HumorSkills: Cognitive, Social and Creative Skills for AI Humor Generation

Sean Kim
Computer Science Department
Columbia University
New York, NY
ssk2245@columbia.edu

Lydia B. Chilton
Computer Science Department
Columbia University
New York, NY
chilton@cs.columbia.edu

Abstract

Humor is a social binding agent. It is an act of creativity that can provoke emotional reactions on a broad range of topics. Humor has long been an outstanding challenge in AI, probably because humans are complex, and humor requires our complex set of skills: cognitive reasoning, social understanding, a broad base of knowledge, creative thinking, and audience understanding. We explore whether giving AI such skills enables it to write humor. We target one audience: Gen Z humor fans. We ask people to rate image caption humor from three sources: 1) highly upvoted human captions, 2) basic LLMs, and 3) LLM captions with humor skills. We find that users like LLM captions with humor skills more than basic LLMs and almost on par with top-rated humor written by people. We discuss how giving AI human-like skills can help it generate communication that resonates with people.

Introduction

Producing humor is a difficult yet quintessential endeavor that underpins the everyday human experience. People use humor to connect with others, to impress others, to point out absurdities, to make light of a bad situation, and to recognize universal struggles - big and small. Despite massive amounts of training data and examples in large language models (LLMs), even state-of-the-art generative AI models are known for being disappointing in their humor abilities (Jentzsch and Kersting 2023). Humor is deeply complex as it depends on cognitive skills like observing details that others might overlook, creative skills such as diverging in multiple different ways (considering multiple viewpoints, multiple narratives, etc.) and social skills such as understanding your audience, understanding human relationship dynamics, and judging what is appropriate to make fun of.

We test whether AI-generated humor that uses social, cognitive and creative skills can come close to human performance. As a research context, we study the specific problem of generating humorous captions for images posted to Instagram (IG). This is a popular humor type within English-speaking Gen Z social media users that is widely made and widely appreciated by a well-defined audience. There is ground truth (in the form of upvotes) for what captions the audience finds funny. It is a fairly complex form of humor that uses both visual and language skills to generate, it applies to a wide variety of input images and requires implicit Gen Z cultural knowledge to generate.

Our method of generating humor with AI uses a divergent and convergent process, injecting it with three different human-like "skills" that current large language models (LLMs) may not emphasize. The input to the system is an image. The first phase of the LLM workflow is divergent thinking. The system starts with the **cognitive skill of observation**: an LLM examines the image and writes a detailed description of what it sees, giving the system a diverse set of potential joke targets. Next, the system generates multiple humorous angles from this description. Instead of having one pathway for creating humor, we use the **creative skill of divergent ideation**, generating jokes along two separate pathways that use different approaches. The first approach is simple: it creates humor strictly related to the image content. The second approach introduces the **social skill of narrative/conflict extrapolation**. Instead of relying solely on the information in the image to derive humor, it finds social conflicts that could be analogous to the image and uses those to drive the humor. Social conflicts could include arguments between a boss and their employees, romantic partners, political parties, parents/children, sibling rivalries, friend group drama, etc. These are all relatable social conflicts that are known to provide fodder for humor. Up to this point, the system will have generated 30 jokes. The workflow then enters a short convergent phase, where an LLM acting as a "Gen Z humor expert" rates the captions and selects the 5 best to return.

We discuss how a skill-based approach can be a useful technique for creative problem solving. For domains like humor where the goal is to create a rich, multidimensional artifact, it is likely that a multidimensional approach is necessary.

Background on Humor Theory

Humor is an intellectual challenge that has been studied by many great Western thinkers including Plato, Kant, and Freud (Morreall 2016). It is a multi-dimensional problem, but it clearly has both social and cognitive elements. Kant originated the incongruity theory of humor, which says that humor happens when we perceive a mismatch between what we expect and what actually occurs—a violation of our mental expectations that resolves in a surprising but sensible way. The act of perceiving an incongruity is a cognitive task

relating to logical reasoning. The benign violation theory of humor (McGraw and Warren 2010) builds on expectation violation by adding a social dimension. Typically the expectations that are violated are social norms—like somebody doing something crude, offensive, or inappropriate. However, not all violations are funny; we must have "psychological safety" from the violation expectation. A joke about someone else's country might be funny, but a joke about your country might be offensive. Thus, to be funny, the social norm being violated must be somewhat benign.

Theories of humor assert that humor has structure. At the most basic level, a joke has a setup and a punchline: the setup introduces an expectation, and the punchline delivers a violation. The semantic script theory of humor (Raskin 1985) adds more structure to this. It holds that a joke works by making two contradictory interpretations of the same text fit at once — the setup invites one reading, and the punchline reveals a second. Take Raskin's classic example: *" 'Is the doctor at home?' the patient whispered. 'No,' the doctor's young wife whispered back. 'Come right in.' "* At first, the reader assumes a sick-patient scene (script 1). After the punchline, the reader comes to a different interpretation of the setup: it's not a sick patient, it's someone having an affair (script 2). Both readings were available all along; the way the joke was told allowed the reader to detect the incongruity and discover the second meaning (Suls 1972).

In addition to structure, there are two levels of creativity needed for jokes: a *premise* and an *execution*. A premise is the core comedic idea—the angle, the unexpected connection, the thing the joke is "about." In script theory's terms, the premise is the collision of the two scripts; in incongruity terms, it is the violation itself. Execution is everything that makes that premise land: the phrasing, the brevity, the framing, the choice of words. A strong premise rendered in flat language falls flat, and flawless phrasing with no underlying premise is just noise. Both are needed. This division recurs across comedy traditions under different names. Improvisers distinguish the "game" of a scene—its core comedic premise—from the *heightening* that develops it (Besser, Roberts, and Walsh 2013). Joke writers separate a joke's underlying angle from the "maximizers" used to sharpen it (Toplyn 2014; Knospe 2014). Across these accounts, the premise supplies the core, but lands only through its execution. Writing a joke requires creative work at both levels.

Although there is no definitive formula for humor, many humorists have described a set of social principles for writing jokes. Key themes include:

Jokes should be relatable (Dean 2000; Holloway 2010; Vorhaus 1994; Kaplan 2013; Carter 2001), listeners have to relate and empathize with them in some way. Jokes typically engage human emotions - fear, hope, curiosity, cringing, and other heightened states that raise the stakes that get us to listen and relate to the material (Carter 2001). Often jokes are for an in-group (Hurley, Dennett, and Adams 2011) - using the shared knowledge and experiences of a group to create exclusive material which only that group relates to.

Jokes have details and observations. Observing the difficulties and absurdities of life is a good way to find relatable material (Kaplan 2013). But obvious absurdities tend not to be surprising, so jokes often come from observing details that others likely missed. (Carter 2001; Vorhaus 1994).

Jokes contain narratives that include a point of view. (Carter 2001) Like stories and narratives, jokes often use multiple points of view to see a situation in a new and surprising way. Understanding peoples' thoughts, actions, and behaviors helps fully "act out" (Carter 2001) the story. For example, *"Question: How do you get an Amish person to change a lightbulb? Answer: What's a lightbulb?"* The punchline takes the Amish person's point of view, which is unexpected. Sometimes the narrative can be based on a metaphor that helps the listener see something in a new way, like the many grad school memes comparing the negative "Reviewer #2" who rejected their paper to a movie villain (Weinberg 2017).

Jokes are creative and can be intentionally constructed Almost all humorists who write books about humor echo the design literature that divergent and convergent cognitive processes (Guilford 1950) are helpful processes in writing humor. Exploration is necessary to expand topics (Holloway 2010) (divergent thinking). Semantic Script theory can be used to structure jokes (Raskin and Hempelmann 2002). Many alternatives for punchlines should be explored to find the one that resonates the most (Dean 2000) (convergent thinking).

Overall, jokes require complex social understanding and logical construction. Methods of producing jokes are likely to need both skills intertwined.

Related Work

Computational Creativity and Humor

Computational humor is an outstanding challenge in AI. Even classifying whether or not something is funny is a computational challenge. This has been tried with some limited success by training deep learning classifiers on New Yorker Captions (Shahaf, Horvitz, and Mankoff 2015), and by "unfunning" jokes to create more training data (Horvitz et al. 2024). Wordplay is a common target for computational humor generation, and it relies on linguistic violations rather than social or cultural ones. Thus, many successful systems have produced computational word play (Binsted and Ritchie 1997; Kiddon and Brun 2011; Toplyn 2023c; He, Peng, and Liang 2019; Taylor and Mazlack 2004), but these techniques do tend to be limited to wordplay-based humor, and do not generalize more broadly.

LLMs have opened this as a possibility but humor continues to be a challenge for LLMs like ChatGPT (Jentzsch and Kersting 2023; Mirowski et al. 2024), even with fine tuning (Anonymous 2024). Recently, ChatGPT has been shown to be funnier than jokes written by Turkers (as rated by other Turkers), but not nearly as funny as professionally published humor (Gorenz and Schwarz 2024). Relatedly, in a study of live improvisational comedy (Winters and Van der Stockt 2025) found that chain-of-thought-prompted GPT-4 jokes were rated comparably to those of human comedians, with audiences only slightly favoring the human jokes. However,

the GPT-4 jokes were both selected and performed by the comedians, which undoubtedly added to their funniness.

A common technique for non-wordplay humor is finding and relating unexpected associations (Toplyn 2023a), using multiple humor strategies for associations (Toplyn 2023b). Adding multi-step reasoning can also be applied (Tikhonov and Shtykovskiy 2024), but is not sufficient. These attempts are largely in the right direction. Using associations, reasoning, and creative processes like divergent and convergent thinking make sense. But there are many more dimensions to humor. Considering more human-centric ideas like point of view, observation of details, social understanding, and understanding the audience is also essential for humor generation.

AI Generation Techniques

There are a number of existing techniques shown to improve LLMs' native abilities to generate text. LLMs' understanding of the goal can be improved by prompt engineering (Radford et al. 2019), fine-tuning (Radford et al. 2018), and few-shot prompting (Brown et al. 2020). LLMs can have more human-like production processes through prompt chaining (Wu, Terry, and Cai 2022) or by using a mixture of experts (Cai et al. 2024). LLMs' reasoning abilities can be increased by techniques like chain-of-thought prompting (Wei et al. 2024), thought experiments (Ma et al. 2023), and related approaches. LLMs can automatically close a feedback loop to improve outputs through reflection (Shinn et al. 2023) and self-evaluation (Brodsky 2024). Improving LLMs' theory of mind can also help with perspective taking (Strachan et al. 2024) and other social problems (Chen et al. 2025). When these techniques are used together as agents (Park et al. 2023), they can even combine these abilities to solve multi-faceted problems like software engineering (Yang et al. 2024) by taking different roles. All of these techniques are potentially relevant to get LLMs to use the right combination of social, logical, and creative skills necessary to make jokes.

System

HumorSkills is a system that takes an input image and outputs 5 image captions. The architecture has three key steps that mimic human skills needed for humor. *Visual Detail Extraction* is a step that describes the image in depth in order to make non-obvious observations about it. *Narrative and Conflict Extrapolation* is a step that finds narratives not in the image that could be related to it, to expand the topic of jokes to things that are not just in the image but also analogous to it. *Fine-tuning* the joke generator with examples of good Gen Z humor helps the jokes be more relatable to the target audience by using references, slang, topics, and insecurities that resonate with this group.

The system generates two types of captions: image-focused captions (which comment directly on the content in the image), and narrative-driven captions. Variety is important to humor. Humor relies on surprise, and jokes that are too similar start to become more predictable. Additionally, with an infinite set of input images with different subjects and situations, there are more strategies needed to find a humorous angle that fits the content.

AI Humor Generation Walkthrough

Figure 1 contains a visual diagram and example of intermediate outputs when generating captions for an image. We describe each phase and implementation in detail.

Most steps use GPT-4o for text and image processing such as describing the image, ideation, and narrative/conflict extrapolation. At the time of running the study, GPT-4o was generally the most capable and stable model. However, for one step we used a fine-tuned GPT-3.5. This is only for generating the captions in Gen Z language. The GPT-4o fine-tuning pipeline rejected our training data because it contained "sensitive content." The content is "edgy" in the way that humor often is, but passes Instagram's content policy. Our GPT-3.5 model was trained before such restrictions were put in place.

Visual Detail Extraction The first phase of the system's workflow involves the Visual Detail Extraction component, which utilizes GPT-4o's vision capabilities to analyze the input image. This system incorporates a prompt that asks for a detailed paragraph that explains the who/what/where of the image, distinguishing between identifying the subject of the image, the main action of the image if it exists, and the background elements of the image. This component is responsible for extracting key visual elements such as objects, human expressions, background settings, and any notable aspects that could serve as the foundation for humor.

For instance, in the demolition site example from the system diagram (figure 1), the system identifies a large industrial demolition excavator and a person with a hose spraying the demolition site.

Visual Humor Ideation On top of the visual detail extraction, the system ideates on possible humorous elements from the visual of the image. This incorporates an additional prompt using GPT-4o that intakes the image and asks it to identify and ideate on potential humorous visual elements in the image, whether they are directly humorous elements, such as funny facial expressions, or more analogous elements. For example, for the system diagram (figure 1) image, the system noted the visual contrast of the excavator and person, reminiscent of a David versus Goliath scenario, which provides a foundational metaphor for generating humorous captions.

Narrative and Conflict Extrapolation In this next step, the system generates a narrative and conflict framework by drawing upon common and relatable Gen Z experiences such as work, school, family, social interactions, relationships, and more. The system chains together the results of the previous steps, into a new prompt sent to GPT-4o. The prompt contains the visual details, the visual humor ideation, and a list of common Gen Z experiences, as well as the instruction to "generate narratives that reflect the essence of the image that is set within the framework of the Gen Z experience." This narrative generation adds depth to the humorous captions by applying relatable themes and conflicts

Figure 1: HumorSkills System Diagram. Given an image, the system first extracts visual details with a visual language model, then performs visual humor ideation to analyze the image and propose humorous angles. It then generates ten potential conflicts that could be used to extrapolate the image into a relatable experience. The system then generates humor with and without the narratives, for diversity. A separate instance of the LLM trained to rank Gen Z humor ranks all the captions and returns the top five.

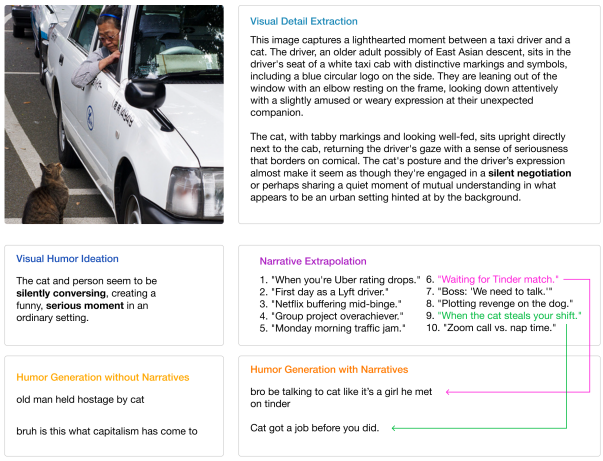


Figure 2: A diagram for how narrative and conflict extrapolation works

to the visual elements identified earlier. For example, in the system diagram 1, jokes like *"me mopping up my last relationship"* are clearly not about the image of a demolition site, it's an analogy to the metaphorical disaster left after a romantic relationship.

Figure 2 shows an example of how narrative/conflict extrapolation differs from more basic visual humor ideation. Both strategies start with a rich visual description of the image. Here, *"the image captures a lighthearted moment between a taxi driver and a cat...the cat's posture and the drivers expression make it seems as though they're engaged in a silent negotiation."*. Whereas the visual humor ideation without narratives lead to plausible jokes *"old man held hostage by cat"* they stay quite close to the image. In contrast, narrative extrapolation ideates multiple types of conflicts to generate jokes that are not about the image, but about social situations analogous to image. The two chosen by the LLM humor rater include: 1) a romantic conflict of waiting for a tinder match, *bro be talking to cat like it's a girl he met on tinder."* And 2) the workplace conflict of getting a shift stolen, *Cat got a job before you did."* Both successfully introduce a social conflict and relate it back to the image.

Humorous Caption Generation Following the narrative and conflict extrapolation, the system generates humorous captions using a fine-tuned GPT-3.5 model trained on Gen Z Instagram comments. It produces 15 image-focused captions (based on the visual description and humor ideation) and 15 narrative-driven captions (which additionally incorporate the narrative/conflict extrapolations), for a total of 30 candidates. The two pathways add variety: image-focused captions like *"bro out here getting paid $8 an hour to spray some water on some bricks"* stay close to the visual, while narrative-driven captions like *"The entitled bro you tried to make the group presentation with"* reach for analogous social situations.

Caption Ranking using Gen Z Agent The final component of the system architecture is the Caption Ranking and Filtering Agent, a GPT-4o-based agent fine-tuned to evaluate humor from a Gen Z perspective. This agent receives the list of 30 total captions from the narrative and visual humor-based caption generations and ranks the captions generated in the previous stage based on humor, relatability, and alignment with the image and narrative.

As illustrated in our system diagram, this agent ranks captions such as *"Me mopping up my last relationship"* and *"me pulling the emotional weight of the friend group"* based on their relevance to Gen Z humor. Captions that fail to meet the humor threshold are filtered out, such as *"Demolition worker really said 1v1 me bro,"* because although a phrase like "1v1 me bro" invokes Gen Z phrases, the content of the caption seems less relevant and relatable than a caption talking about school or relationships, ensuring that only the most effective and relatable captions are presented to the user.

Fine-tuning To make sure the LLM could generate content in the style of Gen Z, we fine-tuned a GPT-3.5 model on a dataset of 80 humorous Gen Z image captions. We first identified three popular Gen Z Instagram accounts that engaged in daily image-captioning contents (each account had over 400,000 followers). These IG accounts post images that are not inherently funny, then let the followers write potential captions for the image in the comment section. The most upvoted comments are typically the funniest captions for that image. We took the top 5 comments for 16 images (80 captions total) and used them to fine-tune GPT-3.5 to replicate their style.

Examples of the visual description of the images in addition to an explanation of potential humorous elements of the image were written in the fine-tuning prompts, then followed by the actual comment itself. This reflected the visual extraction and humor ideation being incorporated into the prompt of our current system.

Study 1: Humor Skills vs. GPT-4o vs. Top Humans

Methods

We ask people to rate humorous captions for images taken from 8 popular Instagram (IG) humor captioning accounts (separate from the training data accounts). For each image, users were shown 15 captions: 1) The top 5 most upvoted captions from IG 2) 5 captions from GPT-4o 3) 5 captions from HumorSkills

Data was collected from an online survey. Users would see an image and then rate 15 captions for it on a scale of 1 to 5, where 1 was "not funny", 3 was "somewhat funny", and 5 was "very funny". For each user, the images were presented in a random order, and the 15 captions for each image were also shown in a random order to avoid any possible ordering effects. Users were not told the condition of each caption.

The survey was distributed through email announcements to clubs and classes at a local university. Users were not told that AI was involved in the humor writing. The announcements advertised being for people who identify as "Gen Z" and who "like Instagram caption humor". These qualifications were added to ensure the raters were part of the target

Table 1: **HumorSkills vs. GPT-4o vs. Top Human Captions.** Comparison of user ratings for captions from HumorSkills, GPT-4o, and the top 5 Instagram comments, using a generalized linear mixed model (controlling for fixed effects of each caption). HumorSkills jokes were rated significantly funnier than GPT-4o jokes (.213 points higher on a scale of 1 to 5). HumorSkills jokes were as funny as the top Instagram posts (only 0.078 points less funny, which is not statistically significant).

Variable	Coef.	Std.Err.	z	P> |z|	[0.025	0.975]
Intercept (HumorSkill rating)	2.273	0.040	56.601	0.000	2.194	2.351
GPT-4o rating	**-0.213****	0.040	-5.285	**0.000**	-0.291	-0.134
Top5 IG rating	**0.078**	0.040	1.934	**0.053**	-0.001	0.157
Group Var	0.323	0.025				

*Note: ** denotes significance at the $p < 0.001$ level.*

audience, and most university students fit this demographic. The survey took about 30 minutes and users were paid $10.

We test the following hypotheses:

H1. HumorSkills captions will receive higher humor ratings than GPT-4o captions.

H2. HumorSkills captions will receive humor ratings comparable to top-rated Instagram captions.

Results

In total, 32 people responded to the survey: 5 male, 9 female, and 18 declined to say their gender. Fourteen of 32 respondents were aged 18–25. One selected "25+", and 17 declined to say their age group.

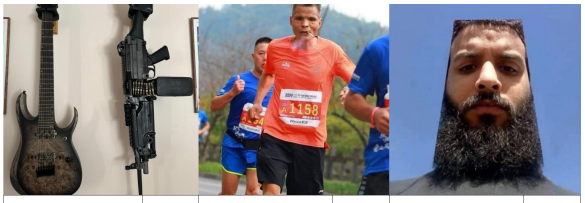
To test the relative humor score of the three conditions, we ran a Generalized Linear Mixed Model (GLMM) to estimate the funniness rating of each condition. The response types were ordinal numbers (ratings), the fixed effects were the caption type (IG, GPT-4o, HumorSkills), and the random effects were individual rater IDs and image IDs.

H1: HumorSkills vs. GPT-4o On the whole, captions from our system were rated as 2.27 (out of 5) for funniness. Captions written by GPT-4o were rated less funny (by 0.213 points), which is statistically significant at a $p < 0.001$ level. This indicates our system writes more humorous captions than a state-of-the-art VLM with prompt engineering. Results are shown in Table 1. Thus **H1 is supported. HumorSkills is rated funnier than GPT-4o.**

H2: HumorSkills vs. top IG captions On the whole, the top IG captions were rated only slightly better (0.078 points on a 5-point scale). But the difference was not statistically significant at the 5% level. However, it was very close ($p = 0.053$). This indicates that the system is competitive with the top IG captions. This effect could disappear with more data, but this dataset had over 6000 observations, and the 0.078 difference in average score is only 2% of the 1-5 scale. Thus **H2 is supported. Based on our sample, HumorSkills was rated as funny as the top 5 Instagram comments.** Or more precisely, HumorSkills was not statistically less funny than the top IG captions.

Qualitative Discussion

Whereas HumorSkills and Instagram comments tied for the highest rated captions, there are still qualitative differences between some of the jokes, and room for improvement. To analyze this, we identified the top-rated captions for each of the caption sources: Instagram, GPT-4o, and HumorSkills, and compared them to the highest-rated captions from the other two sources. We propose possible reasons for the top ratings, and ways to possibly improve the system. Figure 3 shows the images and the captions analyzed.



	Caption	Score	Caption	Score	Caption	Score
IG	American instruments	3.39	bro playing on max difficulty	3.14	Bro ready for his job interview at mojang	3.00
GPT-4o	Which one shreds harder? 🤘🤘	2.50	Running on fumes, literally	2.81	When you accidentally click "Crop" in real life	2.79
Humor Skills	me picking weapons in a video game	2.83	They really said "let's race to the hospital"	2.52	minecraft character looking ahh	3.62

Figure 3: Top rated image captions (marked in green) for Instagram, GPT-4o, and HumorSkills. (The top-scoring captions for the other sources for the same image are shown for comparison). From left to right, the figure includes three images: 1) The image with Instagram's top-rated caption (3.39/5). The image shows a guitar next to a machine gun. 2) The image with GPT-4o's top-rated caption (2.81/5). The image shows man running a race while smoking a cigarette and, 3) the image with HumorSkills' top-rated caption (3.62/5). The image shows a bearded man with his head cropped to a square.

Sharper social critique. The overall top-rated IG caption is "American instruments," for an image which juxtaposes an electric guitar and a machine gun. The joke critiques US gun culture. This caption is good because it's short but also a sharp social critique of a big problem. The HumorSkills caption for this image draws an analogy to video games, critiquing them for having silly weapon choices. It's relatable to Gen Z, but not nearly as sharp a critique as the top IG captions. Perhaps HumorSkills could be expanded to include more social critiques, from topics in the news. The GPT-4o caption is apt, but least funny: "Which one shreds harder?" It connects the two objects through unexpected word play

("shred"), but has no critique.

Learning the appropriate level of seriousness. The top-rated GPT-4o caption says *"running on fumes, literally"*, for an image with a runner smoking a cigarette. This is apt and connects the odd parts of the image (running and smoking), but it again uses wordplay. The IG caption was the highest rating for this image, *"bro playing on max difficulty."* It uses a video game metaphor to critique the runner smoking. The HumorSkills caption did least well for this image, it uses dark humor to critique the runner for smoking and point out the absurdity: *"They really said 'let's race to the hospital."'* Which is apt, but perhaps too serious and not benign enough for some readers. The system should do a better job of screening for jokes that are too "dark" or serious. The appropriate level of seriousness might have to be learned for each audience, or even individual audience member.

Mining more social references and in-group slang. The top-rated HumorSkills caption was the highest-rated caption in the entire study. It reads *"minecraft character looking ahh."* (The word "ahh" is Gen Z slang roughly meaning "ass".) Although this might not be broadly funny, it is funny to Gen Z because of this insider slang, because it's short, and because it is a critique of his looks related to Minecraft. The IG caption was rated less funny (*"bro ready for his job interview at mojang"*). It also relates the image to Minecraft (Mojang the developer of Minecraft). Gen Z grew up on Minecraft, and jokes about it resonate. The GPT-4o also makes fun of his hairstyle but relates it to cropping an image, rather than Minecraft, which is not as socially relevant. Overall, HumorSkills uses social references like Minecraft well, as well as in-group slang. Even more fine-tuning on Gen Z language could improve this ability. In the future, the fine-tuning doesn't necessarily need to come only from humor.

Study 2: HumorSkills performance on non-target images

The HumorSkills approach was designed and fine-tuned on a particular type of humor—Gen Z Instagram image captioning. It is possible that this approach is overfit to one type of image, especially since it uses fine-tuning to understand the types of images and captions this audience finds funny. Moreover, images in this dataset are not randomly selected—they are selected by people for being innately interesting or unusual. Other types of images that are not innately unusual could be harder to caption humorously.

To investigate how well the HumorSkills approach generalizes to other types of images, we compare how well HumorSkills and GPT-4o write captions for two types of images that are less likely to be inherently interesting.

1. **Camera roll images**. Camera roll images are the ordinary photos people take in their day-to-day lives. Typically they contain people, places, and things that are meaningful to the camera owner, but they are not generally interesting, noteworthy, meaningful, or funny.

2. **Museum Art**. Art is very different than photos. Although some of it depicts people doing things. Most of those things are historical, and not particularly relatable. Some paintings are abstract and do not concretely depict any discernible subject. Additionally, many of the art pieces are not paintings, but objects like vases or chairs, which could be challenging to make relatable and create a narrative for.

We randomly selected 30 images from the Flickr Image Dataset(Young et al. 2014) and 30 museum art images to test on our target audience. We take 15 images from the Museum of Modern Art (The MoMA) and 15 images of art from the Metropolitan Art Museum (The Met). For every image, we generate 5 HumorSkills captions and 5 GPT-4o captions. We create two separate surveys for rating the captions - one for art and one for camera roll images. As before, we randomize the order of the images and the captions to mitigate ordering effects. We recruited raters with the same process as the previous study.

We hypothesize that HumorSkills' captions will be rated more funny than GPT-4o captions for both new image sets. We did not compare to top-ranked human-written captions because these images were not taken from Instagram.

Results: Camera Roll Humor Ratings

As before, we ran a GLMM regression with image ID and user ID as random effects. There were a total of 4765 ratings. There were a total of 21 respondents: 10 female, 4 male, and 7 declined to say their gender. Twelve respondents were aged 18–25, 2 were aged "25+", and 7 declined to say their age group.

The results show that HumorSkills is 0.291 points higher than GPT-4o (2.19 vs. 1.9 out of 5). This difference is statistically significant with $p < 0.001$. Results are shown in Table 2.

Results: Museum Art Humor Ratings

There were a total of 4044 ratings. There were a total of 17 respondents, 7 female, 3 male, and 7 declined to say their gender. 8 respondents were aged 18-25, 2 were "25+", and 7 declined to say their age group.

The results show that HumorSkills is 0.18 points higher than GPT-4o (2.18 vs. 2.00 out of 5). This difference is statistically significant with $p < 0.001$. Results are shown in Table 3.

For both image sets, the image-level variance component was estimated at approximately zero, indicating that systematic differences between images explained negligible variance once caption-level fixed effects and rater random effects were accounted for; nearly all unexplained variance was attributable to individual rater differences.

This shows that HumorSkills can perform well on images that do not have obvious humorous qualities.

Qualitative Discussion

Benefits of Narrative/Conflict for Out-of-Domain Images
We expected the out-of-domain images to be harder to caption than the Instagram set, because Instagram humor accounts deliberately select for unusual or visually arresting subjects. Camera-roll photos, by contrast, are ordinary and offer no obvious joke target, and museum art is often abstract or historically distant from everyday life. With little inherent comedic material in the image itself, the system has to

Table 2: **Camera Roll GLMM Results.** Comparison of user ratings for captions generated by HumorSkills and GPT-4o on the Flickr image dataset of camera roll images. Using a generalized linear mixed model (controlling for fixed effects of each caption and random effect for individual differences in humor rating), HumorSkills captions were rated significantly funnier than GPT-4o captions (0.291 points higher on a scale of 1 to 5).

Variable	Coef.	Std.Err.	z	P> $\|z\|$	[0.025	0.975]
Intercept (GPT-4o rating)	1.901	0.054	35.357	0.000	1.796	2.006
HumorSkills rating	**0.291****	0.022	13.207	**0.000**	0.248	0.334
Group Var	0.000					
User_ID Var	1.260	0.116				

*Note: ** denotes significance at the $p < 0.001$ level.*

Table 3: **Museum Art GLMM Results.** Comparison of user ratings for captions generated by HumorSkills and GPT-4o on museum art images. Using a generalized linear mixed model (controlling for caption-level fixed effects and random effects for users), HumorSkills captions were rated significantly funnier than GPT-4o captions (0.175 points higher on a 1–5 scale).

Variable	Coef.	Std.Err.	z	P> $\|z\|$	[0.025	0.975]
Intercept (GPT-4o rating)	2.000	0.061	32.774	0.000	1.881	2.120
HumorSkill rating	**0.175****	0.023	7.677	**0.000**	0.130	0.219
Group Var	0.001					
User_ID Var	1.387	0.147				

*Note: ** denotes significance at the $p < 0.001$ level.*

	Caption	Score	Caption	Score	Caption	Score
Humor Skills	That's one way to deal with the **Space Force** budget cuts.	2.65	Medical staff having a discussion on whether it's **worth it to save the uninsured man**.	2.76	How your **charger** looks when you leave for 5 seconds.	2.76
GPT-4o	he really said "scope out the stars"	1.88	When the surgery team gets a crash course in the break room	2.53	When you attempt abstract art and accidentally create a web of confusion	1.64

Figure 4: Narrative Generation examples. Images across all 3 datasets with notable narrative generation. Instagram (left), Flickr (center), Museum Art (right)

	Caption	Score	Caption	Score	Caption	Score
Humor Skills	**Forehead so high**, it's getting its own zip code	2.39	dude looks like he's about to write a **manifesto**	2.87	bruh looks like they're waiting for a text back from the **shogun**	2.31
GPT-4o	That moment when you're caught off guard mid-snack, and now it's awkward	1.65	That "I came to the bar for solitude" vibe	2.13	That 'I'll start my homework in 5 minutes' energy.	2.15

Figure 5: Visual detail extraction examples. Images across all 3 datasets with notable visual extractions. Instagram (left), Flickr (center), Museum Art (right)

find a premise somewhere else. This is what the cognitive technique of narrative extrapolation does.

This shows up clearly in Figure 4. The first image, an unremarkable telescope, prompts a poorly-rated pun from GPT-4o; HumorSkills instead reaches for a topical political analogy: "That's one way to deal with the Space Force budget cuts." It use an external narrative with a rich/poor conflict to supply the humor the image itself doesn't. The second image, a break-room photo of doctors talking, again draws a pun from GPT-4o, while HumorSkills brings in a narrative of social conflict: *"discussing whether it's worth it to save the uninsured man."* The third, a fully abstract painting of intersecting colored lines, has nothing concrete to make fun of at all; HumorSkills reinterprets it through a relatable metaphor: *"how your charger looks when you leave for 5 seconds."* In each case, the joke comes from a narrative the image gestures toward but does not contain.

Benefits of Observation for Out-of-Domain Images A cognitive technique for finding jokes is observation: noticing details that others would miss. HumorSkills does this with an explicit visual detail extraction phase. GPT-4o is also a vision model and recognizes objects in images natively, but the explicit extraction step seems to make HumorSkills notice more, and notice differently.

Figure 5 shows examples from all three image datasets. In the first (from Instagram), the main subject has an unusually tall forehead; HumorSkills observes this and lands a well-rated joke about it, while GPT-4o reaches for a less interesting detail (eating snacks). In the second (from the Flickr camera-roll set), a man sits alone at a restaurant; HumorSkills picks up on his visual similarity to Hitler and

writes a sharp joke about composing a manifesto, while GPT-4o sees only that he "came for the solitude." In the third (from the Museum Art set), a Japanese woman in a kimono lies reading a book; HumorSkills registers the book, her forlorn expression, and the Japanese style, and writes *"bruh looks like they're waiting for a text back from the shogun."* GPT-4o's caption reads: *"That 'I'll start my homework in 5 minutes' energy"*. It is relatable but ignores the painting's visual cues entirely. In each case, the explicit observation step surfaces a detail the baseline either missed or underweighted, and that detail becomes the joke.

Discussion

Giving Human-like Skills to AI

This paper showed that for one form of humor — Gen Z style Instagram image captioning — HumorSkills was rated significantly funnier than GPT-4o's native humor. Additionally, HumorSkills was rated as funny as the top five most-upvoted human-written captions. This shows that while LLMs are not particularly funny out of the box, they can be made as funny as humans.

We attribute this to a set of explicitly added "skills," each addressing a different shortfall of native LLM humor generation. The cognitive skill of visual detail extraction found sharper, less obvious joke targets, closer to those of the human captions than GPT-4o's. The social skill of narrative and conflict extrapolation broadened the base of relatable targets, moving from the literal objects in the image to using them as metaphors for relatable situations such as relationship disasters, teamwork breakdowns, and the burden of student loans. The creative skill of divergent and convergent thinking allowed the model to expand the input into multiple joke premises and then select the best for the target audience, ranked by an LLM trained on Gen Z humor taste.

Each of these skills replicates a technique humorists have long described as useful: observing non-obvious details, finding relatable narratives and unexpected points of view, and treating joke-writing as a creative act of divergence and convergence. These are techniques current LLMs do not apply on their own. Asked to caption an image, a model like GPT-4o tends to name the most salient object and reach for wordplay; it does not, by default, scan for the non-obvious detail, search for an analogous social situation, or weigh what a specific audience will find funny. We show these skills can be added explicitly — and future foundation models may come to exhibit them inherently.

Humor as a Combination of Skills, Not a Single Trick

HumorSkills treats humor not as a single trick but as a combination of mechanisms working together. Recall from the background that humor operates on two levels: a premise — the core comedic idea — and an execution that makes the premise land. HumorSkills' skills span both. Visual detail extraction and narrative extrapolation expand the space of premises the system can draw on; narrative extrapolation also supplies material for heightening, such as exaggerated characters and emotionally charged situations. The fine-tuned Gen Z joke writer handles execution, rendering premises in the shorter, punchier language the audience rewards. No single one of these is the driver of a funny caption; the gains come from running them together to craft both premise and execution. This is why we are cautious about attributing the result to any single component: humor here looks less like a single capability than like several working together, mirroring the intertwining of logical construction and social understanding that humor theorists describe in human joke-writing.

Limitations and Future Work

This study targeted only one form of humor: Gen Z Instagram captions, which anchor the joke in an input image. Jokes can be anchored in many other ways: news headlines, mundane observations, or formats like stand-up that require generating both setup and punchline. Future work should explore what additional skills these forms demand.

The clearest near-term improvement is a stronger ranker. Many captions fail because they don't tie back to the image or because their logic never resolves into a coherent second script. A ranker that reasons explicitly about whether a joke's incongruity resolves could filter these out.

Most everyday humor is made within a friend group, leveraging insider knowledge — personality quirks, embarrassing moments, shared misery — to heighten relevance in benign ways. Future work could explore how to elicit such knowledge from a target group and personalize humor generation accordingly.

Our baseline used a simple GPT-4o prompt; more elaborate prompting might close some of the gap. Testing prompts with and without skill scaffolding would clarify how much of the gain comes from the skills themselves versus from prompting more carefully.

Finally, we did not measure joke originality. LLMs trained on internet-scale data risk producing jokes that are retellings of material in the training data (Jentzsch and Kersting 2023; Veale 2024). Because our system grounds each joke in the specific details of a given image and a social analogy chosen for that image, its outputs are more likely to be original to the input (Amir 2025) — though a rigorous measure remains future work.

Conclusion

Humor is a highly valued human ability. It has the power to teach, to entertain, and to connect people. Centuries of humor theory argue that humor is complex; it requires social, cognitive, and creative skills. In this paper, we study whether giving LLMs human-like skills for humor generation can improve their ability to write effective humor. Since humor is often about finding relatable connections between people, we focus on generating only one type of humor: Gen Z Instagram caption humor. We run two studies showing that LLMs with humor skills are rated as funnier than LLMs alone, and as funny as the top-rated human-written captions. This points to a future where LLMs can potentially write with similar social skills as people to produce emotional reactions and human bonding.

References

Amir, O. 2025. Are AI-generated jokes truly original? charting the "joke space". In *Proceedings of the International Conference on Computational Creativity (ICCC)*.

Anonymous. 2024. FunLMs: Methods for fine-tuning LLMs to generate humor. In *Submitted to ACL Rolling Review - June 2024*. under review.

Besser, M.; Roberts, I.; and Walsh, M. 2013. *The Upright Citizens Brigade Comedy Improvisation Manual*. New York, NY: Comedy Council of Nicea.

Binsted, K., and Ritchie, G. 1997. Computational humor. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, volume 97, 1089–1094.

Brodsky, S. 2024. Who watches the ai watchers? the challenge of self-evaluating ai. `https://www.ibm.com/think/news/ai-testing-advances`. Accessed: 2025-02-06.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Cai, W.; Jiang, J.; Wang, F.; Tang, J.; Kim, S.; and Huang, J. 2024. A survey on mixture of experts.

Carter, J. 2001. *The Comedy Bible: From Stand-up to Sitcom–The Comedy Writer's Ultimate "How To" Guide*. New York, New York, USA: Touchstone.

Chen, R.; Jiang, W.; Qin, C.; and Tan, C. 2025. Theory of mind in large language models: Assessment and enhancement.

Dean, G. 2000. *Step by Step to Stand-Up Comedy*. Portsmouth, New Hampshire: Heinemann Drama.

Gorenz, D., and Schwarz, N. 2024. How funny is chatgpt? a comparison of human- and a.i.-produced jokes. *PLOS ONE* 19(7):1–13.

Guilford, J. P. 1950. Creativity. *American Psychologist* 5(9):444–454.

He, H.; Peng, N.; and Liang, P. 2019. Pun generation with surprise. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1734–1744. Minneapolis, Minnesota: Association for Computational Linguistics.

Holloway, S. 2010. *The Serious Guide to Joke Writing: How To Say Something Funny About Anything*. Great Yarmouth, UK: Bookshaker.

Horvitz, Z.; Chen, J.; Aditya, R.; Srivastava, H.; West, R.; Yu, Z.; and McKeown, K. 2024. Getting serious about humor: Crafting humor datasets with unfunny large language models.

Hurley, M. M.; Dennett, D. C.; and Adams, Jr., R. B. 2011. *Inside Jokes: Using Humor to Reverse-Engineer the Mind*. Cambridge, MA, USA: The MIT Press.

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

Kaplan, S. 2013. *The Hidden Tools of Comedy: The Serious Business of Being Funny*. Studio City, CA: Michael Wiese Productions.

Kiddon, C., and Brun, Y. 2011. That's what she said: Double entendre identification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, 89–94. Stroudsburg, PA, USA: Association for Computational Linguistics.

Knospe, H. 2014. *Generating Humour: Analysing and Implementing a General Theory of Verbal Humour*. Doctoral dissertation, University of Leipzig.

Ma, X.; Mishra, S.; Beirami, A.; Beutel, A.; and Chen, J. 2023. Let's do a thought experiment: Using counterfactuals to improve moral reasoning.

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

Mirowski, P. W.; Love, J.; Mathewson, K.; and Mohamed, S. 2024. A robot walks into a bar: Can language models serve as creativity support tools for comedy? an evaluation of llms' humour alignment with comedians. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, 15. Rio de Janeiro, Brazil: ACM.

Morreall, J. 2016. Philosophy of humor. The Stanford Encyclopedia of Philosophy. `https://plato.stanford.edu/entries/humor/`.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Raskin, V., and Hempelmann, C. F. 2002. A computer model of joke generation based on script oppositions. In Behrendt, J., ed., *The April Fools' Day Workshop on Computational Humor*. Oldenburg, Germany: BIS-Verlag. 59–71.

Raskin, V. 1985. *Semantic Mechanisms of Humor*. Dordrecht, Netherlands: D. Reidel Publishing Company.

Shahaf, D.; Horvitz, E.; and Mankoff, R. 2015. Inside jokes: Identifying humorous cartoon captions. In *Proceedings of the 21th ACM SIGKDD International Conference on*

Knowledge Discovery and Data Mining, KDD '15, 1065–1074. New York, NY, USA: ACM.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems* 36. Publisher Copyright: © 2023 Neural information processing systems foundation. All rights reserved.; 37th Conference on Neural Information Processing Systems, NeurIPS 2023 ; Conference date: 10-12-2023 Through 16-12-2023.

Strachan, J.; Albergo, D.; Borghini, G.; et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour* 8:1285–1295.

Suls, J. M. 1972. A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. In Goldstein, J. H., and McGhee, P. E., eds., *The Psychology of Humor: Theoretical Perspectives and Empirical Issues*. New York: Academic Press. 81–100.

Taylor, J. M., and Mazlack, L. J. 2004. Computationally recognizing wordplay in jokes. In *In Proceedings of CogSci 2004*.

Tikhonov, A., and Shtykovskiy, P. 2024. Humor mechanics: Advancing humor generation with multistep reasoning.

Toplyn, J. 2014. *Comedy Writing for Late-night Tv: How to Write Monologue Jokes, Desk Pieces, Sketches, Parodies, Audience Pieces, Remotes, and Other Short-form Comedy*. Twenty Lane Media, LLC.

Toplyn, J. 2023a. Witscript 2: A system for generating improvised jokes without wordplay.

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

Toplyn, J. 2023c. Witscript: A system for generating improvised jokes in a conversation.

Veale, T. 2024. From symbolic caterpillars to stochastic butterflies: Case studies in re-implementing creative systems with LLMs. In *Proceedings of the 15th International Conference on Computational Creativity (ICCC)*, 236–244.

Vorhaus, J. 1994. *The Comic Toolbox: How to Be Funny Even If You're Not*. Los Angeles, CA: Silman James Press.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Red Hook, NY, USA: Curran Associates Inc.

Weinberg, J. 2017. How bad is reviewer 2, actually? data from a philosophy journal. Online; Daily Nous blog post. Accessed 20 May 2025.

Winters, T., and Van der Stockt, S. 2025. Evaluating humor generation in an improvisational comedy setting. *Computational Linguistics in the Netherlands Journal* 14:505–523.

Wu, T.; Terry, M.; and Cai, C. J. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. New York, NY, USA: Association for Computing Machinery.

Yang, J.; Jimenez, C. E.; Wettig, A.; Lieret, K.; Yao, S.; Narasimhan, K.; and Press, O. 2024. Swe-agent: Agent-computer interfaces enable automated software engineering. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 50528–50652. Curran Associates, Inc.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *Transactions of the Association for Computational Linguistics*, volume 2, 67–78.