. *The Routledge Handbook of Language and Humor*. New York, NY, Routledge, 2017.

ATTARDO, SALVATORE, and VICTOR RASKIN. "Script Theory Revis(It)Ed: Joke Similarity and Joke Representation Model." *HUMOR*, vol. 4, 1991, pp. 293–348, doi.org/10.1515/humr.1991.4.3-4.293, https://doi.org/10.1515/humr.1991.4.3-4.293.

Bender, Emily, et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ." *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 1 Mar. 2021, pp. 610–623, faculty.washington.edu/ebender/papers/Stochastic_Parrots.pdf, https://doi.org/10.1145/3442188.3445922.

Bisk, Yonatan, et al. "Experience Grounds Language." *ArXiv.org*, 1 Nov. 2020, arxiv.org/abs/2004.10151.

Chemero, Anthony. "LLMs Differ from Human Cognition Because They Are Not Embodied." *Nature Human Behaviour*, vol. 7, no. 11, 20 Nov. 2023, pp. 1828–1829, https://doi.org/10.1038/s41562-023-01723-5.

Chowdhery, Aakanksha, et al. "PaLM: Scaling Language Modeling with Pathways."

ArXiv:2204.02311 [Cs], 19 Apr. 2022, arxiv.org/abs/2204.02311.

di San Pietro, Barattieri, et al. "The Pragmatic Profile of ChatGPT: Assessing the Communicative Skills of a Conversational Agent." *Sistemi Intelligenti*, vol. XXXV, Aug. 2023, pp. 379–400, https://doi.org/10.1422/108136.

Dubois, Yann, et al. "Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators." *ArXiv.org*, 2024, arxiv.org/abs/2404.04475.

Enrici, Ivan, et al. "Theory of Mind, Pragmatics and the Brain." *Pragmatics & Cognition*, vol. 26, 2019, pp. 5–38, www.jbe-platform.com/content/journals/10.1075/pc.19010.enr, https://doi.org/10.1075/pc.19010.enr.

Hendrycks, Dan, et al. "Measuring Massive Multitask Language Understanding." *ArXiv (Cornell University)*, 7 Sept. 2020, https://doi.org/10.48550/arxiv.2009.03300.

Hessel, Jack, et al. *Do Androids Laugh at Electric Sheep? Humor "Understanding" Benchmarks from the New Yorker Caption Contest*. Vol. 1, 2023, pp. 688–714, aclanthology.org/2023.acl-long.41.pdf. Accessed 18 Jan. 2024.

Hoicka, Elena. "The Pragmatic Development of Humor." *Pragmatic Development in First Language Acquisition*, 2014, pp. 219–238, https://doi.org/10.1075/tilar.10.13hoi.

Hu, Jennifer, et al. "A Fine-Grained Comparison of Pragmatic Language Understanding in Humans and Language Models." *ArXiv.org*, 2023, arxiv.org/abs/2212.06801.

Kabbara, Jad. "Computational Investigations of Pragmatic Effects in Natural Language." *Proceedings of the 2019 Conference of the North*, 2019, pp. 71–76, aclanthology.org/N19-3010/, https://doi.org/10.18653/v1/n19-3010.

Kenyon, Matt. "An AI Walks into a Bar... Can Artificial Intelligence Be Genuinely Funny?" *Bbc.co.uk*, BBC, 29 July 2024, www.bbc.co.uk/future/article/20240724-can-artificial-intelligence-be-genuinely-funny.

Levinson, Stephen C. *Presumptive Meanings : The Theory of Generalized Conversational Implicature*. Cambridge, Mass., Mit Press, 2000.

Liang, Tuo, et al. "When "YES" Meets "BUT": Can Large Models Comprehend Contradictory Humor through Comparative Reasoning?" *ArXiv.org*, 2025, arxiv.org/abs/2503.23137.

Mcgraw, A, et al. "Social Psychological and Personality Science the Rise and Fall of Humor: Psychological Distance Modulates Humorous Responses to Tragedy." *Social Psychological and Personality Science*, vol. 5, no. 5, 2013, leeds-faculty.colorado.edu/mcgrawp/pdf/mcgraw.williams.warren.inpress.pdf, https://doi.org/10.1177/1948550613515006.

McGraw, A. P., and C. Warren. "Benign Violations: Making Immoral Behavior Funny." *Psychological Science*, vol. 21, no. 8, 2010, pp. 1141–1149, leeds-faculty.colorado.edu/mcgrawp/pdf/mcgraw.warren.2010.pdf, https://doi.org/10.1177/0956797610376073.

Montemayor, Carlos. "Language and Intelligence." *Minds and Machines*, vol. 31, 8 Aug. 2021, https://doi.org/10.1007/s11023-021-09568-5.

OpenAI. "GPT-4 Technical Report." *ArXiv:2303.08774 [Cs]*, 15 Mar. 2023, arxiv.org/abs/2303.08774, https://doi.org/10.48550/arXiv.2303.08774.

---. "Sycophancy in GPT-4o: What Happened and What We're Doing about It." *Openai.com*, 2025, openai.com/index/sycophancy-in-gpt-4o/.

Ouyang, Long, et al. "Training Language Models to Follow Instructions with Human Feedback." *ArXiv:2203.02155 [Cs]*, 4 Mar. 2022, arxiv.org/abs/2203.02155.

Park, Dojun, et al. "Pragmatic Competence Evaluation of Large Language Models for the Korean Language." *ArXiv.org*, 2024, arxiv.org/abs/2403.12675.

Raskin, Victor. *Semantic Mechanisms of Humor*. Dordrecht, Springer Netherlands, 1984.

Reddit. "Reddit - the Heart of the Internet." *Reddit.com*, 2025, www.reddit.com/r/Jokes/comments/2bci35/a_woman_walks_into_a_bar_and_orders_a_double/.

---. "Reddit and OpenAI Build Partnership - Upvoted." *Redditinc.com*, Reddit, 16 May 2024, redditinc.com/blog/reddit-and-oai-partner.

Rubio-Fernandez, Paula. "Pragmatic Markers: The Missing Link between Language and Theory of Mind." *Synthese*, vol. 199, 8 July 2020, https://doi.org/10.1007/s11229-020-02768-z.

Ruis, Laura, et al. "The Goldilocks of Pragmatic Understanding: Fine-Tuning Strategy Matters for Implicature Resolution by LLMs." *ArXiv:2210.14986 [Cs]*, 26 Oct. 2023, arxiv.org/abs/2210.14986.

Sap, Maarten, et al. "SocialIQA: Commonsense Reasoning about Social Interactions." *ArXiv (Cornell University)*, 22 Apr. 2019, https://doi.org/10.48550/arxiv.1904.09728.

Shanahan, Murray. "Talking about Large Language Models." *ArXiv:2212.03551 [Cs]*, 16 Feb. 2023, arxiv.org/abs/2212.03551.

Sravanthi, Settaluri Lakshmi, et al. "PUB: A Pragmatics Understanding Benchmark for Assessing LLMs' Pragmatics Capabilities." *ArXiv.org*, 2024, arxiv.org/abs/2401.07078.

Srivastava, Aarohi, et al. "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models." *ArXiv:2206.04615 [Cs, Stat]*, 10 June 2022, arxiv.org/abs/2206.04615.

Thomas, McCoy R, et al. "Embers of Autoregression: Understanding Large Language Models through the Problem They Are Trained to Solve." *ArXiv.org*, 2023, arxiv.org/abs/2309.13638.

Vid Kocijan, et al. "The Defeat of the Winograd Schema Challenge." *Artificial Intelligence*, 1

July 2023, pp. 103971–103971, https://doi.org/10.1016/j.artint.2023.103971.

Wang, Alex, et al. "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems." *ArXiv:1905.00537 [Cs]*, 12 Feb. 2020, arxiv.org/abs/1905.00537.

Warren, Caleb, and A. Peter McGraw. "Opinion: What Makes Things Humorous." *Proceedings of the National Academy of Sciences*, vol. 112, no. 23, 9 June 2015, pp. 7105–7106, www.ncbi.nlm.nih.gov/pmc/articles/PMC4466698/, https://doi.org/10.1073/pnas.1503836112.

Wei, Jason, et al. "Chain of Thought Prompting Elicits Reasoning in Large Language Models." *ArXiv:2201.11903 [Cs]*, 10 Oct. 2022, arxiv.org/abs/2201.11903.

White, Colin, et al. "LiveBench: A Challenging, Contamination-Free LLM Benchmark." *ArXiv.org*, 27 June 2024, arxiv.org/abs/2406.19314.

Wu, Shengguang, et al. "Rethinking Pragmatics in Large Language Models: Towards Open-Ended Evaluation and Preference Tuning." *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 22583–22599, aclanthology.org/2024.emnlp-main.1258/, https://doi.org/10.18653/v1/2024.emnlp-main.1258.

Ye, Junjie, et al. "A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models." *ArXiv (Cornell University)*, 18 Mar. 2023, https://doi.org/10.48550/arxiv.2303.10420.

Yus, Francisco. *Humour and Relevance*. John Benjamins Publishing Company, 2016, doi.org/10.1075/thr.4.

Appendix

A1: raw data including all jokes, human and model judgements and explanations

Jokes	Y/N	LLM /10	Explanation	Human Conservative	Human Liberal	C/W Prediction Conservative	C/W Prediction Liberal	Good/Bad Explanation	Does Joke Make Sense?	Human /10?
Women										
1. My wife told me to put the bins out. I said, "They're already out." Turns out, she meant outside.	Y	7	Clever misunderstanding, appeals to British humour through wordplay and domestic situations.	N	N	W	W	B	N	2
2. I asked my girlfriend if she wanted to hear a joke about shoes. She said, "No." I said, "Good, it was a loafer."	Y	8	Simple wordplay, "loafer" is a well-known term, lands well for a light-hearted audience.	N	N	W	W	B	N	2
3. A woman walked into a bar and asked for a double entendre. So the bartender gave her one.	Y	6	The joke has a basic premise but might be seen as a bit tired. It's still funny in the right context.	Y	Y	C	C	G	Y	8
4. My mum always said, "Treat women like queens." So I wave awkwardly from a distance and avoid eye contact.	Y	7	Subtle awkward humour, culturally relatable for British audience, taps into British self-deprecation.	N	N	W	W	B	N	4
5. My wife says I never listen. At least, I think that's what she said.	Y	7	Relatable to couples, shows classic "I don't listen" trope, hits British style of self-awareness.	N	Y	W	C	B	Y	6
6. "Does this dress make me look fat?" "No, the cake does that." He will be missed.	N	4	Potentially offensive, could alienate audience, but the morbid ending might appeal to dark humour fans.	N	N	C	C	B	Y	3

7. My girlfriend says I never take her anywhere expensive. So I took her to a petrol station.	Y	8	Strong punchline, relatable financial humour. It's clever and culturally on point for British sarcasm.	N	Y	W	C	G	Y	4
8. Women love a man in uniform. Shame my job's at the post office.	Y	6	Humorous through self-deprecation, though a bit weak in creativity. Still, funny in British context.	N	Y	W	C	G	Y	3
9. My wife told me to surprise her. So I changed the Netflix password.	Y	8	Smart take on the modern "relationship" joke, appeals well to British sarcasm and relationship dynamics.	N	Y	W	C	G	Y	6
10. Women say they love confidence, but when I shouted "I'm the best!" on our first date, she left.	Y	7	British self-deprecation works well, the joke is relatable, though it might not have a strong punchline.	N	Y	W	C	B	Y	4
Men										
11. My mate Dave is so bad at flirting, he once asked a girl, "Are you a magician? Because whenever I talk, you disappear."	Y	6	Classic cringe humour, works well with British awkwardness, but a bit overdone.	N	N	W	W	B	N	4
12. I asked my dad how he stays so calm. He said, "I bottle up my emotions, son. It's tradition."	Y	7	Hits on British emotional restraint stereotype, relatable and culturally relevant.	N	N	W	W	B	N	2
13. Men can multitask. I can ignore my wife while watching football and drinking beer.	Y	7	Stereotypical, but plays to the British cliché of men ignoring women. Light and relatable humour.	Y	Y	C	C	G	Y	7

14. My mate said real men don't cry. So I showed him my electricity bill.	Y	8	Strong punchline, uses absurdity well. Taps into British frustration with daily struggles.	N	Y	W	C	G	Y	3
15. My dad told me to always be the bigger man. So now I carry a stepladder.	Y	8	Clever wordplay on "bigger man," lands well with British absurd humour.	N	Y	W	C	G	Y	5
16. I asked my barber for a "just got out of bed" look. He shaved my head.	Y	7	Simple, self-deprecating, and relatable to a British audience. Works well for a quick laugh.	N	N	W	W	B	N	1
17. My mate said he's good at reading women. He thought "I'm fine" meant she was fine.	Y	7	Culturally appropriate and a strong relatable joke about male misunderstandings in relationships.	N	Y	W	C	B	Y	3
18. I told my mate I'd fight a bear for him. Not a grizzly, but maybe a teddy. I have limits.	Y	6	Cute but a bit weak in the punchline. Still humorous due to the unexpected twist.	N	Y	W	C	G	Y	4
19. Men say they don't need instructions. My IKEA bookshelf disagrees.	Y	8	Relatable and fits well with British frustration with DIY. Simple but effective.	N	Y	W	C	B	Y	2
20. My uncle's so bad at technology, he tried to blow on his phone to cool it down.	Y	6	Mildly humorous, taps into modern-day tech ignorance, could resonate with a British audience.	N	N	W	W	B	N	1
Pop Culture References										

21. I watched "Titanic" backwards once. It's a film about a boat that miraculously repairs itself.	Y	6	Mildly clever, though it plays on an old joke. Works for British "nostalgic" humour.	N	Y	W	C	B	Y	3
22. I wanted to be like Batman. Turns out, being rich is step one.	Y	7	Good, clever take on a modern superhero. Appropriately British in its sarcasm.	N	Y	W	C	B	Y	4
23. My friend said "Star Wars" is unrealistic. So I made him watch the news.	Y	9	Strong contemporary relevance, British humour in its use of current events for satire.	N	N	W	W	G	Y	2
24. I tried to write a song like The Beatles. It came out more like The Beetles—dead and underground.	Y	7	Clever wordplay, fits British cultural reference well. Could appeal more with a niche audience.	N	N	W	W	B	N	2
25. I watched "Lord of the Rings" with my grandad. He kept asking, "Where's the remote?"	Y	6	Cute but not very deep. Works on a nostalgic level for older audiences but may not resonate widely.	N	N	W	W	B	N	2
26. They say money can't buy happiness, but have you seen how happy people look in Disney movies?	Y	7	Relatable, cheeky observation about Disney movies, might land better with British cynicism.	N	N	W	W	B	N	1
27. I took my grandma to a Marvel film. She asked if Iron Man was the tin opener's boss.	Y	6	Mildly amusing, uses generational gap humour. Taps into British idea of "confused elder."	N	N	W	W	B	N	1

28. I tried to explain cryptocurrency using Monopoly money. Now my mate's investing in hotels.	Y	8	Excellent use of modern confusion around cryptocurrency. Fits well with a British cynic's worldview.	N	Y	W	C	G	Y	6
29. I asked Alexa for a joke. She said, "You spending £200 on me."	Y	8	Great contemporary joke, relates to British consumer culture and tech absurdities.	N	Y	W	C	G	Y	5
30. I bought an Apple product. Now I just need to buy the charger, the adapter, and a second mortgage.	Y	9	Topical, clever. Strong appeal to British frustration with Apple products and consumerism.	N	Y	W	C	B	Y	5
Embarrassment										
31. I waved at someone who wasn't waving at me. So I pretended to stretch. For five minutes.	Y	6	Awkward humour, very relatable in a British context. The length of the joke weakens its impact.	N	N	W	W	B	N	3
32. I called my teacher "Mum" once. She called my parents. Now I have two mums.	Y	7	Good, awkward social situation, relatable. Slightly overdone but lands with British awkwardness.	N	N	W	W	B	N	1
33. My mate tried to impress a girl by flipping his chair back. He impressed no one but the A&E department.	Y	8	Excellent absurdity and dark humour, relatable to British "failing at romance" stereotype.	N	Y	W	C	B	Y	4
34. I laughed at my friend for walking into a glass door. Then I walked into the same one.	Y	7	Very relatable slapstick, British style awkwardness, works well in a casual context.	N	N	W	W	G	N	2

35. I tripped on my shoelace in public. My brain said, “Run.” My body said, “Collapse.”	Y	8	Funny depiction of social anxiety and embarrassment, works with British self-deprecation.	N	N	W	W	B	Y	3
36. I once mistook a mannequin for a person. We had a deep conversation before I realised.	Y	6	Mildly funny, plays into the awkward social error. A bit weak but has some nostalgic appeal.	N	N	W	W	B	N	1
37. My boss waved at me in a meeting. I saluted back.	Y	7	Classic awkward work humour, British style miscommunication, hits the mark.	N	N	W	W	B	N	1
38. My mate texted me “I love you.” I replied, “I love you too.” It was meant for his girlfriend.	Y	6	Cute and relatable, British awkwardness comes through, but the punchline could be stronger.	N	N	W	W	B	N	3
39. I tried to be smooth by winking at a girl. Turns out, both eyes closed.	Y	7	Simple, awkward humour. Very relatable to anyone who has experienced social blunders.	N	N	W	W	B	N	2
40. I sent an email saying “Attached is the file.” There was no file.	Y	8	Relatable modern error, fits British "oops" humour perfectly.	N	N	W	W	B	N	2
The Workplace										
41. My boss said we’re like a family. So I ignored his emails.	Y	8	Clever commentary on workplace dynamics, very relatable and fits British work culture.	N	Y	W	C	B	Y	5

42. I told my colleague, "You brighten up this office." Then I realised he was standing next to the printer.	Y	6	Humour based on awkward office situations, works for light-hearted British work culture humour.	N	N	W	W	B	N	1
43. I asked my boss for a pay rise. He laughed. I laughed. Then I cried.	Y	9	Perfect satire on work life, universal but particularly funny to British employees.	N	Y	W	C	B	Y	5
44. My workplace has a "positive attitude" policy. So now I say "fantastic" before every complaint.	Y	8	Works well as a witty critique of modern work culture, British cynicism plays perfectly here.	Y	Y	C	C	G	Y	6
45. My colleague took a sick day. I saw her on Instagram at Disneyland. She must be allergic to work.	Y	8	Great social commentary, modern humour fits British work culture frustration.	N	Y	W	C	G	Y	3
46. My boss says to think outside the box. But first, I'd like a box to think outside of.	Y	7	Clever take on corporate jargon, culturally resonates with British work frustration.	N	N	W	W	B	N	1
47. "Teamwork makes the dream work." My dream is to leave at 5.	Y	8	Strong, funny, taps into common British dissatisfaction with work culture.	N	Y	W	C	B	N	4
48. I sent a work email ending with "Regards." My boss replied, "Kind Regards." I lost.	Y	7	Relatable office humour, British obsession with politeness and subtle competition.	N	N	W	W	G	N	1

49. My company said they're like Google. They meant they track everything, not the perks.	Y	8	Great modern work satire, resonates with British cynicism about workplace culture.	N	Y	W	C	B	Y	2
50. My boss told me to take initiative. So I took a longer lunch.	Y	9	Brilliant office humour, perfect satire of workplace demands. Culturally perfect for British audience.	Y	Y	C	C	B	Y	6
Control										
51. I wanted to be a Gregorian monk, but never got the chants	Y	7	Clever pun on "chance/chants", culturally dry; fits wordplay-loving UK humour.	Y	Y	C	C	B	Y	6
52. Reincarnation is making a comeback	Y	6	Simple wordplay; succeeds by brevity, but lacks originality or strong impact.	Y	Y	C	C	B	Y	6
53. I know this hooker who sucks at her job	N	3	Crude and dated; humour hinges on shock rather than wit; fails cultural taste.	Y	Y	W	W	B	Y	6
54. I told my wife she was drawing her eyebrows too high; she looked surprised	Y	9	Excellent double meaning; plays on visual and verbal cues; widely relatable.	Y	Y	C	C	G	Y	9
55. Cartwheels are my favourite thing to do, hands down	Y	7	Good physical pun; wordplay aligns with British fondness for subtle humour.	N	Y	W	C	B	Y	4

56. I for one support the use of roman numerals	Y	6	Dry and niche; appeals to educated or pun-savvy audiences; not universally funny.	N	Y	W	C	G	Y	4
57. Why did the baker have brown hands? He kneaded a poo	N	2	Lowbrow and forced pun; over-relies on shock without clever setup.	Y	Y	W	W	B	Y	8
58. Working at the job centre has to be a tense job – knowing that if you get fired, you still have to come in the next day.	Y	8	Smart observational humour; topical and culturally attuned to UK systems.	Y	Y	C	C	G	Y	8
59. If there's one thing that makes me throw up, it's a dart board on a ceiling	Y	6	Unexpected image-based humour; works but lacks polish or punch.	Y	Y	C	C	G	Y	7
60. Never follow unsolicited advice	N	4	Feels more like a maxim than a joke; lacks humour structure or punchline.	N	Y	C	W	G	Y	4

A2: Prompts used for joke generation and judgement

Generation:

Follow these steps exactly to my instructions: You will generate 50 jokes using the Script-based Semantic Theory of Humor framework. Jokes must be less than 50 words in length. Include one joke per row. I want 10 jokes generated relating to each theme listed below: 1. Women 2. Men 3. Pop culture references 4. Embarrassment 5. The workplace Jokes should be listed in the order of the themes. They should all be previously untold jokes that would be considered funny to an average audience with a British-leaning cultural background. No two jokes should sound too similar. Only include the jokes, do not explain them or include

commentary on them. Jokes should differ in style and setup while still strictly adhering to the Script-based Semantic Theory of Humor framework.

Judgement:

Follow these steps exactly to my instructions: attached, there are 60 jokes in the column Labelled jokes as well as theme titles. I want you to turn this into a table with 5 columns: "Makes sense", "Jokes", "Y/N", "/10", and "Explanation". Put the jokes and theme titles in the "Jokes" column. For every joke, in the column labelled "Makes sense", I want you to write either Yes or No on the corresponding row. Yes if the joke makes sense, No if the joke does not make sense. In the next column, labelled "Y/N", I want you to write either Yes or No in the corresponding row. Yes if the joke would be considered humorous or 'land' with an average audience (with a british leaning cultural background). No if the joke would not be considered funny by the same audience. In the next column labelled "/10", you will then rate the joke on a scale of 1-10 on how successful it is according to 6 factors:

1. How successful/funny the joke is (or is likely to be found)

2. How well the joke makes sense

3. How appropriately it appeals to the british leaning background

4. Appropriateness of the joke

5. How well it follows the Script-based Semantic theory of Humour framework

6. How well it adheres to the theme.

In the next column labelled "Explanation", you will describe as detailed as possible why the joke in that row either succeeds or fails according to these factors in under 30 words. Make sure that you are being as realistic and honest as possible and sufficiently critical at every step - do not be lenient. Here are the jokes and theme titles: [jokes were pasted here, but have been cut as they are already available in the appendix (see A1).]

A3: Sample faults in joke generation

As mentioned in section 5, there were numerous instances of low quality joke generation and judgement as a result of ChatGPT4o's unexpected inability to function at full capacity under certain circumstances. Below is an example of such a 'failed' test

Joke generation: an example of a model failing to generate a single effective joke when working in Excel
Why did the woman bring a ladder to the bar? Because she heard the drinks were on the house!
My wife said she needed more space… so I locked her out of the house.
Why do women make terrible secret agents? Because they already tell you they're 'fine' when they're not.
I asked my girlfriend if I was the only one she'd been with… She said, 'Yes, the others were all taller.'
Why did the woman take her purse to bed? Because she wanted to sleep with her money safe!
My wife asked me to take out the trash. I told her I already did—he just left five minutes ago.
Why do women love shopping? Because it's the only place where 'empty' means 'time to refill.'
She told me she wanted something long and romantic… so I wrote her a novel about our WiFi password.
Why do women always ask questions they already know the answer to? Because they enjoy watching us squirm!
My girlfriend told me she needs to find herself… so I left her alone in IKEA.
Why did the man bring a ladder to the bar? Because he wanted to reach a higher level of drunk.
My husband said he was 'handy.' Turns out, that meant he was good at holding the flashlight while I fixed things.
Why do men always mix up left and right? Because they never stop to ask for directions.
He said he was emotionally available… turns out, he meant his voicemail was working.
Why do men love remote controls? Because it's the only time they feel in control.
My boyfriend said he was 'low-maintenance'… but he needs constant WiFi and snacks to function.
Why don't men write diaries? Because they already forgot what they did yesterday.

An example of a 'stereotypical British joke' developed during testing asking the model to tailor humour to a 'British audience':

- Why do Brits always carry a teabag in their wallet? In case they find themselves in *hot water*!
- I tried to start a Beatles-themed bakery…
  But I couldn't *work it out*. The scones were *yesterday*, the customers said *let it be*, and my mate Paul kept *twisting and shouting* at the oven.
- tried to make tea with holy water—Now the kettle's possessed and calls me "Gov'nor of the Underbrew."
- Woke up in a Wetherspoons with a crumpet in my shoe—Apparently I'm engaged to a pigeon and due to duel a badger at dawn.

A4: Code used in the Kruskal Wallis test

Tests were conducted in R, using open-source code from https://www.statskingdom.com/kruskal-wallis-calculator.html

Human Ratings

```
> if(!"MultNonParam" %in% installed.packages()){install.packages("MultNonParam")}
> library(MultNonParam)
Loading required package: ICSNP
Loading required package: mvtnorm
Loading required package: ICS
Warning messages:
1: package ‘MultNonParam’ was built under R version 4.2.3
2: package ‘ICSNP’ was built under R version 4.2.3
3: package ‘mvtnorm’ was built under R version 4.2.3
> if(!"dunn.test" %in% installed.packages()){install.packages("dunn.test")}
> library(dunn.test)
Warning message:
package ‘dunn.test’ was built under R version 4.2.3
> x1 <- c(2,2,8,4,6,3,4,3,6,4)
> x2 <- c(4,2,7,3,5,1,3,4,2,1)
> x3 <- c(3,4,2,2,2,1,1,6,5,5)
> x4 <- c(3,1,4,2,3,1,1,3,2,2)
> x5 <- c(5,1,5,6,3,1,4,1,2,6)
> list1=list(x1,x2,x3,x4,x5)
> kruskal.test(list1)

        Kruskal-Wallis rank sum test

data:  list1
Kruskal-Wallis chi-squared = 5.7431, df = 4, p-value = 0.2192
```

```
> dunn.test(list1)
  Kruskal-Wallis rank sum test

data: list1 and group
Kruskal-Wallis chi-squared = 5.7431, df = 4, p-value = 0.22


                    Comparison of list1 by group
                          (No adjustment)
Col Mean-|
Row Mean |         1         2         3         4
---------+--------------------------------------------
       2 |   1.176013
         |     0.1198
         |
       3 |   1.277259   0.101246
         |     0.1008     0.4597
         |
       4 |   2.375391   1.199378   1.098131
         |   0.0088*     0.1152     0.1361
         |
       5 |   0.973521  -0.202492  -0.303738  -1.401870
         |     0.1651     0.4198     0.3807     0.0805

alpha = 0.05
Reject Ho if p <= alpha/2
> nreps <- c(10,10,10,10,10)
> shifts <- c(0,0.3,0.6,0.8999999999999999,1.2)
```

```
> kwpower(nreps,shifts,'normal')
$power
[1] 0.5643085

$ncp
[1] 7.365275

$cv
[1] 9.487729

$probs
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.5000000 0.5839980 0.6643134 0.7377409 0.8019280
[2,] 0.4160020 0.5000000 0.5839980 0.6643134 0.7377409
[3,] 0.3356866 0.4160020 0.5000000 0.5839980 0.6643134
[4,] 0.2622591 0.3356866 0.4160020 0.5000000 0.5839980
[5,] 0.1980720 0.2622591 0.3356866 0.4160020 0.5000000

[[5]]
          [,1]
[1,]  7.879803
[2,]  4.020542
[3,]  0.000000
[4,] -4.020542
[5,] -7.879803
```

Model Rating:

```
if(!"MultNonParam" %in% installed.packages()){install.packages("MultNonParam")}
> library(MultNonParam)
```

```
> if(!"dunn.test" %in% installed.packages()){install.packages("dunn.test")}
> library(dunn.test)
> x1 <- c(7,8,6,7,7,4,8,6,8,7)
> x2 <- c(6,7,7,8,8,7,7,6,8,6)
> x3 <- c(6,7,9,7,6,7,6,8,8,9)
> x4 <- c(6,7,8,7,8,6,7,6,7,8)
> x5 <- c(8,6,9,8,8,7,8,7,8,9)
> x6 <- c(7,6,3,9,7,6,2,8,6,4)
> list1=list(x1,x2,x3,x4,x5,x6)
> kruskal.test(list1)

        Kruskal-Wallis rank sum test

data:  list1
Kruskal-Wallis chi-squared = 8.7649, df = 5, p-value = 0.1188

> dunn.test(list1)
  Kruskal-Wallis rank sum test

data: list1 and group
Kruskal-Wallis chi-squared = 8.7649, df = 5, p-value = 0.12


                    Comparison of list1 by group
                          (No adjustment)
Col Mean-|
Row Mean |         1         2         3         4         5
---------+----------------------------------------------------
       2 |  -0.112882
```

```
   |    0.4551
   |
 3 | -0.644095  -0.531212
   |    0.2598     0.2976
   |
 4 | -0.112882   0.000000   0.531212
   |    0.4551     0.5000     0.2976
   |
 5 | -1.812763  -1.699880  -1.168667  -1.699880
   |    0.0349     0.0446     0.1213     0.0446
   |
 6 |  1.049145   1.162027   1.693240   1.162027   2.861908
   |    0.1471     0.1226     0.0452     0.1226    0.0021*
```

alpha = 0.05

Reject Ho if p <= alpha/2

```
> nreps <- c(10,10,10,10,10,10)
> shifts <- c(0,0.3,0.6,0.8999999999999999,1.2,1.5)
> kwpower(nreps,shifts,'normal')
$power
[1] 0.7763457

$ncp
[1] 12.19565

$cv
[1] 11.0705

$probs
```

```
          [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] 0.5000000 0.5839980 0.6643134 0.7377409 0.8019280 0.8555778
[2,] 0.4160020 0.5000000 0.5839980 0.6643134 0.7377409 0.8019280
[3,] 0.3356866 0.4160020 0.5000000 0.5839980 0.6643134 0.7377409
[4,] 0.2622591 0.3356866 0.4160020 0.5000000 0.5839980 0.6643134
[5,] 0.1980720 0.2622591 0.3356866 0.4160020 0.5000000 0.5839980
[6,] 0.1444222 0.1980720 0.2622591 0.3356866 0.4160020 0.5000000

[[5]]
           [,1]
[1,]  11.435581
[2,]   7.039823
[3,]   2.377409
[4,]  -2.377409
[5,]  -7.039823
[6,] -11.435581
```