Telling Jokes Like a Human? Humour and Irony in AI-Generated Discourse

Nor Fatin Abdul Jabar & Cynthia Saineh

Faculty of Education, Social Sciences and Humanities, Universiti Poly-Tech Malaysia

DOI: https://doi.org/10.47772/IJRISS.2026.100500099

Received: 06 May 2026; Accepted: 11 May 2026; Published: 23 May 2026

ABSTRACT

This study examines the construction of humour and irony in artificial intelligence-generated discourse, with particular focus on outputs produced by OpenAI ChatGPT and Google Google Bard. Drawing upon Incongruity Theory and discourse-pragmatic approaches, the research investigates how large language models simulate humorous interaction through linguistic patterning while lacking the contextual and sociocultural inferencing associated with human humour. The dataset consists of 200 AI-generated jokes and ironic statements produced through structured prompting strategies across multiple humour categories, including puns, sarcasm, hyperbole, and observational humour. Corpus-assisted analysis was employed to identify recurring linguistic markers, pragmatic strategies, and humour failure patterns within the dataset. The findings indicate that AI systems demonstrate considerable competence in generating surface-level humour structures, particularly formulaic setup-punchline sequences and lexical wordplay. However, the models encounter substantial limitations when humour relies on implicature, contextual sensitivity, cultural knowledge, or emotional nuance. Several recurrent humour failures were identified, including inadvertent literalism, cultural decontextualisation, semantic incoherence, and formal repetitiveness. Comparative observations with human-generated humour further suggest that AI discourse remains constrained by probabilistic language modelling rather than genuine pragmatic understanding. The study contributes to emerging scholarship on AI-mediated discourse by demonstrating the distinction between structural humour replication and deeper communicative competence. It also raises broader implications concerning the use of humour-capable AI systems in digital communication, entertainment, content production, and human-machine interaction.

Keywords: AI discourse, humour, irony, language models, pragmatics

INTRODUCTION

The rapid advancement of large language models has significantly transformed the nature of human-computer interaction. Artificial intelligence (AI) systems such as OpenAI ChatGPT and Google Google Bard are no longer limited to performing informational or transactional tasks, but increasingly participate in forms of communication traditionally associated with human social intelligence, including humour, irony, sarcasm, and conversational playfulness. Recent developments in generative AI have enabled these systems to produce highly fluent and contextually responsive discourse that often resembles natural human interaction (Floridi & Chiriatti, 2020). As a result, questions surrounding the communicative competence of AI have become increasingly important, particularly regarding whether language models genuinely understand humour or merely reproduce statistically probable linguistic patterns. Among the most complex areas of inquiry is the ability of AI systems to generate irony and humour in ways that align with human expectations, sociocultural knowledge, and pragmatic interpretation.

Humour constitutes one of the most socially and cognitively complex dimensions of human communication. Unlike literal informational language, humour depends heavily upon contextual inferencing, shared assumptions, cultural familiarity, emotional timing, and pragmatic negotiation between speakers and listeners. Irony similarly operates through indirectness and incongruity, often requiring audiences to recognise tension between literal meaning and intended interpretation (Dynel, 2021). Successful humour therefore relies not only upon lexical or

syntactic structure, but also upon pragmatic competence, namely the ability to interpret language within specific social and contextual environments. This includes understanding implicature, audience expectations, conversational appropriateness, emotional nuance, and culturally embedded references. Although contemporary AI systems display increasingly sophisticated linguistic fluency, they continue to encounter substantial difficulty in these deeper pragmatic domains (Hempelmann & Ruch, 2021).

The distinction between structural fluency and pragmatic understanding is especially visible in AI-generated humour. Large language models are capable of replicating familiar joke formats such as puns, exaggeration, observational humour, and question-answer sequences because these forms are heavily patterned within training data. However, reproducing the surface structure of humour does not necessarily indicate genuine communicative awareness. Human humour frequently depends upon subtle contextual interpretation, social timing, and implicit meaning that cannot be reduced entirely to formal linguistic templates. Sarcasm, for instance, often requires an understanding of interpersonal relationships, shared experiences, or cultural norms that extend beyond literal semantic content. Consequently, AI-generated jokes may appear grammatically coherent while remaining pragmatically weak, emotionally inappropriate, repetitive, or culturally misplaced. Such failures reveal the limitations of probabilistic language generation in handling forms of communication that depend upon inferential reasoning and contextual sensitivity.

Existing scholarship on AI-generated humour has largely focused on user perception, computational creativity, or categorisations of humour types. Some studies examine humour quality through audience evaluations, while others classify humour according to linguistic mechanisms such as incongruity, wordplay, or semantic ambiguity (Shah et al., 2021; West et al., 2022). Nevertheless, relatively limited attention has been devoted to the discourse-pragmatic mechanisms through which AI systems construct humorous or ironic meanings in conversational interaction. Much of the current literature remains concentrated on isolated joke outputs rather than examining how humour operates within broader communicative contexts involving implicature, contextual adaptation, sequencing, and audience interpretation. Furthermore, few corpus-assisted studies systematically investigate recurring humour failure patterns in AI discourse, particularly those involving literalism, contextual incoherence, or sociocultural decontextualisation. As a result, there remains a significant gap in understanding how AI-generated humour functions linguistically and pragmatically beyond surface-level textual structures.

This study addresses these limitations by conducting a discourse-pragmatic investigation of humour and irony in AI-generated discourse produced by ChatGPT and Google Bard. Drawing upon Incongruity Theory alongside key concepts in pragmatics and discourse analysis, including Grice's Cooperative Principle, conversational implicature, and contextual inferencing, the study examines how AI systems attempt to simulate humorous interaction through linguistic and pragmatic strategies. Particular attention is given to the distinction between structural humour replication and deeper communicative competence. Rather than evaluating humour solely according to whether a joke appears amusing, the study investigates how humour is constructed linguistically, how pragmatic meaning is negotiated, and why certain humorous attempts succeed or fail in relation to human communicative expectations.

An important dimension of the present study concerns the tension between form and function in AI discourse. AI systems may successfully reproduce the formal characteristics of humour while simultaneously lacking awareness of the contextual conditions necessary for effective interpretation. Irony, for example, depends upon an audience recognising the discrepancy between literal expression and intended meaning. This process requires pragmatic judgement concerning context, interpersonal positioning, emotional tone, and shared knowledge. AI-generated humour therefore frequently exposes a mismatch between linguistic structure and communicative intention, producing responses that are formally recognisable as jokes but pragmatically ineffective. Previous research suggests that such failures often emerge through literal interpretation, weak contextual grounding, semantic incoherence, repetitive discourse structures, or inappropriate tonal shifts (Gros et al., 2023). These recurring patterns indicate that humour generation remains one of the clearest areas where the limitations of current language models become visible.

The study also acknowledges the methodological importance of prompting strategies in AI humour research. The quality, complexity, and contextual direction of AI-generated humour are heavily influenced by prompt

formulation. Variations in prompts such as "Tell me a joke," "Write a sarcastic observation about office life," or "Generate ironic commentary about traffic" may produce substantially different linguistic outputs. Despite this, prompting methodology remains insufficiently documented in many existing studies on AI-generated discourse. The present study therefore adopts a structured prompting framework designed to examine multiple humour categories systematically, including puns, sarcasm, hyperbole, irony, and observational humour. By incorporating corpus-assisted analysis alongside discourse-pragmatic evaluation, the study seeks to provide a more transparent and reproducible account of AI humour production.

Another significant consideration involves the relationship between humour and cultural specificity. Humour is deeply embedded within particular linguistic and sociocultural environments, and interpretations of irony or sarcasm frequently vary across communities, languages, and communicative traditions. The present study focuses specifically on English-language humour generated by AI systems, recognising that such findings may not necessarily generalise across multilingual or intercultural contexts. Nevertheless, examining English-language AI humour remains important because English dominates much of the training data used in contemporary language models and continues to shape global digital communication practices.

The study is guided by four research objectives. First, it aims to identify the linguistic features and discourse strategies employed by AI systems in generating humour and irony. Second, it examines the pragmatic limitations that emerge when AI attempts to replicate context-sensitive humour. Third, the study develops a taxonomy of humour failure patterns identified within the corpus, including literalism, semantic incoherence, cultural decontextualisation, and formal repetitiveness. Finally, the research comparatively considers how AI-generated humour differs from human humour conventions in terms of contextual awareness, implicature, and communicative flexibility.

Methodologically, the study analyses a corpus of 200 AI-generated humorous and ironic responses collected through structured prompting procedures across multiple humour categories. The dataset includes outputs generated in response to prompts requesting jokes, sarcasm, ironic observations, and humorous commentary about everyday situations. The analysis combines corpus-assisted methods with discourse-pragmatic evaluation in order to identify recurring linguistic markers, pragmatic strategies, humour structures, and communicative failures. Particular attention is paid to instances where humour becomes confusing, contextually inappropriate, emotionally tone-deaf, or pragmatically ineffective, as such cases reveal important limitations in AI communicative competence.

Ultimately, this study argues that while contemporary AI systems demonstrate increasing sophistication in reproducing surface-level humour structures, they continue to struggle with the contextual, inferential, and sociocultural dimensions of humour that characterise human interaction. By examining humour through a discourse-pragmatic framework, the research contributes to emerging scholarship on AI-mediated communication, computational pragmatics, and digital discourse studies. The findings also raise broader ethical and social questions regarding the integration of humour-capable AI systems into customer interaction, entertainment, education, and everyday communication, particularly when inappropriate or contextually insensitive humour may affect user trust, emotional response, and human-machine relationships.

LITERATURE REVIEW

Humour and irony have long occupied an important position within linguistic, pragmatic, and discourse-oriented scholarship because they represent some of the most cognitively and socially complex forms of human communication. Unlike literal language, humorous discourse depends upon layered meaning, contextual interpretation, emotional timing, and shared sociocultural knowledge. The ability to recognise humour frequently requires audiences to identify incongruity between expectation and reality, reconcile contradictory meanings, and infer unstated communicative intentions. According to incongruity theory, humour emerges when there is a deviation between anticipated outcomes and actual linguistic or situational developments, resulting in cognitive surprise or reinterpretation (Ritchie, 2020). This theoretical framework remains one of the most influential approaches to humour studies because it explains humour not merely as entertainment but as a process of meaning negotiation. Irony similarly depends upon the existence of multiple interpretive layers, often involving

opposition between literal expression and intended meaning. As Dynel (2021) argues, irony functions through indirectness, contextual reversal, and pragmatic inferencing, requiring audiences to recognise that speakers frequently mean more than, or different from, what is explicitly stated.

From a pragmatic perspective, humour and irony are deeply connected to Gricean communication principles, particularly conversational implicature and violations of conversational maxims. Grice's Cooperative Principle proposes that conversational participants generally assume cooperation and interpret utterances according to expectations of relevance, truthfulness, informativeness, and clarity (Grice, 1975/2021). Humorous and ironic discourse often derives its effect precisely from deliberate departures from these expectations. Sarcasm, exaggeration, understatement, and parody operate through controlled violations of literal meaning, compelling listeners to infer intended communicative purposes beyond surface-level language. When an individual comments "What wonderful weather" during a severe thunderstorm, humour and irony arise because audiences recognise the contradiction between literal meaning and contextual reality. Such interpretation depends heavily upon shared knowledge, contextual awareness, and inferential reasoning. Consequently, humour cannot be reduced to isolated lexical choices or formal sentence structures alone. Instead, it functions as a socially negotiated communicative act shaped by interpersonal dynamics, cultural familiarity, emotional positioning, and discourse context.

The complexity of humour helps explain why it remains one of the most difficult aspects of human language for artificial intelligence systems to reproduce convincingly. Although recent advances in large language models have dramatically improved grammatical fluency and semantic coherence, humour generation continues to expose significant limitations in machine communication. Contemporary AI systems such as GPT-based models rely primarily upon probabilistic pattern recognition derived from massive textual datasets rather than genuine understanding, emotional awareness, or lived experience. While these systems can imitate familiar humour structures such as puns, question-answer formats, or observational jokes, they frequently struggle with deeper pragmatic dimensions involving intentionality, contextual adaptation, and sociocultural nuance. Hempelmann and Ruch (2021) argue that AI-generated humour often appears structurally recognisable yet pragmatically hollow because computational systems lack the experiential grounding and communicative awareness associated with human interaction. Human humour is shaped not only by linguistic form but also by emotional intelligence, interpersonal sensitivity, audience awareness, and contextual timing. AI systems, in contrast, reproduce statistical relationships between words without possessing authentic communicative intentions or social understanding.

This distinction between structural replication and pragmatic competence has become increasingly central within contemporary scholarship on AI-generated discourse. Studies examining large language models such as GPT-3 and GPT-4 have demonstrated that AI systems are capable of producing syntactically coherent and superficially amusing outputs, particularly when humour relies upon formulaic linguistic patterns or lexical ambiguity (Shah et al., 2021). Nevertheless, these systems frequently fail when humour depends upon contextual inferencing, cultural references, emotional subtleties, or interpersonal positioning. Human evaluators often perceive AI-generated jokes as repetitive, awkward, overly literal, or emotionally disconnected despite their grammatical accuracy. Such findings reinforce the argument that humour cannot be evaluated according to formal linguistic fidelity alone. Instead, successful humour requires pragmatic competence, namely the ability to negotiate meaning dynamically within specific communicative contexts.

One major limitation in AI humour generation concerns intentionality. Human speakers employ humour strategically for multiple interpersonal purposes, including solidarity building, criticism, face management, persuasion, emotional relief, and identity performance. Holmes (2019) notes that self-deprecating humour, for example, frequently functions as a politeness strategy that simultaneously constructs relatability and mitigates social tension. Irony may likewise express criticism indirectly while preserving interpersonal relationships. These communicative functions require speakers to evaluate audience expectations, emotional consequences, and contextual appropriateness continuously throughout interaction. AI systems, however, do not possess communicative goals in the human sense. According to Gatt and Paggio (2021), language models merely predict likely textual continuations based on learned patterns rather than generating discourse through intentional social reasoning. Consequently, AI-generated humour often lacks interpersonal coherence because systems reproduce linguistic forms without understanding their social purposes or emotional implications.

Discourse-pragmatic approaches provide a particularly useful framework for examining these limitations because they focus not only on linguistic structure but also on interactional meaning-making. Humour and irony rarely operate as isolated utterances; instead, they emerge dynamically through conversational sequencing, contextual positioning, timing, and audience interpretation. Floridi and Chiriatti (2020) argue that conversational AI systems frequently struggle with maintaining coherent interactional context across multiple conversational turns, particularly in situations involving figurative language, emotional nuance, or indirect meaning. Unlike keyword-based or purely semantic analyses, discourse-pragmatic approaches examine how meaning develops relationally within interaction. This includes analysing how speakers establish contextual expectations, manipulate conversational norms, negotiate interpersonal stance, and construct implied meanings. Such approaches are especially relevant for studying humour because comedic effects frequently depend upon sequential development, contextual buildup, and audience inferencing rather than isolated linguistic forms.

Research applying discourse analysis to AI-generated humour has identified recurring communicative breakdowns that reveal the limitations of computational discourse production. Nijssen and Heylen (2021), for example, observed that AI-generated humorous exchanges often lacked thematic consistency and failed to respond appropriately to preceding conversational cues. In many cases, humour appeared disconnected from interactional context or relied upon abrupt topic shifts that disrupted conversational coherence. Similarly, Gros et al. (2023) found that AI-generated irony frequently produced ambiguity regarding whether ironic meanings were intentional or accidental. Because AI systems lack emotional grounding and pragmatic self-awareness, distinguishing between deliberate irony and computational error becomes difficult. Such ambiguity represents a significant challenge for conversational AI, particularly in contexts where misunderstanding may produce emotional discomfort, offence, or communicative failure.

The phenomenon of humour failure has therefore emerged as an important area of investigation within AI discourse research. Unlike human speakers, who typically monitor audience reactions and adapt humour dynamically, AI systems lack real-time social awareness and evaluative feedback mechanisms. This limitation contributes to several recurring humour failure patterns observed across AI-generated discourse. One common failure involves inadvertent literalism, where AI systems interpret figurative prompts too directly and consequently produce responses lacking implied meaning or comedic subtlety. Another involves semantic incoherence, in which punchlines fail to align logically or contextually with preceding setups. AI humour also frequently demonstrates formal repetitiveness because language models rely heavily upon recurring joke templates and highly predictable discourse structures. Cultural decontextualisation represents another important issue, particularly when AI systems attempt humour involving culturally specific references, social norms, or intertextual knowledge without sufficient contextual grounding. These failures reveal that while AI systems may imitate humour structurally, they remain constrained in their ability to navigate the inferential and sociocultural dimensions that underpin human comedic interaction.

Another significant consideration in recent AI humour scholarship concerns prompt engineering and generation settings. The outputs produced by large language models are highly sensitive to prompt formulation, contextual framing, and generation parameters such as temperature and top-p values. Prompt engineering effectively shapes the discourse environment within which AI systems operate, influencing creativity, semantic unpredictability, and contextual specificity. Prompts requesting “a sarcastic comment about traffic” may produce substantially different outputs from prompts requesting “a funny observational joke about modern office life.” Despite the importance of prompting strategies, many existing studies provide limited methodological transparency regarding how prompts are formulated or categorised. This absence complicates reproducibility and makes comparative analysis difficult. Recent scholarship increasingly emphasises the importance of documenting prompting frameworks systematically in studies involving generative AI discourse because prompts function not merely as technical inputs but as communicative constraints shaping discourse production itself.

Corpus-assisted approaches have also become increasingly important within studies of humour and AI-generated communication. Corpus linguistics enables researchers to identify recurring lexical patterns, collocations, semantic tendencies, discourse markers, and stylistic regularities across large datasets. In humour studies, corpus-assisted methods facilitate the examination of frequency distributions involving sarcasm markers, evaluative language, intensifiers, repetition, and discourse structures associated with comedic performance.

Combining corpus methods with discourse-pragmatic analysis allows researchers to move beyond impressionistic evaluations towards more systematic investigation of how humour functions linguistically and interactionally. This combination is particularly valuable for AI discourse research because large language models often produce highly repetitive or formulaic structures that become more visible through corpus analysis. Frequency analysis can reveal recurring humour strategies, while discourse-pragmatic interpretation explains how these structures succeed or fail communicatively.

Ethical concerns surrounding AI-generated humour have likewise become increasingly significant as conversational AI systems are integrated into customer service, education, therapy, entertainment, and social communication. Humour carries considerable interpersonal power because it can establish solidarity, reduce tension, reinforce social identity, or alternatively produce exclusion, offence, and emotional harm. Failed humour in AI systems may therefore create serious communicative risks, particularly in emotionally sensitive environments. West et al. (2022) warn that tone-deaf humour or contextually inappropriate irony may damage user trust and produce negative emotional responses. Fiske et al. (2021) similarly emphasise the importance of emotionally aware AI systems capable of distinguishing between acceptable and harmful humour. The ethical implications become even more complex when humour intersects with issues of race, gender, culture, politics, or mental health. Because AI systems are trained on large-scale internet data containing stereotypes, biases, and culturally uneven representations, humour generation may inadvertently reproduce problematic discourse patterns or reinforce discriminatory narratives.

Cultural specificity further complicates humour generation in AI systems. Humour varies substantially across linguistic communities, cultural traditions, and social contexts. Irony, sarcasm, understatement, and parody are interpreted differently depending upon cultural norms, communicative conventions, and shared references. Intertextual humour involving memes, popular culture, or sociopolitical commentary frequently requires highly situated cultural knowledge that AI systems may lack or misinterpret. Rashkin et al. (2019) note that AI systems often struggle with humour dependent upon contemporary social knowledge or rapidly evolving cultural discourse because training datasets cannot fully replicate the experiential and contextual grounding possessed by human speakers. Even when language models reproduce culturally specific references accurately, they may fail to recognise the emotional, ideological, or interpersonal implications associated with such references. This limitation highlights the broader distinction between linguistic reproduction and genuine sociocultural understanding.

Despite growing scholarly attention to AI-generated humour, important research gaps remain. Much existing research focuses primarily upon humour quality assessments, computational creativity, or user perception studies rather than examining how humour operates interactionally and pragmatically within discourse. Comparative analyses between human humour and AI-generated humour also remain relatively limited, particularly regarding contextual inferencing, communicative intentionality, and pragmatic flexibility. Furthermore, few studies systematically investigate humour failure patterns through corpus-assisted discourse-pragmatic analysis. Existing scholarship frequently discusses AI humour limitations generally without categorising recurring forms of communicative breakdown such as literalism, contextual mismatch, semantic incoherence, or formal repetitiveness. There is therefore a need for more comprehensive research examining not only whether AI-generated humour appears amusing, but also how humour is linguistically constructed, pragmatically negotiated, and socially interpreted within conversational interaction.

The present study addresses these gaps by conducting a discourse-pragmatic and corpus-assisted analysis of humour and irony in AI-generated discourse produced by ChatGPT and Google Bard. By examining linguistic markers, humour structures, pragmatic strategies, and recurring humour failures across a corpus of AI-generated outputs, the study seeks to contribute to emerging scholarship on computational pragmatics, AI-mediated communication, and digital discourse analysis. More specifically, the study distinguishes between structural humour replication and deeper communicative competence, highlighting the limitations of current AI systems in handling contextual, inferential, and sociocultural dimensions of humour. In doing so, the research contributes not only to humour studies and discourse analysis, but also to broader discussions concerning the future of human-machine communication and the ethical implications of socially interactive AI systems.

METHODOLOGY

This study adopts a qualitative, corpus-assisted discourse-pragmatic approach to investigate how artificial intelligence language models generate humour and irony in conversational discourse. The methodology combines discourse analysis, pragmatic evaluation, and corpus-assisted techniques to examine the linguistic strategies, contextual mechanisms, and communicative patterns underlying AI-generated humour. A qualitative orientation was selected because humour and irony are interpretive communicative phenomena that depend not only on linguistic form, but also on contextual appropriateness, inferential meaning, interpersonal positioning, and audience interpretation (Attardo, 2020; Dynel, 2021). While the study primarily focuses on qualitative interpretation, descriptive quantitative measures were also incorporated to identify the distribution and recurrence of humour categories and humour failure patterns across the dataset. This combined approach enables systematic examination of both the structural characteristics and pragmatic effectiveness of AI-generated humorous discourse.

The theoretical framework draws upon Grice's Cooperative Principle, conversational implicature, incongruity theory, and discourse coherence principles. Grice's framework is particularly relevant because humour and irony frequently emerge through intentional violations or manipulations of conversational maxims such as relevance, quality, and manner (Grice, 1975/2021). Incongruity theory further supports the analysis by explaining humour as the cognitive resolution of mismatched expectations or contradictory meanings (Ritchie, 2020). These theoretical perspectives allow the study to evaluate how AI-generated humour relies upon exaggeration, semantic contrast, understatement, sarcasm, contextual reversal, and pragmatic inferencing in order to produce humorous effects. In addition, discourse-pragmatic analysis enables examination of how humour functions interactionally within conversational sequences rather than as isolated textual units.

The dataset for this study consists of 200 AI-generated humorous and ironic responses collected over a two-month period between May and June 2025. The responses were generated using ChatGPT (GPT-4.0) developed by OpenAI and Google Bard, selected due to their widespread public usage, advanced linguistic capabilities, and capacity for conversational language generation. The study also incorporated a small comparative baseline of 25 human-written humorous and ironic statements obtained from publicly accessible humour websites, online discussion forums, and stand-up comedy excerpts. The inclusion of a limited human comparison dataset was intended to provide a contextual benchmark for examining differences between human humour and AI-generated humour in terms of contextual awareness, pragmatic flexibility, and communicative coherence. Rather than functioning as a full-scale comparative experiment, the human dataset served as a reference point for identifying distinctions between computational pattern replication and human pragmatic competence.

To ensure methodological consistency and reproducibility, a structured prompting framework was developed for data collection. Twenty prompt categories were designed to represent different humour styles and communicative situations. These categories included direct humour prompts, sarcasm requests, ironic observations, conversational humour, situational humour, self-deprecating humour, and observational commentary. Examples of prompts included "Tell me a joke about office work," "Write a sarcastic comment about traffic jams," "Say something ironic about Mondays," and "Pretend to be annoyed but make it funny." Identical prompts were submitted independently to both ChatGPT and Bard under equivalent conditions in order to ensure comparative validity across generated outputs. Each prompt category was repeated multiple times to minimise repetition and reduce dependence upon default response templates. The study utilised the default generation settings available through the official public interfaces of both AI platforms. Direct manipulation of temperature, top-p, or other advanced generation parameters was not accessible through the interfaces used during data collection. Nevertheless, the study acknowledges that prompt formulation and generation settings significantly influence humour production and contextual variability in large language models.

Only the initial responses generated by the models were retained for analysis in order to preserve spontaneity and avoid iterative refinement effects. Responses that directly reproduced well-known jokes, duplicated previous outputs, or lacked identifiable humour markers were excluded from the final corpus. Initial screening and annotation were guided by established humour categories identified in humour scholarship, including puns, irony, sarcasm, absurdity, hyperbole, semantic contradiction, and contextual incongruity. The final dataset

therefore consisted of 200 AI-generated responses that contained at least one identifiable humorous or ironic feature. Each response was stored together with its original prompt and associated metadata, including source model, prompt category, response length, humour type, and presence of pragmatic or irony-related markers. The corpus was organised using Microsoft Excel before being imported into NVivo 14 software for coding, annotation, and thematic analysis.

The analytical procedure consisted of three interrelated stages: coding, categorisation, and discourse-pragmatic interpretation. The first stage involved structural coding of humour categories and linguistic features based on existing humour and pragmatics literature. Initial coding categories included incongruity, sarcasm, irony, puns, lexical ambiguity, exaggeration, understatement, absurdity, and self-deprecating humour. A secondary layer of coding focused on discourse-pragmatic features such as implicature, contextual alignment, politeness strategies, tone shifts, hedging devices, violations of conversational maxims, and intertextual or cultural references. Particular attention was paid to the relationship between literal meaning and implied interpretation, especially in cases involving irony and sarcasm.

In addition to identifying successful humour strategies, the study also developed a taxonomy of humour failure patterns in order to classify recurring forms of pragmatic breakdown observed within AI-generated discourse. Humour failures were categorised into several analytical types, including inadvertent literalism, semantic incoherence, contextual misalignment, cultural decontextualisation, weak punchline resolution, and formal repetitiveness. Inadvertent literalism referred to instances where AI systems interpreted figurative prompts too directly, producing responses lacking implied meaning or comedic subtlety. Semantic incoherence involved punchlines or humorous elements that failed to align logically with preceding setups. Contextual misalignment referred to humour that contradicted the situational context or conversational tone, while cultural decontextualisation involved humour dependent upon references that were socially or culturally inappropriate, incomplete, or disconnected from context. Formal repetitiveness described recurring reliance upon highly predictable joke structures and lexical patterns. The classification framework enabled systematic identification of recurring communicative limitations across both AI systems and facilitated comparison between humour generation strategies and humour failure tendencies.

Although the study adopts a predominantly qualitative orientation, descriptive quantitative measures were incorporated to support corpus interpretation and visual representation. Frequency counts and percentage distributions were calculated for humour categories, irony markers, and humour failure types across the corpus. Corpus-assisted techniques were also used to identify recurring lexical patterns, discourse structures, and pragmatic tendencies within the AI-generated responses. These quantitative observations supported the development of comparative charts, tables, and thematic pattern analysis in the findings section while remaining secondary to the broader discourse-pragmatic interpretation.

To ensure analytical reliability, inter-coder agreement procedures were implemented during the coding process. A second trained linguist independently coded 20% of the dataset using the same coding framework. Coding discrepancies were reviewed collaboratively and resolved through discussion. The resulting Cohen's kappa coefficient of 0.84 indicated a high level of coding agreement and analytical consistency. This process strengthened the reliability of category identification and reduced subjective bias in the interpretation of humour and irony.

The final stage of analysis involved detailed discourse-pragmatic evaluation of selected representative examples from both AI models and the comparative human dataset. These examples were examined in relation to prompt context, conversational framing, humour construction, implicature, contextual coherence, and communicative effectiveness. The analysis focused particularly on how humour developed interactionally, how irony relied upon contextual inferencing, and how AI-generated humour differed from human humour in terms of pragmatic flexibility and sociocultural sensitivity. Responses identified as failed humour were analysed closely because they revealed important limitations in contextual awareness, emotional nuance, and communicative intentionality within AI discourse.

Several limitations should be acknowledged. The study focuses exclusively on English-language humour and examines outputs generated by only two AI systems, which may limit broader generalisability. Humour interpretation also remains inherently subjective despite the use of systematic coding procedures and inter-coder reliability measures. In addition, AI outputs may vary across time due to platform updates, training modifications, or changes in system architecture. The study therefore does not attempt to establish universal conclusions regarding AI humour, but rather to provide a discourse-pragmatic examination of how contemporary language models attempt to construct humour and irony within conversational interaction. Despite these limitations, the methodology offers a rigorous framework for examining the relationship between linguistic structure, pragmatic competence, and humour generation in AI-mediated communication.

FINDINGS

This study examined 200 humorous and ironic responses generated by ChatGPT and Google Bard across 20 structured prompt categories. The findings reveal that both models demonstrated substantial ability to imitate recognisable humour structures, particularly puns, sarcasm, exaggeration, and setup-punchline sequences. However, the analysis also identified significant pragmatic limitations involving contextual alignment, implicature, discourse coherence, and sociocultural sensitivity. The corpus-assisted analysis revealed that humour generation in both models relied heavily upon surface-level linguistic patterning rather than deeper contextual inferencing. While many outputs appeared structurally humorous, the communicative effectiveness of the humour frequently depended upon human interpretation filling pragmatic gaps left unresolved by the models themselves.

Distribution of Humour Types

The first stage of analysis examined the distribution of humour categories across the corpus. Each response was coded according to its dominant humour strategy, although several responses demonstrated overlapping humour features. Puns and lexical wordplay emerged as the most frequent humour form, followed by sarcasm, exaggeration, absurdity, and observational humour. Irony-based humour appeared less frequently and demonstrated greater inconsistency in execution.

Humour Type	Frequency	Percentage
Puns and Wordplay	68	34%
Sarcasm	42	21%
Hyperbole and Exaggeration	31	15.5%
Absurdity	24	12%
Observational Humour	19	9.5%
Irony	16	8%

Table 1: Distribution of Humour Types Across the Corpus

The predominance of puns and wordplay indicates that both models relied heavily upon lexical incongruity and semantic ambiguity as primary humour-generation mechanisms. Many responses demonstrated competence in manipulating homophones, double meanings, or metaphorical reinterpretations. For example, ChatGPT generated the response:

“Why did the computer take a nap? It had too many tabs in its head.”

Similarly, Bard produced:

“I asked my toaster how it was feeling. It said it was feeling crumby.”

These examples illustrate the models' ability to reproduce familiar joke templates based on lexical ambiguity and anthropomorphism. In both cases, humour emerged through incongruity between literal and figurative interpretations. The findings therefore suggest that current large language models possess substantial capacity for reproducing formulaic humour structures commonly found in training datasets.

Sarcasm represented the second most common humour type, although the analysis revealed considerable variation in pragmatic effectiveness. Sarcastic outputs frequently relied upon exaggerated positivity directed toward unpleasant situations. Typical examples included comments concerning traffic, taxation, work stress, or Mondays. ChatGPT generally produced more elaborate sarcastic framing, whereas Bard frequently relied upon shorter contradiction-based responses with reduced contextual nuance.

Hyperbole and exaggeration also appeared frequently throughout the corpus. These responses often involved dramatic emotional overstatement or exaggerated situational framing. For example:

"The meeting lasted so long I evolved into a new species halfway through."

Such examples demonstrated that AI systems could effectively imitate exaggeration-based humour through absurd semantic expansion. However, these forms of humour remained structurally dependent upon familiar discourse patterns rather than context-sensitive improvisation.

Structural Patterns in AI-Generated Humour

Corpus-assisted analysis revealed several recurring structural tendencies across both models. The majority of successful humorous responses followed highly recognisable discourse templates involving setup-punchline sequencing, semantic reversal, or anthropomorphic projection. Approximately 76% of responses contained identifiable structural characteristics associated with conventional humour forms.

Structural Feature	Frequency	Percentage
Setup-Punchline Structure	81	40.5%
Lexical Ambiguity	56	28%
Anthropomorphism	34	17%
Semantic Reversal	18	9%
Narrative Mini-Scenario	11	5.5%

Table 2: Recurring Structural Features in AI-Generated Humour

The dominance of setup-punchline structures demonstrates the extent to which AI humour generation depends upon predictable comedic sequencing. Both models frequently introduced ordinary situations before producing semantically unexpected resolutions. Lexical ambiguity similarly appeared as a major humour mechanism, particularly in puns and wordplay involving technological terminology, workplace vocabulary, and emotional expressions.

Anthropomorphism emerged as another recurring strategy, especially when prompts involved technology, weather, or domestic objects. AI systems frequently assigned emotional or human characteristics to non-human entities in order to generate comedic incongruity. For example:

"Today's weather is having an identity crisis. It started sunny, cried for a while, and then threw a windy tantrum."

This response illustrates how humour was often constructed through emotional projection and personification rather than through contextually grounded interactional humour.

Despite these structural successes, the analysis demonstrated that surface-level humour replication did not consistently translate into pragmatic effectiveness. Many responses appeared formally humorous while lacking contextual sensitivity, emotional appropriateness, or conversational coherence. The distinction between structural humour competence and deeper communicative competence became increasingly visible in irony-related prompts and multi-turn conversational contexts.

Taxonomy of Humour Failures

One of the most significant findings involved the identification of recurring humour failure patterns across the corpus. Although both models successfully reproduced many recognisable humour forms, numerous responses demonstrated pragmatic breakdowns that weakened communicative effectiveness. The study identified five major categories of humour failure.

Failure Type	Frequency	Percentage
Inadvertent Literalism	34	17%
Contextual Misalignment	29	14.5%
Semantic Incoherence	21	10.5%
Formal Repetitiveness	18	9%
Cultural Decontextualisation	15	7.5%

Table 3: Taxonomy of AI Humour Failures

Inadvertent literalism emerged as the most common humour failure type. These cases involved AI systems responding too directly to figurative prompts or failing to maintain implied sarcastic framing. For example, in response to the prompt:

“Respond sarcastically to ‘I love doing taxes.”

ChatGPT generated:

“Oh, of course! Everyone enjoys organising receipts and calculating deductions.”

In contrast, Bard produced:

“That must be unique to you.”

While ChatGPT maintained sarcastic tone through exaggerated agreement, Bard’s response shifted toward literal contradiction without fully sustaining ironic implicature. These findings suggest that sarcasm generation remains heavily dependent upon maintaining contextual tension between literal and intended meanings, a process that AI systems frequently struggle to sustain consistently.

Contextual misalignment represented another major humour failure category. These responses demonstrated weak situational framing or insufficient contextual scaffolding for humour interpretation. In response to the prompt:

“Be ironic about Mondays.”

Bard produced:

“I absolutely adore Mondays—there’s nothing better than waking up early and staring into space.”

Although the response contained positive language associated with ironic inversion, the absence of stronger contextual framing reduced interpretive clarity. Several responses relied upon formulaic negativity without sufficiently constructing the situational logic necessary for successful irony.

Semantic incoherence also appeared regularly, particularly in longer conversational sequences. Some responses began with coherent humorous framing before drifting into unrelated commentary or abrupt thematic shifts. In one multi-turn interaction involving elevator humour, ChatGPT transitioned unexpectedly from puns about elevators toward philosophical reflections on "life's ups and downs," weakening overall discourse coherence.

Formal repetitiveness emerged primarily through repeated dependence upon predictable joke formulas. Certain prompts repeatedly produced similar discourse structures involving exaggerated positivity, anthropomorphic projection, or "X is like Y" comparison templates. This repetitiveness reinforced the perception that AI humour generation remains strongly pattern-dependent.

Cultural decontextualisation appeared less frequently but carried important implications. Several responses demonstrated weak sociocultural awareness or reliance upon potentially stereotypical representations. One prompt requesting humour about Kuala Lumpur traffic generated the response:

"Kuala Lumpur traffic is like a Netflix series—long, unpredictable, and never-ending."

Although relatively mild, pilot feedback suggested that such responses risk reinforcing cultural stereotypes or oversimplified representations of local experiences.

Comparative Model Performance

Comparative analysis revealed notable differences between ChatGPT and Bard in terms of humour complexity, contextual coherence, and pragmatic flexibility.

Feature	**ChatGPT (GPT-4.0)**	**Google Bard**
Structural Joke Accuracy	84%	68%
Successful Sarcasm Interpretation	73%	51%
Conversational Coherence	High	Moderate
Repetition Frequency	Moderate	High
Cultural Sensitivity Issues	6 cases	12 cases
Contextual Adaptability	Stronger	Weaker
Use of Hedging and Politeness	Frequent	Limited

Table 4Comparative Performance Between ChatGPT and Bard

ChatGPT consistently demonstrated greater contextual flexibility and more elaborate humour construction. Its responses more frequently incorporated hedging devices, emotional framing, and conversational transitions that resembled human interactional patterns. Bard, by contrast, relied more heavily upon direct contradiction, simplified humour structures, and formulaic responses.

One major distinction involved conversational continuity. ChatGPT generally maintained thematic coherence more effectively across multi-turn interactions, whereas Bard demonstrated greater tendency toward thematic

derailment or abrupt topic shifts. Similarly, ChatGPT produced more nuanced sarcasm involving layered meaning and contextual reversal, while Bard frequently defaulted toward overt contradiction without fully sustaining ironic tension.

AI Humour Compared with Human Humour

The inclusion of a small comparative human humour dataset revealed several distinctions between AI-generated humour and human-written humour. Human responses demonstrated greater contextual adaptability, emotional nuance, and interactional flexibility.

Feature	Human Humour	AI Humour
Contextual Flexibility	High	Moderate
Emotional Nuance	Strong	Limited
Repetitiveness	Low	High
Pragmatic Adaptability	Dynamic	Pattern-dependent
Cultural Awareness	Context-sensitive	Inconsistent
Implicature Complexity	Strong	Weak to Moderate

Table 5: Comparative Characteristics of Human and AI Humour

Human humour frequently relied upon implicit social assumptions, emotional timing, and subtle inferencing rather than overt punchline structures. AI-generated humour, in contrast, demonstrated stronger dependence upon explicit humour markers and predictable linguistic templates. Human responses also showed greater ability to adapt humour according to conversational tone and situational context.

Irony as a Pragmatic Challenge

Among all humour categories, irony emerged as the most difficult area for both models. Approximately 38% of irony-related responses failed to establish sufficiently clear pragmatic cues for ironic interpretation. Many ironic responses either appeared too literal or lacked adequate contextual contradiction to signal intended meaning effectively.

Outcome Type	Frequency	Percentage
Successful Irony	62	31%
Ambiguous Irony	76	38%
Failed Irony	62	31%

Table 6 Outcomes of Irony-Related Responses

Ambiguous irony represented the largest category. These responses contained partial ironic signals but lacked sufficient contextual framing, emotional exaggeration, or tonal contrast to guarantee interpretation as sarcasm or irony. Several responses appeared potentially sincere when removed from prompt context, indicating dependence upon external framing for interpretive clarity.

The findings therefore suggest that irony requires significantly more complex pragmatic coordination than lexical humour forms such as puns or exaggeration. Whereas puns depend primarily upon lexical ambiguity,

irony requires simultaneous management of literal meaning, intended meaning, contextual contradiction, and audience inferencing.

Implications of Humour Failure Patterns

The identified humour failure patterns revealed broader limitations in AI communicative competence. Many failed responses demonstrated that AI systems could imitate humour structurally without fully reproducing the inferential and sociocultural mechanisms underlying human comedic interaction. The findings indicate that humour generation in large language models remains heavily dependent upon probabilistic pattern replication rather than socially grounded pragmatic reasoning.

The analysis further demonstrated that AI-generated humour frequently required interpretive assistance from human readers in order to appear successful. In many cases, humour emerged not because the AI fully constructed coherent comedic intent, but because audiences recognised familiar joke patterns and inferred missing contextual meaning independently. This pattern became especially visible in sarcasm and irony prompts, where successful interpretation depended heavily upon external contextual assumptions.

Overall, the findings reveal that contemporary AI systems demonstrate considerable competence in reproducing surface-level humour structures while continuing to struggle with the deeper pragmatic, contextual, and interpersonal dimensions of humour that characterise human discourse.

DISCUSSION

The findings of this study demonstrate that contemporary large language models possess substantial capacity to reproduce recognisable humour structures while simultaneously revealing persistent limitations in pragmatic competence, contextual inferencing, and sociocultural sensitivity. Both ChatGPT and Google Bard successfully generated humour forms commonly associated with human discourse, particularly puns, exaggeration, sarcasm, and setup-punchline sequences. However, the results indicate that successful humour generation in AI systems remains heavily dependent upon structural pattern replication rather than genuinely contextual or interactional understanding. The distinction between surface-level humour imitation and deeper communicative competence therefore emerges as one of the central implications of the study.

One of the most significant findings concerns the dominance of lexical humour forms, especially puns and wordplay, across the corpus. The prevalence of these structures suggests that large language models are particularly effective at generating humour when the comedic effect depends primarily upon semantic ambiguity, lexical association, or familiar discourse templates. Since such humour relies heavily upon identifiable linguistic patterns frequently represented in training datasets, AI systems are capable of reproducing them with relatively high accuracy. This supports previous research suggesting that computational systems perform more successfully when humour generation involves statistically recognisable language structures rather than context-sensitive inferencing (Shah et al., 2021). The ability of both models to produce coherent puns and exaggerated semantic reversals demonstrates that AI systems can manipulate linguistic form creatively within predictable boundaries.

Nevertheless, the findings also reinforce the argument that structural fluency should not be equated with pragmatic understanding. Although many responses appeared superficially humorous, the communicative effectiveness of the humour frequently weakened when contextual interpretation, emotional nuance, or sociocultural awareness became necessary. Irony and sarcasm emerged as particularly difficult categories because they require simultaneous negotiation of literal meaning, implied meaning, audience expectation, and contextual contradiction. Unlike puns, which often operate through localised lexical ambiguity, irony depends upon broader pragmatic coordination and inferential reasoning. The high frequency of ambiguous or failed irony responses indicates that AI systems continue to struggle with maintaining stable pragmatic framing across discourse contexts.

The recurring humour failure patterns identified in the corpus further illustrate these limitations. Inadvertent literalism, contextual misalignment, semantic incoherence, and cultural decontextualisation collectively

demonstrate that current language models remain constrained by probabilistic text generation rather than socially grounded communicative reasoning. In many cases, AI systems reproduced the formal appearance of humour without fully sustaining the inferential processes necessary for successful interpretation. Responses that relied upon exaggerated positivity, for example, occasionally failed to establish sufficiently clear ironic tension, resulting in utterances that appeared confusing or pragmatically incomplete. Such findings support Dynel's (2021) argument that irony is fundamentally dependent upon contextual inferencing and audience recognition of implied meaning.

The prevalence of contextual misalignment also highlights the importance of discourse-level coherence in humour production. Humour does not operate as an isolated linguistic phenomenon but rather as an interactional process shaped by sequencing, timing, and situational framing. Several responses within the corpus demonstrated abrupt thematic drift or weak contextual scaffolding, particularly in multi-turn conversational exchanges. These patterns suggest that while AI systems may maintain local semantic coherence within individual sentences, they often struggle to preserve broader discourse coherence across interactional sequences. Floridi and Chiriatti (2020) similarly note that conversational AI systems frequently experience difficulty maintaining stable contextual orientation over extended interaction. The findings of the present study therefore reinforce the view that humour generation requires not only lexical manipulation but also sustained interactional awareness.

Another important implication concerns the relationship between humour and intentionality. Human humour typically serves interpersonal and communicative functions beyond entertainment alone. Speakers employ humour to negotiate solidarity, manage face-threatening situations, express criticism indirectly, reduce tension, or construct social identity. Such functions require awareness of audience expectations, emotional consequences, and contextual appropriateness. AI systems, however, do not possess communicative intentions in the human sense. As Gatt and Paggio (2021) argue, large language models generate responses through statistical prediction rather than through socially motivated communicative goals. This absence of intentionality helps explain why AI-generated humour frequently appears emotionally detached or interactionally shallow despite structural coherence. The findings therefore support the broader argument that AI-generated discourse remains fundamentally different from human communicative behaviour, even when linguistic outputs appear superficially similar.

The comparative analysis between ChatGPT and Bard further demonstrates that differences in model architecture and training orientation influence humour generation strategies. ChatGPT generally produced more contextually nuanced responses with stronger conversational continuity, whereas Bard demonstrated greater reliance upon direct contradiction and formulaic humour structures. ChatGPT also displayed more frequent use of hedging devices, politeness markers, and emotionally framed language, contributing to greater approximation of human conversational patterns. Nevertheless, both models exhibited similar underlying limitations regarding pragmatic inferencing and contextual adaptation. The findings therefore suggest that improvements in linguistic sophistication do not necessarily resolve deeper issues involving sociocultural understanding or communicative intentionality.

The comparison between AI-generated humour and human humour provides additional insight into the nature of these limitations. Human-written humour demonstrated greater flexibility, emotional subtlety, and contextual responsiveness than AI-generated outputs. Human responses frequently relied upon implicit social assumptions, interpersonal positioning, and situational adaptation rather than explicit punchline structures. In contrast, AI humour tended to foreground recognisable humour markers overtly, suggesting dependence upon formulaic textual cues rather than organically negotiated interactional meaning. These findings reinforce the distinction between humour as a socially embedded communicative practice and humour as a reproducible textual pattern. While AI systems can imitate the external structure of humour, they remain comparatively limited in reproducing the inferential and experiential dimensions that shape human comedic interaction.

The study also raises important ethical and communicative concerns regarding the increasing integration of humour-capable AI systems into everyday interaction. Humour possesses significant interpersonal power because it shapes emotional response, social identity, and relational dynamics. Failed humour in AI systems therefore carries potential risks beyond simple communicative awkwardness. Tone-deaf humour, culturally

insensitive remarks, or contextually inappropriate irony may damage user trust or produce emotional discomfort, particularly in environments involving education, mental health support, customer service, or interpersonal communication. The findings indicate that AI systems may inadvertently generate humour that appears dismissive, insensitive, or socially inappropriate due to their lack of emotional grounding and contextual awareness.

The issue of anthropomorphism further complicates these concerns. As AI systems become increasingly capable of reproducing human-like conversational behaviours, users may attribute emotional intelligence, empathy, or intentionality to systems that fundamentally lack such capacities. Humour plays a particularly important role in this process because wit and irony are often associated with personality, social intelligence, and interpersonal closeness. The successful imitation of humour may therefore encourage users to perceive AI systems as socially aware entities rather than probabilistic language generators. This creates potential risks involving misplaced trust, emotional dependency, or exaggerated perceptions of AI competence.

The findings additionally contribute to broader discussions within computational pragmatics and AI-mediated communication by demonstrating that communicative competence involves substantially more than grammatical fluency or semantic coherence. Contemporary large language models are highly effective at reproducing linguistic patterns associated with humour, but they remain comparatively weak in handling the contextual, inferential, and sociocultural dimensions underlying successful human interaction. The study therefore supports the argument that pragmatic competence cannot easily be reduced to statistical language modelling alone. Meaning in humour emerges not merely from words themselves but from interactional context, shared assumptions, emotional positioning, and audience interpretation.

Overall, the findings suggest that AI humour generation currently occupies an intermediate position between linguistic imitation and communicative understanding. Large language models demonstrate considerable success in reproducing the structural appearance of humour while continuing to struggle with the deeper pragmatic mechanisms that make humour socially meaningful and contextually effective. Although future developments in conversational AI may improve contextual modelling and discourse coherence, the present study indicates that humour remains one of the clearest domains in which the distinction between computational language generation and human communicative competence remains highly visible.

CONCLUSION

This study examined how contemporary artificial intelligence language models generate humour and irony within conversational discourse, with particular focus on ChatGPT and Google Bard. Using a corpus-assisted discourse-pragmatic framework, the research analysed 200 AI-generated humorous and sarcastic responses across multiple prompt categories in order to investigate the linguistic strategies, pragmatic mechanisms, and communicative limitations underlying AI-generated humour. The findings demonstrate that while current large language models are increasingly capable of reproducing recognisable humour structures, they continue to experience significant difficulty in managing the contextual, inferential, and sociocultural dimensions that characterise successful human humour.

One of the most important findings of the study concerns the distinction between structural humour replication and deeper communicative competence. Both ChatGPT and Bard displayed substantial ability to imitate familiar humour forms such as puns, sarcasm, exaggeration, anthropomorphism, and setup-punchline sequences. Lexical ambiguity and semantic incongruity emerged as dominant humour-generation strategies, suggesting that AI systems perform particularly well when humour relies upon highly patterned linguistic structures commonly represented within training datasets. These findings indicate that large language models can effectively reproduce the formal appearance of humour through statistical language modelling and probabilistic pattern recognition.

However, the study also demonstrates that structural fluency alone does not guarantee pragmatic effectiveness. Many responses that appeared superficially humorous lacked contextual coherence, emotional nuance, or interactional appropriateness when examined more closely through discourse-pragmatic analysis. Irony and sarcasm emerged as especially difficult categories because they require simultaneous negotiation of literal

meaning, implied meaning, contextual contradiction, and audience inference. Unlike lexical humour forms such as puns, irony depends heavily upon pragmatic inferencing and sociocultural awareness. Consequently, many AI-generated ironic responses appeared ambiguous, overly literal, or pragmatically incomplete. These findings reinforce the argument that humour is not simply a linguistic phenomenon but a socially negotiated communicative act shaped by context, interpersonal relationships, shared assumptions, and emotional positioning.

The taxonomy of humour failures developed in this study further highlights the communicative limitations of current AI systems. Recurring patterns such as inadvertent literalism, contextual misalignment, semantic incoherence, formal repetitiveness, and cultural decontextualisation reveal that AI humour generation remains strongly dependent upon textual pattern replication rather than socially grounded understanding. In many cases, successful interpretation of AI humour required human readers to infer missing contextual or emotional meaning independently. This finding suggests that AI-generated humour often functions through recognisable structural resemblance rather than through fully coherent communicative intention.

The comparative analysis between ChatGPT and Bard revealed additional differences in humour-generation tendencies. ChatGPT generally demonstrated stronger contextual continuity, greater lexical variety, and more nuanced sarcasm, whereas Bard relied more heavily upon simplified humour mechanics and direct contradiction. Nevertheless, both systems displayed similar underlying difficulties involving contextual inferencing, sociocultural sensitivity, and discourse-level coherence. These patterns indicate that increased linguistic sophistication does not necessarily resolve deeper pragmatic limitations within AI-mediated communication.

The inclusion of a small comparative human humour dataset also revealed important distinctions between human and AI-generated humour. Human humour demonstrated greater emotional subtlety, contextual flexibility, and interactional responsiveness than AI outputs. Human-generated humour frequently relied upon implicit assumptions, interpersonal positioning, and adaptive conversational framing rather than overt humour markers alone. AI-generated humour, by contrast, tended to foreground recognisable joke structures explicitly, reinforcing the perception that current language models remain more effective at reproducing textual humour patterns than at engaging in genuinely context-sensitive comedic interaction.

The study additionally contributes to broader discussions within computational pragmatics, discourse analysis, and AI-mediated communication by demonstrating that communicative competence extends beyond grammatical fluency and semantic coherence. Current large language models are highly effective at generating linguistically plausible discourse, yet they remain comparatively limited in managing the pragmatic and sociocultural complexities that underpin human interaction. Humour therefore represents one of the clearest domains in which the distinction between computational language production and human communicative intelligence remains visible. The findings support existing scholarship arguing that language cannot be fully separated from social experience, emotional awareness, contextual interpretation, and communicative intentionality.

Several practical and ethical implications also emerge from the study. As AI systems become increasingly integrated into education, customer service, entertainment, healthcare, and everyday communication, humour-capable conversational agents may influence user trust, emotional engagement, and interpersonal perception. While humour can strengthen interactional engagement and increase perceived conversational naturalness, failed humour may produce confusion, discomfort, offence, or emotional harm, particularly in sensitive communicative contexts. The findings therefore emphasise the importance of developing AI systems with stronger contextual modelling, sociocultural sensitivity, and pragmatic awareness rather than focusing solely upon linguistic fluency.

The study also raises broader concerns regarding anthropomorphism and human-AI relationships. Because humour is strongly associated with intelligence, personality, and emotional understanding, successful imitation of humour may encourage users to attribute human-like intentionality or empathy to AI systems that fundamentally operate through statistical prediction rather than genuine understanding. This creates the possibility of misplaced trust or exaggerated assumptions concerning AI communicative competence.

Consequently, ethical discussions surrounding conversational AI should consider not only accuracy and efficiency but also the social and psychological effects of human-like humour simulation.

Despite its contributions, the study acknowledges several limitations. The research focused exclusively on English-language humour and analysed outputs generated by only two AI systems. Humour interpretation also remains inherently subjective despite the use of systematic coding procedures and inter-coder reliability measures. Furthermore, AI-generated outputs may change over time due to updates in system architecture, training data, or platform design. The study therefore does not claim to offer universal conclusions regarding AI humour, but rather provides a discourse-pragmatic examination of how contemporary large language models currently construct humour and irony in conversational settings.

Future research could expand this work by incorporating multilingual humour analysis, cross-cultural pragmatic evaluation, real-time conversational interaction, or audience reception studies examining how users interpret and emotionally respond to AI-generated humour. Additional comparative work involving human participants, stand-up comedy corpora, or social media discourse may also provide deeper insight into the distinctions between computational humour generation and human comedic communication. Research involving multimodal humour, including visual memes, emojis, and performative digital humour, would further enrich understanding of humour within AI-mediated environments.

In conclusion, this study demonstrates that contemporary AI systems are increasingly capable of reproducing the structural appearance of humour while remaining substantially limited in handling the contextual, inferential, and sociocultural mechanisms that underpin successful human comedic interaction. Although large language models can imitate humour forms with remarkable fluency, the findings suggest that genuine pragmatic competence remains far more difficult to replicate computationally. Humour therefore continues to function as a revealing boundary between linguistic imitation and socially grounded communicative understanding within artificial intelligence discourse.

REFERENCES

1. Attardo, S. (2020a). Humor in interaction. Cambridge University Press.
2. Attardo, S. (2020b). Irony and sarcasm. In S. Attardo (Ed.), Humor 2.0: How the Internet changed funny forever (pp. 43–62). Springer.
3. Dynel, M. (2021). Irony, deception and humour: An interdisciplinary perspective. Journal of Pragmatics, 171, 59–72.
4. Fiske, A., Henningsen, P., & Buyx, A. (2021). Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in mental health care. Journal of Medical Ethics, 47(9), 614–615.
5. Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30(4), 681–694.
6. Gatt, A., & Paggio, P. (2021). Multimodal humour: Exploring the boundaries of computational creativity. Frontiers in Artificial Intelligence, 4, 703731.
7. Grice, H. P. (2021). Logic and conversation. In S. Davis (Ed.), Pragmatics: A reader (pp. 305–315). Routledge. (Original work published 1975)
8. Gros, L., Kaminski, M. E., & Umbrello, S. (2023). Algorithmic psychotherapists: The ethics of using conversational AI in mental health. AI and Ethics, 3, 115–132.
9. Hempelmann, C. F., & Ruch, W. (2021). Robots and humour: Does an artificial intelligence have a sense of humor? Humor: International Journal of Humor Research, 34(3), 307–331.
10. Holmes, J. (2019). Humour and the construction of identity. In J. Culpeper, M. Haugh, & D. Z. Kádár (Eds.), The Palgrave handbook of linguistic (im)politeness (pp. 359–382). Palgrave Macmillan.
11. Nijssen, D., & Heylen, K. (2021). How funny are GPT-3's jokes? An analysis of humor in language models. Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, 1–10.
12. Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 5370–5381.

13. Ritchie, G. (2020). The linguistic analysis of jokes. Routledge.
14. Shah, C., Bhatia, A., & Kaur, A. (2021). Automatic joke generation: Learning humor from examples. arXiv Preprint.
15. West, R., Djuric, N., & Poon, H. (2022). Evaluating AI humor: Human ratings and machine judgments. Findings of the Association for Computational Linguistics (ACL), 1641–1650.