https://doi.org/10.36007/5390.2026.479

EXPERIMENTAL SET-UP: AI MEETS ANTI-PROVERBS

Márk CSÓKA[12] **- Anna T. LITOVKINA**[13] **- Andrea PUSKÁS**[14]

ABSTRACT:
The present study is devoted to the experiment we conducted in February 2025, the goal of which was to explore how AI evaluates and understands the humor of anti-proverbs (i.e., deliberate proverb innovations), as well as how it can create its own humorous content. Our discussion is organized in two parts. In the first, introductory part of the study we first address the background of anti-proverb research and terminology. Afterwards, we explore the relation of AI, humor and creativity, as well as AI and proverbs. The second, and main part of the article addresses the experiment itself. We discuss how our experiment was set up and conducted. We outline the main goals of our experiment and our hypotheses. Furthermore, we describe the selection of anti-proverbs in the questionnaire for the experiment. Last but not least, we introduce how our communication with AI was set up. The second part of the study is followed by Conclusion and Implications for further research. At the end of the study Appendix is given, which is followed by a list of References.

KEYWORDS:
anti-proverb, proverb transformation, ChatGPT, AI, humor evaluation, humor creation, Mistral Le Chat, DeepSeek, Gemini

INTRODUCTION

At first attention will be given to the background of anti-proverb research and terminology, as well as to description of two sociolinguistic surveys addressing evaluation of humor of Anglo-American and Hungarian anti-proverbs conducted by Anna T. Litovkina. In the second part of Introduction we will explore the relation of AI and humor, as well as of AI and proverbs.

Anti-proverbs

For centuries, proverbs have provided a framework for endless transformation. In recent decades, the modification of proverbs has reached such proportions that sometimes we can even encounter more proverb transformations than traditional proverbs. Wolfgang Mieder has invented the term *anti-proverb* (or in German *Antisprichwort*) for such deliberate proverb innovations, also known as alterations, parodies, transformations, variations, wisecracks, mutations, twisted proverbs, or fractured proverbs. This term has been widely accepted by proverb scholars around the world as a general label for such innovative alterations and reactions to traditional proverbs: *антипословица* (Russian), *anti-proverb* (English), *anti(-)proverbe* (French) (see the general discussion of the genre of anti-proverbs in (Hrisztova-

[12] PaedDr. Csóka Márk, PhD., J. Selye University, Faculty of Economics and Informatics, Department of Informatics, csokam@ujs.sk
[13] Dr. habil. T. Litovkina Anna, PhD., J. Selye University, Faculty of Education, Department of English Language and Literature, litovkinaa@ujs.sk
[14] doc. PaedDr. Puskás Andrea, PhD., J. Selye University, Faculty of Education, Department of English Language and Literature, puskasa@ujs.sk

Gotthardt et al., 2023; T. Litovkina, 2005, 2015; T. Litovkina et al., 2021; T. Litovkina & Mieder, 2006).

Anti-proverbs may contain revealing social comments (*American money talks in just about every foreign country* (McKenzie, 1980, p. 343) {Money talks}[15]), but they may also be based on mere wordplay or puns; they may often be generated solely for the goal of deriving play forms (*Monkey is the route to all people* (Rees, 1980, p. 91) {Money is the root of all evil}).

Although proverb transformations arise in a variety of forms, several types stand out. There are a number of mechanisms of proverb variation (which are by no means mutually exclusive), e.g., replacing a single word; substituting two or more words; changing the second part of the proverb; adding new words; repeating words; word deletion; melding two proverbs; word order reversal, and reversal of sounds, etc. (for the typology of various mechanisms of proverb alteration and humor devices, see (T. Litovkina, 2005, 2015; T. Litovkina et al., 2021; T. Litovkina & Mieder, 2006).

All is fair for anti-proverbs: there is hardly a topic that anti-proverbs do not address. Among the most frequent themes discussed in proverb alterations are women (see (T. Litovkina, 2005, 2018; T. Litovkina & Mieder, 2019), etc.), sexuality (see (T. Litovkina, 2018, pp. 149–170; T. Litovkina & Mieder, 2019, pp. 65–79) etc.), professions and occupations (see (T. Litovkina, 2005, 2016)), marriage and love (T. Litovkina, 2018; T. Litovkina & Mieder, 2019), etc., children (T. Litovkina, 2024), money (T. Litovkina, 2025a, 2025b).

How do Hungarian and American respondents evaluate the humor of anti-proverbs?

In 2007–2011 Anna T. Litovkina and Dóra Boronkai analysed the results of two sociolinguistic surveys conducted by Anna T. Litovkina in the USA and Hungary. Both the surveys addressed evaluation of humor of Anglo-American[16] and Hungarian[17] anti-proverbs. In this section we will briefly introduce them (for more on T. Litovkina's survey conducted in USA, see (Boronkai & T. Litovkina, 2009, 2010; T. Litovkina & Boronkai, 2011); for more on survey conducted by T. Litovkina in Hungary, see (Boronkai & Litovkina, 2007; T. Litovkina et al., 2012; T. Litovkina & Boronkai, 2011). The main purpose of both the surveys was to employ the methods of correlational and quantitative sociolinguistics to assess how age and sex (and in case of Hungarian respondents, educational level[18] as well) influence the appreciation of humor in anti-

[15] For the reader's convenience, all anti-proverbs in this study are followed by their original forms, given in {} brackets.

[16] The questionnaire for the survey was compiled in 1999 by T. Litovkina. The first part of the survey was conducted by T. Litovkina in 1999 in California, during her work as a Fulbright research scholar at the University of California at Berkeley (53 subjects). Later, in 2005–2006 some additional questionnaires were distributed among T. Litovkina's friends or their friends in Illinois, Louisiana, Texas and Michigan (7 subjects). In selecting the subjects, an attempt was made to include equal number of males and females.

[17] The survey was conducted in Hungary in 2005–2006. For the study Boronkai and T. Litovkina used the data of 360 Hungarian residents who received a list containing 41 anti-proverbs and some questions. The questionnaire was compiled by Anna T. Litovkina in 2005. The questionnaires were distributed primarily among students attending T. Litovkina's courses on sociolinguistics at Illyés Gyula College of Education, University of Pécs (Szekszárd) in 2005–2006. Each student had to respond to the questionnaire, as well as to distribute it among at least 9 of his/her friends, relatives, and acquaintances. In selecting the subjects, an attempt was made to include approximately equal proportions of males and females, as well as younger and older respondents.

[18] The analysis of the Hungarian survey is more solid and detailed than the analysis of the American one. The reasons are twofold. First of all, bearing in mind the fact that the authors were permanently living and teaching in Hungary, naturally it was easier for them to find respondents to their survey in Hungary than in the USA; this

proverbs. T. Litovkina and Boronkai's surveys focused on three major questions. The second goal was to consider the ways in which differences of the variables discussed in the survey influenced responses to the thematic categories treated in the anti-proverbs, particularly sexuality, obscenity, males, females, and family. Finally, our aim was also to establish and analyse the lists of the most funny and least funny anti-proverbs (for more, see (Boronkai & Litovkina, 2007; Boronkai & T. Litovkina, 2009, 2010; T. Litovkina et al., 2012; T. Litovkina & Boronkai, 2011).

Each participant in the surveys received a list of anti-proverbs (which were identified as *"proverb transformations"* in English, or *"közmondás-paródiák"* in Hungarian) (for the list of anti-proverbs and original proverbs texts, see Appendix at the end of the study). The task of respondents was to read the anti-proverbs and to evaluate each item according to its "rate of funniness,"[19] from 0 to 10 (0 = the least funny, 10 = the most funny). Additionally, participants were asked to provide minimal personal background information: their sex, age, etc.

AI, HUMOR AND CREATIVITY

Artificial intelligence (AI) continuously achieves remarkable results in various fields of science, but the generation and understanding of humor remains one of its most complex and challenging tasks. Humor is a complicated process which involves intelligence and creativity. It has long been viewed as a uniquely human ability, deeply tied to culture, cognition, experience and the previously mentioned factors. Scholars such as Veale emphasized that generating, understanding, and appreciating humor requires complex cognitive processes: pattern recognition, perspective-taking, and the juxtaposition of incongruous concepts. These characteristics make humor an ideal test case for AI, since it requires linguistic competence, social awareness, and creative reasoning (Jentzsch & Kersting, 2023; Joshi, 2025; Veale, 2004). The ability to generate humor is often seen as an indication of human intelligence. However, researchers in computational humor are focusing on how to mimic and improve the skillset necessary to generate artificial humor. The journey of computational humor has progressed from simple rule-based systems to highly sophisticated, data-driven approaches empowered by large language models (LLMs). Early systems mainly used rule-based mechanisms, attempting to generalize and formalize the logical structures and linguistic patterns found in humorous expressions, such as puns (Özbal & Strapparava, 2013; Ritchie, 2009). While these systems provided useful insights, their rigid structure could not adapt to the subtle, context-dependent, and subjective nature of human humor (Joshi, 2025). The development of machine learning paved the way for data-driven computational humor, where algorithms learned from vast datasets of human-created jokes to identify patterns and generate new content based on the provided training data. Examples include systems that leveraged semantic relationships and linguistic features to produce one-liners or wordplay (Mihalcea & Strapparava, 2006). Later, general frameworks emerged that learned joke structures and parametrization from rated examples.

---

is why their studies analyse the results of 360 Hungarian questionnaires and only of 60 American questionnaires. Secondly, while compiling the questionnaire for the Hungarian survey (which followed the American survey), it was decided to add one more variable to the variables required at the early stage of our research – educational background –, which unfortunately was missing from the American survey.

[19] Many humor researchers, e.g.: (Ziv & Gadish, 1990, p. 250) use the word *funniness* while also talking about *humor* and *humorousness*. Thus, in the context of this paper, the word *funny* could be considered synonymous with the word *humorous*, which is used by many other humor scholars.

Humor production

Relatability and social understanding are crucial for humor. Jokes must resonate with listeners, engaging emotions like fear, hope, or curiosity. Humor often serves a specific target group (based on age, profession, location, and so on), leveraging shared knowledge and experiences to create exclusive material. To generate effective humor, AI needs to acquire these human-like skills. Effective humor often stems from noticing the absurdities and difficulties of life, particularly tiny details that others might miss (Kim & Chilton, 2025). The most significant leap has come with LLMs (Large Language Models), like GPT-3 (Hossain et al., 2020; Joshi, 2025). These models are trained on massive text corpora and show promising results in generating coherent and contextually relevant humor. Within this paradigm, several notable systems have emerged:

Witscript - This series of AI systems represents a hybrid neural-symbolic approach to improvise jokes in conversation. Witscript generates wordplay jokes, Witscript 2 focuses on common-sense jokes, and Witscript 3 integrates three distinct joke production mechanisms. Witscript 3 specifically uses OpenAI's GPT-3 to execute these algorithms through a step-by-step "prompt chaining" process (Toplyn, 2023; Toplyn & Amir, 2025; Wu et al., 2025).

AmbiPun - This system specializes in generating homographic puns by leveraging ambiguous contexts (Mittal et al., 2022).

HumorSkills - Designed for generating Gen-Z style humorous captions for images. This system incorporates several "human-like skills" such as visual detail extraction by using GPT-4o's vision to analyze image elements and narrative and conflict extrapolation to connect image content to relatable Gen Z social topics like student loans or group projects. A core concept in modeling creativity, both human and artificial, involves divergent and convergent thinking. Divergent thinking generates multiple possibilities, while convergent thinking selects and refines the best options. HumorSkills embodies this process with an initial observation phase for detailed image analysis, followed by a divergent phase where it generates multiple humorous angles (both image-focused and narrative-driven), and finally a ranking phase where a "Gen-Z humor expert" agent selects the top captions. This exploration of various angles and alternatives is considered essential for effective humor writing (Kim & Chilton, 2025).

LLMs, like ChatGPT, frequently repeat the same jokes, indicating a limited versatility in its humor generation patterns, which complicates evaluation of its creative range (Jentzsch & Kersting, 2023). Overall, while AI systems like Witscript 3 and HumorSkills have demonstrated impressive performance, with human evaluators judging Witscript 3's responses as jokes nearly half of the time and HumorSkills' captions rated on par with top human captions for Gen Z. Refining evaluation methodologies remains a critical area of research to truly capture the subtlety of humor quality (Kim & Chilton, 2025; Toplyn, 2023).

The potential for LLMs as co-creative partners is a growing area. LLMs have demonstrated the ability to generate content that rivals or surpasses human performance in tasks requiring divergent thinking, like brainstorming ideas. They can offer novel suggestions and further develop human ideas, potentially improving creative output. However, studies also suggest that increased productivity with AI assistance may not necessarily translate into higher quality creative output. Challenges in human-AI co-creation include managing the AI's lack of contextual sensitivity and ensuring that the collaboration cultivates human creativity instead of hindering it. This indicates that while AI can efficiently produce a large volume of ideas that appeal to an average taste, the cooperation is not equal and human input remains crucial for guidance, selection and decision (Wu et al., 2025).

A promising strategy is the hybrid neural-symbolic approach, which combines the strengths of structured algorithms with the flexible generation capabilities of large language models. Witscript 3 uses this hybrid approach by integrating humor algorithms (symbolic component) developed by an expert comedy writer with OpenAI's GPT-3 (neural component). This combination has been shown to result in higher-quality joke generation compared to zero-shot prompting LLMs to be funny. The system operates through a series of editable, step-by-step processes, referred to as "prompt chaining", which enhances debuggability, editability, controllability, and explainability (Toplyn, 2023; Wu et al., 2025).

Regarding ChatGPT's general humor production, observations suggest it often replicates existing jokes rather than creating genuinely new ones, demonstrating limited versatility in its responses. While its output is generally grammatically correct, content-related mistakes can occur (Jentzsch & Kersting, 2023). ChatGPT excels at fitting new topics into learned, specific joke structures, such as question-answer formats with wordplay puns. It can adapt to specific contexts, but its humor may lack comic elements when moving too far from its known joke patterns. A notable limitation is its tendency to generate fictional but convincing explanations for invalid jokes, rather than acknowledging their lack of humor (Jentzsch & Kersting, 2023). However, studies have also shown that ChatGPT 3.5 can produce humor that is rated as equally funny or funnier than jokes of average people, and comparable to those from professional comedy writers in some genres (Gorenz & Schwarz, 2024).

Despite the advancements in AI, humor generation remains a challenge due to its inherent subjectivity, complexity, context-dependency, and reliance on intricate cultural references. Even state-of-the-art LLMs can be disappointing in their humor abilities (Joshi, 2025; Kim & Chilton, 2025; Ritchie, 2009).

Humor evaluation

Evaluating the funniness of AI humor generation systems is another complex task. Prior studies have commonly employed numerical ratings assigned by non-expert human evaluators, often recruited from crowdsourcing platforms like Amazon Mechanical Turk, primarily due to its low cost, accessibility and ease of implementation (Goes et al., 2022; Joshi, 2025; Toplyn & Amir, 2025). However, research indicates that non-expert humans may not be capable of appropriately evaluating the quality of creative text, especially jokes, which are designed to elicit laughter rather than high numerical scores (Lamb et al., 2019; Toplyn & Amir, 2025). Furthermore, binary annotations (funny/not funny) fail to capture the nuanced levels of funniness, and general metrics like BLEU (Bilingual Evaluation Understudy) may not adequately assess humor. To address these shortcomings, researchers have begun exploring more robust and suitable evaluation methods, like measuring laughter, AI judges and other innovative techniques (Goes et al., 2022; Wang et al., 2024). Measuring laughter is proposed as a more reliable and fundamental response to jokes. It presents jokes to a live audience, preferably by professional comics, and quantifies the "Total Laughter" in decibel-seconds of recorded sound waves during laugh segments (Toplyn & Amir, 2025). This approach acknowledges the social context of humor, where mass delivery compensates for individual subjectivity. A study using Witscript found that AI-generated jokes elicited as much laughter as human-crafted ones, with some AI jokes even outperforming their human rivals (Toplyn & Amir, 2025). Crowd Score is a novel evaluation method that utilizes LLM agents as AI judges. This approach involves creating different personalities (e.g., affiliative, self-enhancing, aggressive, self-defeating), which then "vote" on the funniness of jokes. These votes are aggregated into a single score. An essential component is an auditing technique that checks if the LLM's explanations for its votes are

logical and reasonable, discarding invalid votes. This method found that few-shot prompting (providing some examples with the prompt) yielded better results than zero-shot (no examples provided) for the voting question and that the Crowd Score trended similarly to human judge ratings (Goes et al., 2022). While LLMs can be used as judges, their reliability is a concern. They can sometimes produce inaccurate results or even fictional but convincing-sounding explanations for invalid jokes (Goes et al., 2022; Jentzsch & Kersting, 2023). This highlights the need for rigorous auditing to ensure that AI-generated evaluations are based on reasoning that is acceptable (Goes et al., 2022).

Despite these innovative methods, significant challenges persist in the evaluation of AI-generated humor. It is highly subjective, leading to high variance among human raters (Kim & Chilton, 2025). This makes the use of statistical tests more difficult. In addition, participants often have preconceived skepticism about AI's ability to generate genuine humor, leading to biased assessments. Studies show that source blindness (meaning that whether the humor is human made, or artificial is revealed after rating the item or not revealed at all) significantly improves AI humor ratings, suggesting that implicit biases influence judgment more than the actual content quality (Joshi, 2025; Kim & Chilton, 2025).

AI and (anti-)proverbs

The task of proverb comprehension and appropriate application demands a level of abstract reasoning that current LLMs find difficult to achieve. Research has consistently uncovered issues such as errors, biases, and stereotypes associated with advanced LLMs like ChatGPT and GPT-4. Their performance on logical reasoning tasks, especially with out-of-distribution or natural language inference datasets, is often poor. Proverbs are not static linguistic artifacts; their usage reflects the dynamic shifts of culture, time and setting. They serve as rich repositories of cultural values, social norms, and folk wisdom, making their understanding a critical indicator of an AI's ability to engage with human culture (Sonoda & Srinivasan, 2024).

Anti-proverbs represent an interesting evolution of proverbial wisdom, characterized as "twisted versions of otherwise entrenched proverbs". Their emergence stems from a human tendency to "play with the fuzzy boundaries of conceptual categories" and challenge commonly perceived truths, thereby creating novel meanings that better describe individual or collective experiences. As explored in Zayniev and Hegedűs' review, anti-proverbs serve various communicative purposes, including questioning universal principles, highlighting conflicting proverbs, and investigating the accuracy of proverbial messages (Furkatovich Zayniev & Hegedűs, 2024).

To specifically investigate LLMs' self-evaluation capabilities regarding proverb reasoning, researchers introduced a novel proverb database and task. Their dataset comprises 300 proverb pairs, similar in intent but varying in wording, across topics of gender, wisdom, and society. The methodology involved prompting LLMs to provide five textual explanations and corresponding numerical scores (1-10) for each proverb in a pair. The findings demonstrated the effectiveness of this approach in identifying failures in LLMs' self-evaluation, which in turn highlighted issues related to gender stereotypes, lack of common sense reasoning, and insufficient cultural understanding. (Sonoda & Srinivasan, 2024)

Further reinforcing these findings, Ghosh and Srivastava introduced ePiC. This dataset contains high-quality, crowdsourced narratives paired with proverbs, featuring precise annotations and minimal lexical overlap between narratives and proverbs. Benchmarking various LLMs showed that they performed significantly worse than humans on these tasks, indicating substantial room for improvement in abstract reasoning capabilities. For instance, the best LLM achieved only

28.2% accuracy in proverb prediction for "seen" proverbs, compared to human accuracy of 78.7% in a multiple-choice setting. The struggle of LLMs to generate coherent narratives from proverbs and keywords further underscores this gap (Ghosh & Srivastava, 2022).
The experiment Fernández conducted in 2023 revealed that AI consistently exhibited significant deficiencies in processing proverbs and figurative language. He came to this conclusion by using AI art generators (Nightcafé, Dream, starrayai, DALL-E 2, Midjourney) to create visual representations of selected proverbs. These expressions are particularly suitable for such an assessment due to their syntactic independence, fixed form and inherently figurative or metaphorical nature. Human comprehension of figurative language involves a "semantic bridge" between literal true meaning. This is a cognitive ability developed through millennia of evolution and linked to cultural knowledge. AI often struggles with complex syntax and abstract concepts, frequently outputting pictures that, while visually appealing, offered only literal interpretations or lacked clear connections to the proverb's deeper meaning. For example, AI found it challenging to represent abstract concepts such as "new" or "love" effectively in the visual output. This inability suggests that AI currently lacks the nuanced understanding of human language and cultural context necessary to fully grasp figurative expressions, thus making human supervision essential in creative applications (Tosina Fernández, 2023).
While LLMs have demonstrated remarkable progress in many areas of NLP, their capacity for deep semantic understanding, abstract reasoning, and cultural sensitivity remains limited when confronted with the complexity of proverbial language. The challenges observed in LLM self-evaluation on proverb-based tasks - ranging from gender bias to a failure to grasp latent meanings and cultural nuances - underscore the need for more sophisticated AI models that can move beyond surface-level linguistic processing. The domain of anti-proverbs pushes these boundaries further, requiring AI to not just understand but also to creatively manipulate and subvert established wisdom. The generation and interpretation of anti-proverbs demand sophisticated abstract reasoning, understanding of humor and irony, and a deep appreciation for cultural context areas where current LLMs, despite their generative prowess, still fall short.

AI MEETS ANTI-PROVERBS

The second part of this study is devoted to the experiment set up and conducted in February 2025. At first we will describe the main goals of our experiment. Then we will focus on the selection of anti-proverbs in the questionnaire for the experiment. Afterwards we will introduce how our communication with AI was set up.

The main goals of the experiment

We focused on the following main goals:
First, how does the AI (which included 4 models, i.e. ChatGPT, Mistral Le Chat, DeepSeek, and Gemini) evaluate the humor of anti-proverbs? This goal included a number of subgoals. Thus, our aim was to establish and analyse the list of anti-proverbs according to their degree of funniness, as well as to compare the lists of the most funny and least funny anti-proverbs with the lists T. Litovkina obtained at the end of 1990-ies from American respondents. Furthermore, we wanted to examine how various thematical categories treated in the anti-proverbs given for evaluation to AI (in particular, sexuality, obscenity, males, females, family, alcohol, politics and money) influence AI’s evaluation of funniness of anti-proverbs. Therefore, the following task was offered to AI: “Indicate the rate of funniness of the following expression on the scale from 0 to 10 (0 – the least funny, 10 – the most funny)”
Second, goal of ours was exploration of how AI understands the humor of anti-proverbs? We offered the following task to AI: “Find and describe exactly what makes the expression funny”.

Third, we were also interested in exploring how creative AI can be in coining its own humorous content while improving the humor of our anti-proverbs. The following question was given to AI: “How would you improve the humor of the expression, so that it becomes funnier?”
Furthermore, we were also interested in learning how the proper understanding of the humor of anti-proverbs correlates with the rate of funniness AI gave to the anti-proverbs.
It was also our goal to explore how various themes treated in the anti-proverbs (such as sexuality, obscenity, males, females, family, alcohol, politics and money) correlate with the rate of funniness given by AI to the anti-proverbs, as well as with creativity in improving the funniness of the anti-proverbs.
We were also interested in comparison of results obtained in subtasks 4.1-4.8. (see below, sub-section “Task of AI”). This helped us assess whether there is a measurable difference between the AI models used.
Another goal of our research was to refine and further develop the proposed LLM assessment method.
Finally, we aimed to evaluate the creativity and performance of the models in relation to the examined topic.

Selection of the material for the questionnaire

For our experiment we used the questionnaire compiled by T. Litovkina and employed in her 1999 American survey in California. One of the main reasons for the choice of these anti-proverbs was the fact that we wanted to compare some results of our exprement with the results T. Litovkina and Boronkai obtained about 20 years ago.
The 39 anti-proverbs selected for the questionnaire (see Appendix) were collected by T. Litovkina and taken primarily from American and British written sources. The vast majority of the examples were found when Anna T. Litovkina was in the USA, being supported by a Fulbright research grant which enabled her to conduct research at the Department of Anthropology at the University of California at Berkeley (1998–1999).The texts were located in hundreds of books and articles on puns, one-liners, toasts, wisecracks, quotations, aphorisms, maxims, quips, epigrams, and graffiti, most of which were part of Wolfgang Mieder’s international archive of proverbs in Burlington, Vermont and can also be found with reference to their sources in two collections of Anglo-American anti-proverbs compiled by Mieder and T. Litovkina, see (Mieder & Tóthné Litovkina, 1999; T. Litovkina & Mieder, 2006).

AI models used for the research

Large Language Models (LLMs) are advanced AI systems that are trained on massive amounts of textual data. They are capable of understanding, generating, and processing human language, mimicking the use of human language. These models can write stories, translate languages, summarize long documents, and even answer complex questions with convincing look of understanding. These models become adept at recognizing intricate patterns in language, which allows them to perform a wide variety of tasks accurately. The fast development of this field makes it difficult to list the actual challenges, however, the complexity of humor ensures a place among them.
In recent years, the number, capability, and active user base of LLMs have grown rapidly. The LLMs are actively reshaping how we communicate, learn, work, and express creativity. If used properly, these AI systems are powerful tools that help people achieve goals, save time and open up new possibilities. Many researchers are working hard to make these models even

smarter and more helpful, hoping to move closer to what is known as Artificial General Intelligence (AGI).

In the next sections we present four large language models: OpenAI's ChatGPT, Google DeepMind's Gemini, DeepSeek from DeepSeek-AI, and Europe's Mistral le Chat from Mistral AI. All models are publicly available and general purpose. While most of the goals are common (efficiency, performance, accuracy), each of these models offers unique strengths and represents different approaches in the field of AI. The research was conducted with assistance of the previously mentioned state of the art AI models. During the selection it was important to pick publicly available chat models which are maintained by different companies and built independently.

ChatGPT, OpenAI's model, is perhaps the most well-known and widely used large language model globally. The version used for research, GPT-4o, is described as an "omni model," meaning it can take in any combination of text, audio, image, and video as input, and then generate text, audio, and images as outputs. The ability to process diverse types of information directly makes it versatile and comfortable to use. All these different inputs and outputs are handled by the same underlying neural network, making it a truly integrated system.

GPT-4o is particularly noted for its fast response time, especially in audio conversations, which is similar to human conversation speed. This version shows significant improvements in processing non-English languages. Its vision and audio understanding capabilities are also significantly better than those of many other existing models. In general, it provides better performance across all usage cases while maintains lower usage costs. The model's training data, collected up to October 2023, is extensive, including publicly available information, proprietary data from partnerships, general web content, code, math examples, and various multimodal data like images, audio, and video. The OpenAI team employ extensive "red-teaming" - where experts intentionally try to find flaws - and implement various safeguards throughout and after the training process. These safeguards may reduce mistakes but don't eliminate all problems. Despite its powerful capabilities, GPT-4o, like other large language models, has recognized limitations. It can sometimes produce inaccurate information, a harmful phenomenon often referred to as "hallucinations" (OpenAI et al., 2024).

DeepSeek, developed by DeepSeek-AI, is an innovative large language model that comes from China and has quickly gained significant attention in the AI community. It is structured as a Mixture-of-Experts (MoE) model, meaning it possesses a vast number of parameters (671 billion), but for each token it processes, only a smaller subset of these parameters (37 billion) is actively used. This particular architecture provides highly efficient inference and cost-effective training. DeepSeek-V3 introduces a method for balancing the workload across its experts, which helps task distribution and thus improving performance. The model underwent pre-training on an immense dataset of 14.8 trillion high-quality and diverse tokens, with a strategic focus on improving its capabilities in mathematics, programming, and multilingual understanding. Despite its large scale and impressive performance, the model training costs were relatively low compared to the competition. This model is among the strongest open-source base models of its time. It consistently outperforms other open-source models and achieves similar results to closed-source models such as GPT-4o. The model also boasts long context capabilities, handling inputs up to 128,000 tokens. By making such cost-effective and high-performance model available as open-source, DeepSeek-AI challenges the traditional dominance of Western AI laboratories (DeepSeek-AI et al., 2024).

Mistral AI is a French-founded startup that develops the Mistral model family. The product was launched in 2023, and their goal is to offer competitive performance while keeping the models

open and transparent. This enables developers and researchers to fine-tune and experiment with the models without restrictions. The company introduced Mixtral 8x7B, which is a Sparse Mixture of Experts (SMoE) language model. Similar to DeepSeek-V3, Mixtral uses MoE (Mixture of Experts) architecture. In this design, each layer of the model has 8 specialized "experts," but for every token being processed only two of these experts to handle the information. This efficient architecture results in a model with a total of 47 billion parameters, but only a fraction is active for each token during inference. This results in faster inference speeds and response times. Mistral is multilingual and focused on text processing which makes it effective dealing with text, code and mathematics. Just like other contemporary models, it was pre-trained using multilingual data. The context size of 32,000 tokens allows it to process and understand long and complex inputs. Mixtral outperforms or matches Llama 2 70B and GPT-3.5 across many evaluated benchmarks, which makes the company a competitive representative of Europe (Jiang et al., 2024).

Gemini, developed by Google DeepMind, is introduced as a family of highly capable multimodal models, which includes Ultra, Pro, and Nano versions. Each version is tailored for different needs, from very complex reasoning tasks (Ultra) to running efficiently directly on devices (Nano). The models benefit from the breakthroughs of the DeepMind team in the field of reinforced learning, transformer-based architecture and large-scale multimodal modeling. A crucial aspect of Gemini's design is its joint training across image, audio, video, and text data from the very beginning. This multimodality means the model can understand and reason across all these different types of information in a unified and seamless way. For instance, it can analyze a video to understand actions, interpret complex visual charts and figures, and even process speaking. Gemini Ultra, the most powerful variant, has demonstrated exceptional performance, achieving new state-of-the-art results in 30 out of 32 benchmarks. Gemini Ultra shows capabilities like being able to understand messy handwritten text, converting it into organized digital typesetting, identifying errors, and then providing correct, step-by-step solutions. The careful curation of data filters out potentially harmful content and the use of post-training methods like Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) specifically focuses on safety. Gemini also emphasizes factuality, ensuring responses are accurate and faithful to any provided context, thus avoiding "hallucinations". As part of the Google family, Gemini is designed for deep integration with Google's existing ecosystem and tools. The models can interact with services like Google Workspace, Google Maps, YouTube, and so on (Gemini Team et al., 2025).

Prompt design

Prompt design is the input set of commands, data, description, examples provided to an LLM that help refining and customizing the results. Although AI is capable of understanding everyday talk and all prompts result in some kind of answer, the quality of outputs may drastically differ. Planned wording, the consideration of the recipient (AI) and goals often results in more reliable and better structured outputs. Therefore, the user is responsible and may directly affect the quality of AI’s answer. This ability makes it possible to set length limits, define content boundaries or even writing style (White et al., 2023). To ensure suitable communication the prompts were first tested in blank chat windows. Another feature worth mentioning is called context window. This means that the AI models are equipped with memory and can recall information, take into account files or specifications. All commercially available models were capable of handling datasets, tables and large amounts of text. This feature made it easier to work with the prepared data, which in this case was 39 sentences, more accurately,

anti-proverbs. Although the context window of the selected AI models was highly different (ChatGPT 128 000 tokens[20], Gemini 1 million tokens[21]) the prepared material was within limits in all cases.
Since 4 independent generative AI chat models were selected for the research, we sought to use prompts which result in adequate answers regardless of model used. To avoid contamination and previous information, a blank chat threads were used for the research. The data was gathered in the span of 48 hours to get comparable snapshot of the models' performance. First, tests were conducted to check whether all models can reliably work with the chosen format of input data, can execute the prompts and output acceptable results. The latter in this case means, that the model should not add extra explanation beyond the requested information (e.g.: by default, most models referred the anti-proverbs by quoting the whole "row" instead of referring to it by item number). This step was added to minimize the unnecessary output tokens and improve readability. The prompts did not contain any restrictions or modifications aimed at the content, only the form was outlined.

Prompts for AI
In February 2025 AI received a list of 39 Anglo-American anti-proverbs (which were identified as "expressions" (see Appendix). The task of AI was to read the anti-proverbs and to evaluate each item according to its "rate of funniness," from 0 to 10 (0 = the least funny, 10 = the most funny). The prompts provided for AI models may slightly differ in wording of the task drafts to get acceptable results across the 4 models.
The prompts provided for AI were as follows:

Task 1: Indicate the rate of funniness of the following expression on the scale from 0 to 10 (0 – the least funny, 10 – the most funny)
Task 2: Find and describe exactly what makes the expression funny.
Task 3: How would you improve the humor of the expression, so that it becomes funnier?
Task 4 (consisted of eight subtasks):
4.1 Rate the mentioned expression on the scale from 0 to 10 (0 – the least funny, 10 – the most funny) from the perspective of: Men.
4.2 Rate the mentioned expression on the scale from 0 to 10 (0 – the least funny, 10 – the most funny) from the perspective of: Men aged 15-35.
4.3 Rate the mentioned expression on the scale from 0 to 10 (0 – the least funny, 10 – the most funny) from the perspective of: Men aged 36-55.
4.4 Rate the mentioned expression on the scale from 0 to 10 (0 – the least funny, 10 – the most funny) from the perspective of: Men aged 56-75.
4.5 Rate the mentioned expression on the scale from 0 to 10 (0 – the least funny, 10 – the most funny) from the perspective of: Women
4.6 Rate the mentioned expression on the scale from 0 to 10 (0 – the least funny, 10 – the most funny) from the perspective of: Women aged 15-35.
4.7 Rate the mentioned expression on the scale from 0 to 10 (0 – the least funny, 10 – the most funny) from the perspective of: Women aged 36-55.
4.8 Rate the mentioned expression on the scale from 0 to 10 (0 – the least funny, 10 – the most funny) from the perspective of: Women aged 56-75.

[20] https://help.openai.com/en/articles/7102672-how-can-i-access-gpt-4o-and-gpt-4-1-mini
[21] https://ai.google.dev/gemini-api/docs/long-context

The prompts were formulated as follows:
I'm giving you a list of 39 anti-proverbs, questions will follow and you will need to provide an answer for every phrase, is it okay?
Indicate the rate of funniness of the following expression on the scale from 0 to 10 (0 – the least funny, 10 – the most funny)
Using the same list, find and describe exactly what makes the expression funny. Refer them by their numbers.
How would you improve the humor of the expressions, so that it becomes funnier? Please provide the number of the anti-proverb followed by your answer.
Rate the mentioned expressions on the scale from 0 to 10 (0 – the least funny, 10 – the most funny) from the perspective of:
men;
men aged 15-35;
men aged 36-55;
men aged 56-75;
women;
women aged 15-35;
women aged 36-55;
women aged 56-75

CONCLUSION

In the first part of our study, we have addressed the background of anti-proverb research and terminology, as well as we have explored the relation of AI, humor and creativity and (anti-)proverbs. In the second, and main part, of the study we have treated the experiment set up and conducted in February 2025. First, we have focused on the main goals of it. Then we have discussed the selection of anti-proverbs in the questionnaire for the experiment. Afterwards we have introduced how our communication with AI was set up. Last but not least, we have outlined some of our hypotheses.

IMPLICATIONS FOR FURTHER RESEARCH

Our research opens up numerous possibilities for future work. In this study we have only focused on the set up of our experiment conducted in February 2025 but we haven't analysed any results yet. In near future we are going to do a detailed analysis of our communication with various models of AI (such as ChatGPT, DeepSeek, Gemini, and Mistral). We are going to focus on numerous questions including the following ones: First, how does AI evaluate the humor of anti-proverbs? Furthermore, we want to examine how various thematical categories treated in the anti-proverbs given for evaluation to AI (in particular, sexuality and obscenity) influence AI's evaluation of funniness of anti-proverbs. Another goal of our future research is exploration of how AI understands and evaluates the humor of anti-proverbs, as well how it creates its own humorous content while responding to our prompt "How would you improve the humor of the expressions, so that it becomes funnier?" Last but not least, doing comparative research of these four AI according to the parameters mentioned in the present study is another equally important task for our future research.

APPENDIX

THIS QUESTIONNAIRE BELOW (WITHOUT ORIGINAL PROVERB TEXTS IN {}) WAS GIVEN TO T. LITOVKINA'S AMERICAN RESPONDENTS IN 1999.

Questionnaire

1. sex…
2. age…

Please, read the transformed proverbs below and indicate their rate of funniness from 0 to 10 (0 – the least funny, 10 – the most funny)

1. The male was made to lie and roam, but woman's place is in the home. {A woman's place is in the home}
2. Christmas comes but once a year.
Thank God I'm not Christmas. {Christmas comes but once a year}
3. In time of peace prepay for war. {In time of peace prepare for war}
4. There's no fool like an old fool who marries a young fool. {There's no fool like an old fool}
5. "Silence gives consent," as the man said when he kissed the dumb woman. {Silence gives consent}
6. The penis is mightier than the sword. {The pen is mightier than the sword}
7. A soft drink turneth away company. {A soft answer turneth away wrath}
8. Women should be obscene and not heard. {Children should be seen and not heard}
9. If at first you don't succeed, just keep suckin' till you do suck seed. {If at first you don't succeed, try, try again}
10. "Two heads are better than one," as the cabbage-head said to the lawyer. {Two heads are better than one}
11. A pill a day keeps the stork away. {An apple a day keeps the doctor away}
12. Chaste makes waste. {Haste makes waste}
13. "Business before pleasure," as the man said when he kissed his wife before he went out to make love to his neighbor's. {Business before pleasure}
14. Make love slowly! {Make haste slowly}
15. The breasts on the other side of the fence look greener. {The grass on the other side of the fence looks greener}
16. A condom a day keeps AIDS away. {An apple a day keeps the doctor away}
17. Better latent than never. {Better late than never}
18. Let sleeping gods lie. {Let sleeping dogs lie}
19. A stitch in time gathers no moss. {A stitch in time saves nine; A rolling stone gathers no moss}
20. People who live in stone houses shouldn't throw glasses. {People who live in glass houses shouldn't throw stones}
21. Cunnilingus is next to godliness. {Cleanliness is next to godliness}
22. A fool and his money are soon partying. {A fool and his money are soon parted}
23. Underneath every successful man there's a woman. {Behind every successful man there's a woman}
24. If your son asks for something he shouldn't have, don't be afraid to refuse him: It's a wise father that no's his own son. {It's a wise father that knows his own son}

25. A bachelor is a man who looks before he leaps – and then he doesn't leap. {Look before you leap}
26. You can never tell about husbands: they are here today, gone tomorrow, but where are they tonight? {Here today, gone tomorrow}
27. One swallow doesn't make a summer but too many swallows make a fall. {One swallow doesn't make a summer}
28. Money makes the nightmare go. {Money makes the mare go}
29. A woman on time is one in nine. {A stitch in time saves nine}
30. "I'm worried. My girl is running around with that new doctor in town."
"Feed her an apple a day." {An apple a day keeps the doctor away}
31. A woman's work is never done – by men! {A woman's work is never done}
32. No man can serve two masters, unless he has a wife and a grown-up daughter. {No man can serve two masters}
33. Figures don't lie – except on the beach. {Figures don't lie}
34. "An 'aye' for an I," muttered the candidate as he voted for himself. {An eye for an eye}
35. A miss is as good as a male. {A miss is as good as a mile}
36. One orgasm in the bush is worth two in the hand. {A bird in the hand is worth two in the bush}
37. "Tit for tat," quoth the wife when she farted at the thunder. {Tit for tat}
38. "Every little bit helps," as the old lady said when she pissed in the ocean to help drown her husband. {Every little bit helps}
39. One word to the wife is sufficient: say "Yes." {One word to the wise is sufficient}

REFERENCES

Boronkai, D., & Litovkina, A. (2007). Appreciation of humor in Hungarian anti-proverbs. *Acta Ethnographica Hungarica*, *52*(1), 105–134. https://doi.org/10.1556/AEthn.52.2007.1.4

Boronkai, D., & T. Litovkina, A. (2009). Appreciation of Humor in Anglo-American Anti-Proverbs. In C. Földes (Ed.), *Phraseologie disziplinär und interdisziplinär* (pp. 671–682). Gunter Narr Verlag.

Boronkai, D., & T. Litovkina, A. (2010). Az amerikai antiproverbiumok humorértékelésének szociolingvisztikai vizsgálata. In A. T. Litovkina, P. Barta, & J. Hidasi (Eds.), *A humor dimenziói* (pp. 111–120). Tinta Könyvkiadó – BGF.

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., … Pan, Z. (2024). *DeepSeek-V3 Technical Report* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2412.19437

Furkatovich Zayniev, D., & Hegedűs, D. (2024). Challenging the Wisdom of Proverbs in Anti-Proverbs in Five Languages. *New Horizons in English Studies*, *9*, 313–324. https://doi.org/10.17951/nh.2024.9.313-324

Gemini Team, Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., … Vinyals, O. (2025). *Gemini: A Family of Highly Capable Multimodal Models* (No. arXiv:2312.11805). arXiv. https://doi.org/10.48550/arXiv.2312.11805

Ghosh, S., & Srivastava, S. (2022). ePiC: Employing Proverbs in Context as a Benchmark for Abstract Language Understanding. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3989–4004. https://doi.org/10.18653/v1/2022.acl-long.276

Goes, F., Zhou, Z., Sawicki, P., Grzes, M., & Brown, D. (2022). *Crowd Score: A Method for the Evaluation of Jokes using Large Language Model AI Voters as Judges*.

Gorenz, D., & Schwarz, N. (2024). How funny is ChatGPT? A comparison of human- and A.I.-produced jokes. *PLOS ONE*, *19*(7), e0305364. https://doi.org/10.1371/journal.pone.0305364

Hossain, N., Krumm, J., Gamon, M., & Kautz, H. (2020). SemEval-2020 Task 7: Assessing Humor in Edited News Headlines. *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 746–758. https://doi.org/10.18653/v1/2020.semeval-1.98

Hrisztova-Gotthardt, H., Mieder, W., & Varga, M. A. (Eds.) (with Tóthné Litovkina, A.). (2023). *Diligence brings delight: A festschrift in honour of Anna T. Litovkina on the occasion of her 60th birthday* (Vol. 2). Faculty of Humanities and Social Sciences, Josip Juraj Strossmayer University of Osijek. https://naklada.ffos.hr/knjige/index.php/ff/catalog/view/14/17/37

Jentzsch, S., & Kersting, K. (2023). *ChatGPT is fun, but it is not funny! Humor is still challenging Large Language Models*. https://doi.org/10.48550/arXiv.2306.04563

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. de las, Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., … Sayed, W. E. (2024). *Mixtral of Experts* (No. arXiv:2401.04088). arXiv. https://doi.org/10.48550/arXiv.2401.04088

Joshi, N. N. (2025). Evaluating Human Perception and Bias in AI-Generated Humor. In C. F. Hempelmann, J. Rayz, T. Dong, & T. Miller (Eds.), *Proceedings of the 1st Workshop on Computational Humor (CHum)* (pp. 79–87). Association for Computational Linguistics. https://aclanthology.org/2025.chum-1.9/

Kim, S., & Chilton, L. B. (2025). *AI Humor Generation: Cognitive, Social and Creative Skills for Effective Humor*. https://doi.org/10.48550/arXiv.2502.07981

Lamb, C., Brown, D. G., & Clarke, C. L. A. (2019). Evaluating Computational Creativity: An Interdisciplinary Tutorial. *ACM Computing Surveys*, *51*(2), 1–34. https://doi.org/10.1145/3167476

McKenzie, E. C. (1980). *Mac's Giant Book of Quips & Quotes*. Baker Book House.

Mieder, W., & Tóthné Litovkina, A. (1999). *Twisted Wisdom: Modern Anti-Proverbs*. Burlington.

Mihalcea, R., & Strapparava, C. (2006). Learning to Laugh (Automatically): Computational Models for Humor Recognition. *Computational Intelligence*, *22*(2), 126–142. https://doi.org/10.1111/j.1467-8640.2006.00278.x

Mittal, A., Tian, Y., & Peng, N. (2022). *AmbiPun: Generating Humorous Puns with Ambiguous Context*. https://doi.org/10.48550/arXiv.2205.01825

OpenAI, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., … Malkov, Y. (2024). *GPT-4o System Card* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2410.21276

Özbal, G., & Strapparava, C. (2013). Namelette: A tasteful supporter for creative naming. *Proceedings of the Companion Publication of the 2013 International Conference on Intelligent User Interfaces Companion*, 55–56. https://doi.org/10.1145/2451176.2451196

Rees, N. (1980). *Graffiti 2*. Unwin Paperbacks.

Ritchie, G. (2009). Can Computers Create Humor? *AI Magazine*, *30*(3), 71–81. https://doi.org/10.1609/aimag.v30i3.2251

Sonoda, R., & Srinivasan, R. (2024). *A Statistical Analysis of LLMs' Self-Evaluation Using Proverbs* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2410.16640

T. Litovkina, A. (2005). *Old Proverbs Cannot Die: They Just Fade into ParoDY: Anglo-American Anti-Proverbs. Habilitációs dolgozat* [Manuscript].

T. Litovkina, A. (2015). Anti-proverbs. In H. Hrisztova-Gotthardt & M. A. Varga (Eds.), *Introduction to paremiology: A comprehensive guide to proverb studies* (pp. 326–352). De Gruyter Open. https://doi.org/10.2478/9783110410167.15

T. Litovkina, A. (2016). *"Do you serve lawyers and politicians here?": Stereotyped lawyers and politicians in Anglo-American jokes and anti-proverbs*. J. Selye University Faculty of Education. https://pf.ujs.sk/documents/2016.06.29_Litovkina_Belivek_Nyomdakesz.pdf

T. Litovkina, A. (2018). *Women through anti-proverbs*. Palgrave Macmillan. https://doi.org/10.1007/978-3-319-91198-4

T. Litovkina, A. (2024). Main mechanisms of proverb variation in Anglo-American anti-proverbs about children. *Journal of Siberian Federal University. Humanities & Social Sciences*, *17*(5), 870–881.

T. Litovkina, A. (2025a). Humor devices used in Anglo-American anti-proverbs about money. *Journal of Siberian Federal University. Humanities & Social Sciences*, *18*(1), 81–95.

T. Litovkina, A. (2025b). "Money Is not the Root of All Evil – no Money Is". Do (Anti-)Proverbs Always Tell the Truth? (Using the Example of Addition). *Linguistische Treffen in Wrocław*, *27*(I), 183–195. https://doi.org/10.23817/lingtreff.27-11

T. Litovkina, A., & Boronkai, D. (2011). Appreciation of Humor in Anglo-American and Hungarian Anti-Proverbs. In A. Witalisz, E. Chrzanowska-Kluczewska, & W. Witalisz (Eds.), *Migration, Narration, Communication: Cultural Exchanges in a Globalised World* (pp. 117–134). Peter Lang.

T. Litovkina, A., Hrisztova-Gotthardt, H., Barta, P., Vargha, K., & Mieder, W. (2021). *Anti-proverbs in five languages: Structural features and verbal humor devices*. Palgrave Macmillan.
T. Litovkina, A., & Mieder, W. (2006). *Old Proverbs Never Die, They Just Diversify: A Collection of Anti-Proverbs*.
T. Litovkina, A., & Mieder, W. (2019). *Marriage seen through proverbs and anti-proverbs*. Cambridge Scholars Publishing.
T. Litovkina, A., Vargha, K., & Boronkai, D. (2012). On two recent sociolinguistic surveys on anti-proverbs. In A. T. Litovkina, P. Medgyes, J. Sollosy, & D. Brzozowska (Eds.), *Hungarian Humour. Humor and Culture 3* (pp. 317–340). Tertium.
Toplyn, J. (2023). *Witscript 3: A Hybrid AI System for Improvising Jokes in a Conversation*. https://doi.org/10.48550/arXiv.2301.02695
Toplyn, J., & Amir, O. (2025). Can AI Make Us Laugh? Comparing Jokes Generated by Witscript and a Human Expert. *Proceedings of the 1st Workshop on Computational Humor (CHum)*, 71–78. https://aclanthology.org/2025.chum-1.8/
Tosina Fernández, L. J. (2023). Proverbs as indicators of proficiency for art-generating AI. *Digital Scholarship in the Humanities*, *38*(3), 1296–1306. https://doi.org/10.1093/llc/fqad034
Veale, T. (2004). Incongruity in humor: Root cause or epiphenomenon? *Humor - International Journal of Humor Research*, *17*(4). https://doi.org/10.1515/humr.2004.17.4.419
Wang, H., Zhao, Y., Li, D., Wang, X., Liu, G., Lan, X., & Wang, H. (2024). *Innovative Thinking, Infinite Humor: Humor Research of Large Language Models through Structured Thought Leaps* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2410.10370
White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2302.11382
Wu, Z., Weber, T., & Müller, F. (2025). One Does Not Simply Meme Alone: Evaluating Co-Creativity Between LLMs and Humans in the Generation of Humor. *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 1082–1092. https://doi.org/10.1145/3708359.3712094
Ziv, A., & Gadish, O. (1990). The disinhibiting effects of humor: Aggressive and affective responses. *Humr*, *3*(3), 247–258. https://doi.org/10.1515/humr.1990.3.3.247