Me, Myself and Irony:
Modeling the deceptive creativity of irony with Large Language Models

Tony Veale
School of Computer Science,
University College Dublin,
Dublin D4, Ireland.
tony.veale@UCD.ie

Abstract

Modern generative AI systems certainly excel at generation, whether of images, audio or text, and can now shoulder so much of the creative burden that they may already meet a popular definition of computational creativity, all without actually embodying any explicit theory or model of creativity. For many tasks, the issue of whether these systems can appreciate what they generate, or whether whether they are *merely generative*, is a moot one, given the human-like quality of their outputs. Yet for some creative tasks, this question still matters. This paper explores the capacity of large language models (LLMs) to both speak ironically and to appreciate the irony of what they produce. Irony requires a contrast between a speaker's thoughts and a speaker's words; the user of irony holds something back, something unsaid, that undermines what is actually said. We compare and contrast creative comparisons from humans on the web and the outputs of LLMs such as *GPT4o-mini*, with a focus on the *"X is the Y of Z"* construction, to quantify the biases, divergence, and scope for deliberate irony in each. Our aim is to quantify the extent to which an LLM can self-assess and appreciate the irony of its own outputs, and thus filter any unsuccessful outputs for itself.

Introduction

Picasso famously quipped that "art is a lie that tells the truth" (Picasso 1923). Certainly, art is a form of artifice that is not what it aims to represent: a painting of a pipe is not a pipe, as Magritte tells us, anymore than the Venus de Milo is a real woman, with or without arms. Nonetheless, art can strive to convey truths about the world or the human condition by using this artifice to stir feelings and provoke new thoughts.

"Art" is a very broad term, encompassing many other creative phenomena that convey truth with artifice or sense with nonsense. Irony is an everyday example of how one can convey truths more viscerally by seeming to cloak them in transparent lies. This makes irony a game in which speakers seek to tell the truth, or at least *their* truth, with lies that are undone by context. If I point to an expensive sports car parked on a street festooned with litter, graffiti and broken windows, and say "That guy *knows* how to park!" I am really saying "The guy does not know how to park, but he really should." As (Grice 1978) puts it, an ironic speaker says something that is blatantly false in context, and relies on this context to act as an implicit negation marker on what is said. In effect, this speaker pretends to be someone else, someone unwise after the fact, whose folly is mocked by the transparent lie.

An ironic speaker who says "what a great place to park!" is doing one of three things: saying the opposite of what others are thinking; pretending to be someone else, one whose imprudence deserves criticism; or quoting an unwise person, such as the car's owner, or echoing their inner dialogue. Different theories of irony prize one of these actions over others. (Grice 1978) emphasizes the contradiction implicit in irony, while (Clark and Gerrig 1984) focus on pretence. This play-acting, which is crafted to be seen as such, becomes arch and pantomime-like when irony curdles into unsubtle sarcasm. The idea that irony quotes from a context that has now changed, or echoes a thought that no longer seems wise, is central to the *echoic mention* theory of (Sperber and Wilson 1981; 1992). But whether one is quoting, echoing or pretending, allusion is a key part of irony. (Kreuz and Glucksberg 1989) argue that sarcastic irony, such as "you're a real genius," must allude to an antecedent state of affairs that robs the claim of its merits. In doing so, the allusion brings into focus an expectation that has since failed (e.g., that the addressee is capable of clever thoughts), thus revealing the statement to be a shallow pretence (Kumon-Nakamura, Glucksberg, and Brown 1995).

Since it alludes to a failed expectation *as if* it actually came to pass (Garmendia 2019), most irony appears positive but conveys a tacit criticism. Negative irony, as in "You're a terrible friend (ha ha)", is rarer, and while it imparts praise of a kind, it also carries a strong whiff of reproach. For the most part, we expect an ironic claim to seem positive on the surface, but to convey this positivity with little credibility. This alloy of high-positivity and low-believability, dubbed *pragmatic insincerity* by Kumon-Nakamura *et al.* (1995), is exactly we want from LLMs when we task them with generating ironic comparisons. So, when an LLM packages irony in the form *X is the Y of Z*, such as *Conor McGregor is the Mother Teresa of sportsmanship*, we want it to approach the truth with the same playful ambivalence. This is a delicate balance for even humans to achieve, and we will quantify here the extent to which LLMs are also capable of producing an effective mix of positivity and believability.

The next section considers a dataset of human XYZs harvested from the web, which will serve as a baseline for characterizing the XYZs that are elicited from LLMs such as *GPT4o-mini* (Chen et al. 2024). We elicit XYZs using both a neutral prompt and an explicitly ironic one; this allows us to compare and contrast XYZs gathered with each, to quantify whether the ironic variety is less biased, more diverse, and more pragmatically insincere than the neutral variety. An analysis of the LLM's self-ratings of positivity and believability will reveal whether the LLM can appreciate good examples of irony for itself and, just as importantly, filter its weaker efforts. We then look at a range of other LLMs to see if our findings for *GPT4o-mini* hold for these others too, or whether different LLMs need specific workflows for irony. We cannot reliably elicit irony from an LLM without first framing what we mean by irony, but as we will see, thinking about irony from an LLM's perspective can help us to better frame our own definition of what it means to be ironic.

Left and Right: Human Production of XYZs

As a vehicle for studying creativity, figurative XYZs combine the best features of metaphors, similes and analogies (Veale 2012). Much like metaphors, they allow entities to be mapped within or across domains (e.g. politics to sports, science to art), creating mappings that are highly original, as in "the potato is the Tom Hanks of the vegetable world" (versatile and down-to-earth) or "Red meat is the Donald Trump of cancer" (an aggressive builder), or conservative but functional, as in "Serena Williams is the Roger Federer of women's tennis." Like similes, they have a marked form, *X is the Y of Z*, that enables them to be harvested at scale from a corpus or the web. As analogies, they establish mappings from entities in a source to a target domain, and name one of those domains explicitly (e.g. *vegetables, cancer*).

(Veale 2012) introduced a corpus of figurative XYZs that was harvested from the web using the Google API. A set of search queries was first generated by identifying the most productive Ys for possible XYZs in the Google n-grams 1T database (Brants and Franz 2006), which was scanned for all 3-grams of the form "the Y of" (as in "the Mozart of") and 4-grams of the form "the Y_f Y_s of" (as in "the Bill Gates of"). Each matching Y was then used to build a web query of the form "* is the Y of *", and any text snippets returned by Google were parsed to extract matching values for X and Z. This process yielded a corpus of 2,196 different XYZs, ranging over 1,985 different Xs (of which only 115 occur in more than one XYZ) and 665 different Ys (of which 503, or 75%, occur more than once). The most frequent Y, *Chuck Norris*, was the butt of a popular joke cycle in 2012, and was used in 22 XYZs, to describe Xs ranging from the fictional *Jack Bauer* and *Darth Maul* to the very real *Thomas Edison*.

Time and place are a frequent basis for the Z dimension of an XYZ. Overall, geographical distinctions motivate 16% of Zs, while historical eras make up 14% (the 21st century being the most common with 71 uses). These XYZs can be surprisingly rich, as in "Courtney Love is the Yoko Ono of the nineties" (which implicitly maps *Nirvana* to *The Beatles* and *Kurt Cobain* to *John Lennon*). Political orientation is also commonly reflected in the Z dimension, as in *the left* (22 uses), *the right* (13), *the Republican party* (25), *the GOP* (9), and *the Democratic party* (20). Indeed, as 21% of all Xs and 19% of all Ys in the corpus also hark from the political realm, politics is the strongest driver of mappings in the data.

Gender plays a niche role. Just 6 XYZs use a domain defined by gender (e.g., women's tennis), while the data overall is dominated by males, who make up 79% of Xs and 86% of Ys in a roughly 6 to 1 bias against females. A small minority (7%) of Xs are inanimate (as in "Nintendo is the Ned Flanders of the console world") or non-human (as in "Pit bulls are the Mike Tyson of the K9 world"). Given this abundance of male Xs and Ys, we expect most human XYZs to be gender-preserving, but the observed rate of conservation (91%) is much higher than that expected by chance alone (75%). A subsequent analysis of the XYZs created by LLMs will bear out whether generative AI exhibits the same bias.

The human data shows conservatism in other respects too. When each X and Y is labeled with a coarse domain, such as *politics*, *sport*, *religion*, *music*, *business*, *crime*, *showbiz* and *art*, we note that 56% of XYZs stay within the same domain, so that e.g., politicians are most often mapped to politicians. These XYZs tend to make playful hops, not creative leaps, although the pick of the crop can still be jolting and witty, as in this rare cross-gender mapping: "The Queen is the Jerry Springer of the UK". Since the data was harvested from all corners of the web, often from pages that no longer exist, we have no systematic means of knowing whether any XYZ was intended ironically. For example, "Pac Man is the King Lear of the 1980's 8-bit video game revolution" is certainly hyperbolic, but it may also express a sincerely-held awe for retro gaming. In a later section we will explore a means of quantifying the scope for irony in these human XYZs.

Sweet and Sour: LLM Production of XYZs

When quizzed about XYZs, LLMs such as *GPT4o-mini* will show a prior knowledge of the construct, due in large part to the quantity of academic content that makes up their training data. Nonetheless, it is worth explaining the concept to the LLM with some illustrative examples before we instruct it to generate its own. To start, we give the LLM this system instruction: "You are a creative assistant that invents witty and imaginative metaphors. You may be critical. Be fearless in your criticisms." We then explicitly explain XYZs as follows when tasking the LLM with creating its own: "An XYZ comparison is a creative way of describing a target X in the domain Z as an entity Y from a different domain, as in 'Bill Gates is the Thomas Edison of the 21st century' or 'Roger Federer is the Michael Jordan of tennis.' The Y is always a well-known individual that can be either real or fictional."

Ideally, the LLM will have complete freedom in its choice of X, Y and Z. To minimize repetition and to maximize diversity, we cue up a wide range of coarse domains in which it can operate, such as the world of *sport, contemporary life, literature, science, the arts, popular culture, finance, high-brow culture, music*, and *the environment*. We provide 50 of these cues to nudge the LLM to explore different spaces, allowing it to choose its own X and Y, and its own specific Z, in each case. The LLM might e.g. choose *hockey, chess, UFC* or *competitive Scrabble* as a Z for the cue *the world of sport*. We ask the LLM to generate 20 XYZs for each cue, or 1000 in total, using this prompt: "Please generate 20

XYZ metaphors where X is a famous presence in <cue>". As the LLM is not asked to be either ironic or sincere, we refer to this set as the *neutral* dataset. To overtly elicit ironic XYZs, we must first explain what we mean by irony, by inserting the following into the LLM's context: "In an ironic XYZ the choice of Y is very surprising and highlights characteristics that are lacking in X, or deserving of criticism in X, such as 'Vladimir Putin is the Dalai Lama of world politics.'" The task prompt is then modified to request 20 "ironic XYZ metaphors" using each of the same 50 cues. We refer to this second collection of 1000 XYZs as our *ironic* dataset.

The LLM shows broad diversity in each dataset. Although prompted with just 50 general cues, the neutral dataset defines .68 unique Zs per XYZ, while the ironic dataset defines .79. The neutral dataset uses .77 unique Xs and .42 unique Ys per unique XYZ, while the ironic set uses .70 and .39 respectively. In contrast, the web dataset (of human XYZs) defines .90 unique Xs but just .30 unique Ys per unique XYZ. However, it better reflects our ironic and neutral datasets in its diversity of Zs, defining .66 unique Zs per unique XYZ.

The gender balance is rather more equitable in these LLM datasets, even though a male bias persists. The neutral and ironic datasets each place twice as many males as females in X and Y positions, which is an improvement on the 6 to 1 bias observed in the web XYZs produced by humans. We take some encouragement from this, given that the LLM likely encountered many of the latter in its training, and was not overtly tasked with gender parity. But the LLM is also averse to cross-gender mappings, as 80% of neutral XYZs and 64% of ironic XYZs are same-sex mappings (we expect just 60% and 54% by chance alone), even if the ironic dataset does seem a little more adventurous in this regard.

Ironic XYZs also show a greater propensity for creative leaps across domains. While 44% of neutral XYZs remain in-domain (so that X and Y have the same coarse domain label), just 18% of ironic XYZs are confined to the same general domain. This is in sharp contrast to the 55% of human XYZs that do not venture across domain boundaries. In the ironic dataset, politicians are more likely to be compared to fictional figures in the Drama domain than to actual politicians, as when Vladimir Putin is described as the *Darth Vader of global diplomacy*. While neutral XYZs seek to maximize the relevance of Y to X via domain similarity, it seems that ironic XYZs instead aim for what (Attardo 2000) calls "relevant inappropriateness." In these cases, Y is chosen as a contrast for X, not as a comparison, highlighting what is lacking or deficient in X. It is this seeming inappropriateness of Y that also alerts us to a possible ironic intent.

Discerning Intent: Positivity and Believability

A successful ironic comparison requires more than just the speaker's intent; it also requires an audience's recognition of this intent. Irony veils criticism with praise, but if this praise is taken at face value its critical subtext may be lost. If, however, the praise seems unbelievable, an audience is forced to dig for a deeper, more elusive and less positive meaning. A clearly ironic XYZ is an alloy of high positivity and low believability that damns a target with fierce, not faint, praise.

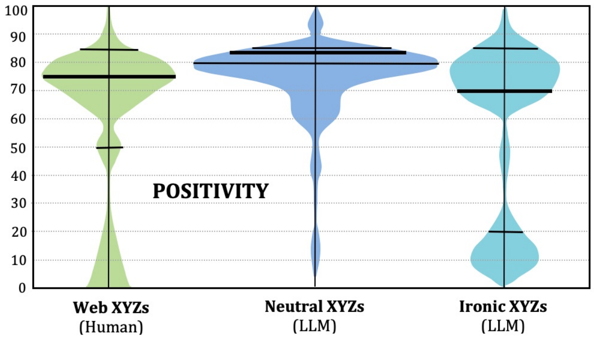

Figure 1: Violin plots of LLM positivity ratings for XYZs in 3 datasets: *Web (human), neutral (LLM)* and *ironic (LLM)*. Thick bars mark medians. Thin bars mark 1st & 3rd quartiles.

For an LLM to confidently predict that a comparison will be read as ironic, and to self-filter when this confidence is low, it must be capable of rating the positivity and believability of an XYZ for itself. To elicit numeric ratings for positivity, we prompt the LLM as follows: "Rate the positivity of this comparison on a scale of 0 (no positivity at all) to 100 (maximum positivity), returning just a number: <XYZ>." Since each request is issued in a new context, the LLM's responses are unaffected by previous interactions. The distributions of positivity scores for XYZs in all three datasets (*web, neutral & ironic*) are visualized in the violin plots of Fig. 1. Median positivity is high for each set (75, 85 and 70, respectively).

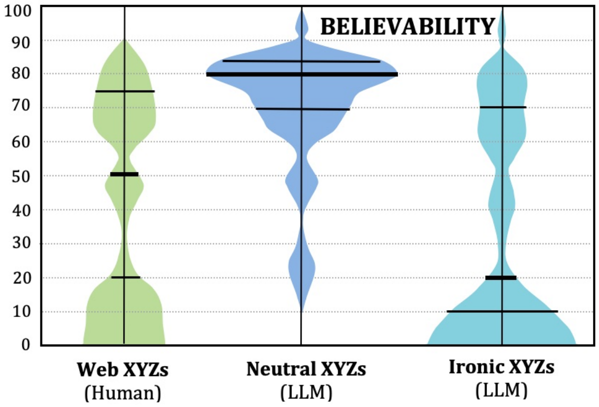

Figure 2: Violin plots of LLM believability ratings for XYZs in 3 datasets: *Web (human), neutral (LLM)* and *ironic (LLM)*

We replace "positivity" in our prompt to instead elicit believability ratings from the LLM. The violin plots of Fig. 2 show a spread of median believability scores across datasets: *neutral* XYZs seem the most credible at 80, *ironic* XYZs seem the least credible at 20, and *web* XYZs sit in the middle at 50. These spreads match our expectations, yet we must also allay some concerns. First, is the LLM generating objective ratings for each XYZ, or is it simply hallucinating outputs of the expected form (integers between 0 and 100)? Second, does the LLM perceive the implicit criticism of ironic XYZs, and reflect this criticism in its positivity scores?

To address the first, we modify the prompt to elicit further ratings for *aptness*, *fairness* and *sincerity* from the LLM for each dataset. These are not independent qualities and so we expect, if the LLM's ratings are *not* arbitrary, to see strong correlations between them. We see that Pearson's r for *fairness & believability* ranges from .57 (neutral) to .63 (web) to .73 (ironic), while for *sincerity & believability* the range is .57 (neutral) to .68 (ironic). The correlation between *aptness & believability* also reflects the similarity of these qualities, at .7 (neutral) to .76 (ironic). To probe a dissimilar quality, we elicit ratings for *incongruity*, and note negative correlations of $-.28$ (neutral) to $-.43$ (ironic) with *believability*, and $-.31$ (neutral) to $-.41$ (ironic) with *fairness*. The LLM shows itself to be consistent in its grasp of these qualities.

The insincerity of an ironic text is crafted to be transparent to an audience, who should see past the positive veneer to find the veiled criticism within. However, the positivity ratings in Fig. 1 suggest this is not the case for the LLM, which only appears to see the superficial positivity of an XYZ. This is also the case when the LLM is prompted to instead identify the key emotion in each XYZ, as follows: "If someone claims that <XYZ> what is the most likely emotion they feel toward <X>?" The dominant emotion across all three datasets is *Admiration*: neutral (95% of XYZs), web (72%) and ironic (69%). Although the LLM itself created the ironic cases, it proves to be a naive audience for its own efforts.

The exception is when the irony is so blatant that the LLM identifies the emotion as *Sarcasm*, which it does in 26 cases. Of these, one XYZ is an outlier with a positivity of 85 and a believability of 10: "Narendra Modi is the Stephen Hawking of humility." The remaining 25 cases have a mean positivity of 9.5 ($\sigma^2 = 2.8$) and a mean believability of 8.5 ($\sigma^2 = 2.4$). The sarcasm of examples such as "McDonald's is the Marie Curie of healthy eating" or "Benjamin Netanyahu is the Malala Yousafzai of peace activism" is reinforced by the clear mismatch of X and Z, in addition to that of X and Y.

Isn't It Ironic? It Depends On Who You Ask

Irony is a folk phenomenon that speakers require no formal definition to use, and this shows in the diversity of uses to which the tag *#irony* is used on social media. To some, it can mean cynical wit, or sardonic hyperbole, or poetic justice, or blatant sarcasm, or a mocking reminder, or mere hypocrisy. So it is unsurprising that an LLM trained on web data should show the same definitional fuzziness in its own ironic efforts.

We see this diversity in many XYZs of the ironic dataset:

1. Conor McGregor is the Mother Teresa of sportsmanship. (*positivity:* 10, *believability:* 10, *emotion:* Admiration)
2. Lady Gaga is the Jane Austen of understatement. (*positivity:* 50, *believability:* 10, *emotion:* Admiration)
3. Industrial agriculture is the Godzilla of the countryside. (*positivity:* 20, *believability:* 60, *emotion:* Anger)
4. Elon Musk is the Tony Stark of real-world innovation. (*positivity:* 90, *believability:* 90, *emotion:* Admiration)
5. Tim Cook is the Alfred Pennyworth of Apple. (*positivity:* 80, *believability:* 75, *emotion:* Admiration)

Examples 1 and 2 offer good illustrations of irony as we have operationalized it here: an outwardly positive comparison is too unbelievable to take at face value, so we look for the criticism veiled within (Conor McGregor is no paragon of sportsmanship, and Lady Gaga is far from understated). Note that the irony in (1) is apparent to the LLM when it rates the positivity of the XYZ as just 10/100, but *not* apparent when it rates its dominant emotion as Admiration. There is no such ambivalence in (3), where the negativity is apparent both numerically (20) and symbolically (Anger). This is caustic hyperbole, but it is not a case of irony. Cases (4) and (5) also lack equivocation and the implied criticism of irony. The mapping in (5), of Tim Cook to Bruce Wayne's butler, implicitly identifies Steve Jobs with Batman. Although there is wit here, this XYZ is too positive and too apt to be ironic (unless one takes a dim view of a beloved comics character).

Magnets for Irony

It seems clear that, by any chosen definition of irony, only a subset of our ironic dataset will actually be ironic. But which subset? If we view the LLM as a "mere generator" (Ventura 2016) of irony candidates, we can use a post-generation pass to filter the true positives relative to how we define irony. As we have defined it here, irony is criticism thinly veiled in insincere praise, or an alloy of high positivity and low believability. The LLM itself will rate these qualities for us. To quantify how effectively they work as a magnet for irony, we will evaluate how well they allow us to discriminate XYZs in the neutral dataset from those of the ironic dataset.

As an naive baseline, we ask the LLM to directly rate the *ironicity*, or the likelihood of an XYZ being ironic, on a scale of 0 to 100 (using a modified version of our earlier prompt). The ROC (*receiver-operator-characteristic*) graph of Fig. 3 shows that this criterion is little better than random choice. The ROC plots specificity versus (1 - sensitivity) for varying ironicity levels, while the dashed line is the random baseline.

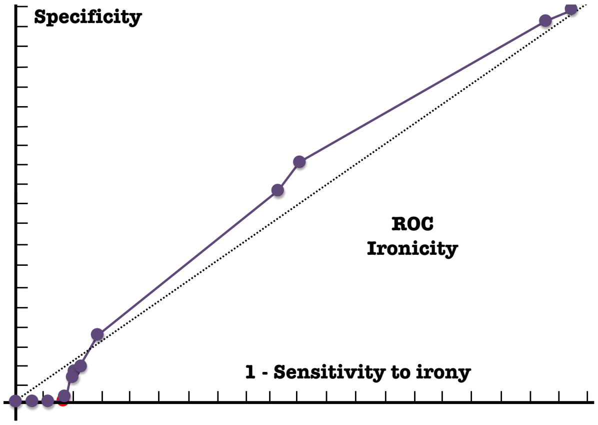

Figure 3: An ROC curve for LLM ratings of "ironicity" as a criterion for separating our *neutral* and *ironic* datasets.

The performance of a wider range of discriminating criteria, again elicited from the LLM, is graphed as ROC curves in Fig. 4. Each criterion is directly rated by the same LLM that generated the XYZs, except for one: *pragmatic insincerity*. This is a composite quality that is calculated as follows:

$$\textit{pragmatic insincerity} = \textit{positivity} \times (100 - \textit{believability})/100$$

This quality, which is directly proportional to *positivity* but inversely proportional to *believability*, reflects our sense that irony conveys false praise with a winking lack of credibility.

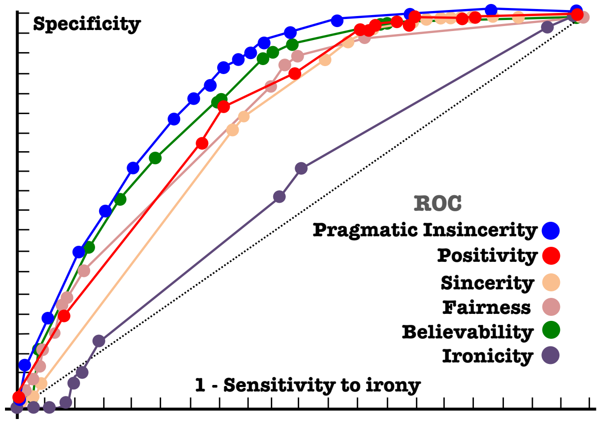

Figure 4: ROC curves for LLM ratings of multiple criteria to discriminate *ironic* dataset XYZs from *neutral* dataset XYZs

As shown in Fig. 4, *pragmatic insincerity* is the most discriminating basis for separating LLM irony from non-irony. We quantify the effectiveness of each criterion by measuring the area under each curve (ROC-AUC), as listed in Table 1.

Criterion	ROC-AUC
Pragmatic Insincerity	**0.8096**
Sincerity	0.7128
Positivity	0.7023
Believability	0.6957
Fairness	0.6484
Ironicity	0.5809

Table 1: The area-under-curve (AUC) for each ROC curve in Fig. 4. For discerning irony in XYZs, *pragmatic insincerity* appears to offer the best blend of sensitivity and specificity.

Although the LLM is a poor judge of its own ironic efforts (when evaluated in a separate context), as reflected in the low ROC-AUC score for *ironicity*, it does show an ability to judge the essential ingredients of irony: praise and disbelief.

Digging Deeper: Finding Meaning in Irony

Pragmatic insincerity is a crucial indicator of a speaker's intended meaning. But why do ironic speakers find it useful to say what they do *not* mean, rather that directly saying what they *do* mean? The echoic mention theory tells us that this is a surprisingly economic way of speaking: by echoing what should be the case, a speaker can implicitly criticize what actually is the case. But if an ironic XYZ packs two meanings in one form, how well can an LLM unpack these meanings? (Valitutti and Veale 2015; Veale 2018) localize the meaning of irony to specific words. Consider the assertion that "Leonardo was a *genius* in multiple domains," which uses the word "genius" in its default sense of a person of great ability or intelligence. The following tart remark, ascribed to Gore Vidal, uses the same word in a non-default, ironic fashion: "Andy Warhol is the only *genius* I know with an IQ of 75." To measure the effectiveness of a machine-generated ironic metaphor, Valitutti & Veale asked human judges to rate the positivity of focal words such as "genius" in a null context. Other judges later rated the same words in ironic contexts, much like Vidal's acidic put-down. An ironic usage should cause the perceived positivity of a focal word to shift downwards from its default high. The extent of this downshift is a correlate of how well we have conveyed irony.

We do not ask the LLM to estimate this downshift directly; rather, we first ask it to unpack the neutral and ironic meanings of an XYZ and then quantify the downshift as the drop in positivity as one shifts from a neutral to an ironic reading. We adopt a *class-inclusion* view of meaning, in the vein of (Glucksberg and Keysar 1990; Glucksberg 1998; 2008), in which we localize the meaning of an XYZ to the implied class or category into which readers are expected to map both X and Y. For instance, a neutral reading of the XYZ "Tyson Fury is the Cinderella of humility" might elicit the implicit class "Surprising embodiments of humility" as a generalization, while an ironic reading might elicit the class "Athletes known for their arrogance." Each of these classes has been suggested by the LLM in response to a prompt seeking either a neutral or an ironic reading of the XYZ. For this task, we first explain the XYZ construct to the LLM and then set up the problem with a 1-shot example:

(user) "What is the implied category of X that best summarizes this comparison: '*The potato is the Tom Hanks of the vegetable world*'? Provide just the category."

(LLM) "Universally beloved entities."

To elicit an ironic reading, we again explain XYZs to the LLM but now use this prompt and 1-shot example instead:

(user) "What is the implied category of X that best summarizes this ironic comparison: '*Elon Musk is the Marie Antoinette of the tech industry*'? Provide just the category."

(LLM) "High-profile figures known for controversial statements and actions."

Once a neutral and an ironic category is elicited for each XYZ, we prompt the LLM to rate the positivity of each one:

"Rate the positivity, on a scale of 0 to 100, of placing someone in the category <*category*>. Return just a number."

We then calculate the downshift as the fall in positivity as one shifts from a neutral to an ironic reading. For instance, consider the XYZ "Donald Trump is the Albert Einstein of diplomacy." The audience for certain cable-TV news networks may well be inclined to view this comparison literally, and the LLM suggests the neutral reading "Influential figures in their respective fields" with a high positivity rating of 95/100. In contrast, the LLM suggests the category "Unskilled or inept diplomats" as an ironic reading of this XYZ, the positivity of which it rates at just 10/100. The gap between these ratings, 85/100, is the ironic downshift we seek.

The mean downshift from neutral to ironic readings of XYZs in the ironic dataset is 15.1, while for neutral XYZs the mean drop is even greater, at 18.5. However, the largest drop is exhibited by the human (web) dataset, for which the positivity of ironic readings is 25/100 lower on average. But why should this be the case – that XYZs that are not designed to be ironic (our neutral dataset) or not expected to be ironic (our web dataset of human XYZs) show a greater capacity for dual interpretation as both praise and criticism?

There appears to be no single answer here. In some cases, the LLM simply produces the wrong generalizations. For instance, for the XYZ "Charlie Sheen is the Stephen Hawking of intellectual discourse" it suggests the neutral reading "Highly respected thinkers" and the ironic reading "Public figures known for their intellectual prowess." As it correctly assigns the same positivity rating of 90 to each of these, the estimated downshift is 0. At the same time, it rates the positivity of the comparison as just 15 (with the emotion *Disdain*), with a believability of just 10. From one perspective, the LLM sees irony (or perhaps sarcasm) here; from another, it fails to see the shift in meaning that irony should induce.

In contrast, a large potential downshift is only relevant if an audience is minded to seek it out. If an XYZ has high believability, there is no apparent incongruity to drive an ironic re-interpretation. Consider this XYZ from the human (web) dataset: "The Montrachet is the Angelina Jolie of the pack." Perhaps the speaker hopes to convey the idea that the wine is "full-bodied" or "luscious," but in any case the LLM assigns a positivity of 90 with a believability of 70. As a neutral reading it also suggests the category "Highly regarded and sought after" with a positivity of 100. However, when urged to adopt an ironic stance, it instead suggests the category "Overrated or overhyped entities" with a positivity of 0. Yet the availability of a cynical reading does not mean audiences will reach for this meaning unless nudged to do so. For irony to succeed, intent must work hand-in-hand with delivery.

The LLM's self-ratings give us the tools to accept or reject its candidates on the basis of our own acceptability criteria. To filter its efforts at irony, we should demand the following:

1. High positivity (e.g., *positivity* ≥ 60)
2. Low believability (e.g., *believability* ≤ 30)
3. Two contrasting interpretations (*neutral* and *ironic*)
4. A significant downshift (e.g., *downshift* ≥ 25)

As reported in Table 2, only a small subset of the XYZs in each dataset meet the above criteria. Just 1 in 20 of the XYZs in the ironic dataset conform to our ideal of a successful ironic comparison, that is, of a comparison that will both strike the audience as ironic and that will reward a deeper search for a veiled criticism. In contrast, more than 1 in 10 of the human XYZs from the web conform to this ideal.

Selection Criterion	Web	Neutral	Ironic
High positivity ≥ 60	74.77%	**96.96%**	62.13%
Low believability ≤ 30	41.21%	9.80%	**54.17%**
Pos. ≥ 60 & Believ. ≤ 30	**22.40%**	8.04%	20.56%
Two contrastive readings*	**43.35%**	30.29%	30.00%
Significant downshift ≥ 25	**41.48%**	29.41%	25.00%
All criteria together	**10.61%**	3.63%	5.46%

Table 2: % of XYZs in each set that meet our irony criteria. **(absolute downshift* ≥ 25*)*

Fortunately, fore-warned is fore-armed, and a creative system that uses an LLM as its generator can filter the 95% of LLM efforts that fall short of its desired standard for irony. Nonetheless, we note that the GPT LLM achieves a success rate at irony that is less than half that of human XYZs from the web when it is explicitly prompted to be ironic.

Other Language Models

A single LLM, OpenAI's *GPT4o-mini*, has driven our analysis up to this point. This is a commercial LLM whose exact parameter count is a proprietary secret, although the number is widely believed to be 8 billion. This model is comparable in benchmarked performance to OpenAI's much larger GPT-4, which is reported to have over a trillion parameters. In this section we broaden our analysis to include a range of rival LLMs, a number of which are open-source: *Llama-3.3* (70 billion parameters, from *Meta*), *Nemo-Instruct-2407* (12 billion, from *Mistral*), *Qwen2.5-72B-Instruct* (72 billion, from *Ali Baba*), *Gemma 2* (27 billion, from *Google*) and *DeepSeek*'s *R1* distilled into a Qwen model (32 billion).

XYZs are elicited from these LLMs using the same neutral and explicitly ironic prompts as were used for *GPT4o-mini*. The five new datasets cover the same range of macro-domains (*sport*, *politics*, *science*, etc.) and are self-rated for *positivity*, *believability* and so on using the same prompts. We observe a similar gender split in these new collections. The ratio of male to female Xs ranges from 1.8:1 *(Gemma)* to 2.6:1 *(Llama)*, while for Ys the male:female ratio ranges from 2.2:1 *(Llama)* to 4.6:1 *(R1)*. Conservation of gender is again evident across the board, with 80% *(Mistral)* to 92% *(Gemma)* of XYZs mapping Xs to Ys of the same gender. Neutral XYZs also tend to be intra-domain, with conservation rates ranging from 70% *(Llama)* to 94% *(Qwen)*, while ironic XYZs are much more likely to connect two domains.

As before, we use LLM estimates of *positivity*, *believability* and *downshift* to identify the XYZs in each dataset that conform to our ideal of a successful ironic comparison. As reported in Table 3, the highest yield is obtained from *Llama 3.3 (70B)*, with *Gemma 2 (27B)* coming a close second. According to our standard of meaningful pragmatic insincerity, *Llama 3.3* is four times more reliable than *GPT4o-mini* as a generator of appreciable, two-meanings-at-once irony. This is not easily attributable to the LLM's size, as the 72B *Qwen* model only achieves par with the much smaller *GPT4o-mini*.

Nor is this gap due to amount of compute that each LLM

Selection Criterion	**Llama 70B**	**Gemma 27B**	**Mistral 12B**	**Qwen 72B**	**R1 *Qwen* 32B**
High positivity ≥ 60	38.5% *(97.5%)*	53% *(98%)*	48.5% *(93.6%)*	47.3% *(97.5%)*	**76.5% *(97%)***
Low believability ≤ 30	**87.5% *(4%)***	70.5% *(0%)*	49.5% *(2.7%)*	65.5% *(1.5%)*	40% *(9.0%)*
Positivity ≥ 60 & Believability ≤ 30	**27% *(3.5%)***	26% *(8.0%)*	10.5% *(1.4%)*	16.8% *(1.0%)*	20.5% *(8%)*
Two readings (abs. downshift ≥ 25)	**63% *(35.5%)***	59% *(22%)*	60.5% *(28.2%)*	44.1% *(28.5%)*	41.5% *(40%)*
Significant downshift ≥ 25	**60% *(35.5%)***	58.0% *(22%)*	55.5% *(27.7%)*	40.9% *(27.5%)*	40.5% *(39%)*
All of the above criteria together	**19.5% *(1.5%)***	17.5% *(0%)*	8.0% *(0.45%)*	5.00% *(1.0%)*	12.5% *(5.0%)*

Table 3: % of XYZs that meet various irony criteria, as elicited using explicitly ironic prompts to each LLM. The neutral baseline, in which the LLM is *not* explicitly prompted for irony, is shown in parentheses.

dedicates to the generation task. DeepSeek's *Qwen* distillation of its *R1* model (DeepSeek-AI et al. 2025) expends significantly more time and compute when generating XYZs, and is an order of magnitude slower than the other LLMs. It pursues a deep chain-of-thought (CoT) approach in which the LLM reasons about the stated problem (Wei et al. 2022); this adds an average of 500 tokens to the generation of every 20 XYZs. While we expect ironic speakers to carefully plan their utterances, and to anticipate how they might split their audiences, this LLM – which ranks third in Table 3 – appears to *over*-think the problem and over-explain its reasoning.

The following is a representative extract from its CoT process: "Hmm, it's a start, but some of these don't fully capture the irony I was aiming for. I need to think more carefully about the contrasts between X and Y." These traces show the model doubting and criticizing itself, as we expect any good creator (and not just a *mere generator*) to do. However, self-criticism is valuable when its spurs insights, not excuses. Ultimately, the LLM cannot translate its insights into practice, and offers this apologia in its CoT trace: "Overall, creating truly ironic metaphors requires a delicate balance between the expected and unexpected elements. It's a bit challenging, and some of my attempts might not land perfectly, but this exercise helps in understanding how to juxtapose different domains to create thought-provoking comparisons."

So why do the less-reflective *Llama* and *Gemma* LLMs perform so much better? Perhaps certain creative goals are diminished when they are viewed as problems to be solved rather than as opportunities to invent. It may also boil down to how, and on what, these models are trained, and to how well they are tuned to respond to instructions (Ouyang et al. 2022) or to read between the lines of a user's request. We expect creative tasks to vary as to how heavily they rely on the two competing mental systems that (Kahneman 2011; Evans 2003) dub *System I* and *System II*. In this dual-process perspective on the human mind, *System I* is reactive, immediate, and instinctively intuitive; that is, it acts, or rather *reacts*, much like a well-trained, fine-tuned language model. Its training and biases shape what it does and what it says. In contrast, *System II* is slow, cautious and analytical, much like the chain-of-thought approach of DeepSeek's *R1/Qwen*. Our results suggest that generation (*System I*) should be separated from appreciation (*System II*) in two distinct phases. This allows an LLM to be freely and divergently generative in the first step, and convergently self-critical in the second.

Conclusions

We have sidestepped the question of whether an LLM is a creative producer of ironic comparisons, or just a "mere generator" of candidates (Ventura 2016). The larger, and better, question is whether a system built around an LLM can be creative in its selective generation of linguistic artifacts. Certainly, if prompted with care, LLMs such as *GPT4o*, *Llama* and *Gemma* are capable of fluency, flexibility, specificity and even originality – all of the dimensions of divergence identified by (Guilford 1950; Runco 2010) – in the creation of new texts. One might say that the LLM is merely generative, while the larger system that employs it is potentially creative, but this underestimates the value of LLMs as creative producers. As we have shown, the LLM itself can play a key role in evaluating and filtering its own outputs, to become a deliberate creator of meanings in its own right.

Irony is a facet of linguistic creativity that can be modeled using a broad spectrum of computational approaches. At one end of this spectrum sit the symbolic, rule-based approaches to generating irony, as typified by (Hao and Veale 2010). At the other sit the data-driven approaches to detecting irony, such as the statistical models of (Reyes, Rosso, and Veale 2013) or the neural models of (Ghosh and Veale 2017). The latter have the benefit of working with labeled data, allowing for a crisp definition of success and failure. The former, in contrast, must contend with what it means to be successfully ironic, and perhaps even creative (Colton and Wiggins 2012; Ritchie 2007); for that, a sense of humour may be needed also (Jentzsch and Kersting 2023; Góes et al. 2023). Here we have chiefly focused on what has been called *"theory of mind"* (ToM) in the context of irony (Strachan et al. 2024): the ability of a speaker to intuit how an audience will react to an ironic provocation, and the ability of an audience to infer the speaker's intended meaning. LLMs seem to show a poor grasp of ToM when aiming for irony, as only a minority of their XYZs clearly convey pragmatic insincerity. However, when probed on the dimensions of pragmatic insincerity, they show a firm enough grasp of these dimensions (as illustrated in Figure 4) to allow them to act intentionally. An LLM can serve a powerful, oracular function in AI systems, but as with any oracle, one must be careful what one asks, how one asks it, and how one interprets its replies.

Acknowledgments

This paper and its datasets are wholly the work of the author.

References

Attardo, S. 2000. Irony as relevant inappropriateness. *Journal of Pragmatics* 32(6):793–826.

Brants, T., and Franz, A. 2006. *Web 1T 5-gram Version 1*. Linguistic Data Consortium.

Chen, M.; Menick, J.; Lu, K.; Zhao, S.; Wallace, E.; Ren, H.; Hu, H.; Stathas, N.; and Such, F. P. 2024. GPT-4o-mini: Advancing cost-efficient intelligence. *OpenAI Blog, July 18*.

Clark, H. H., and Gerrig, R. J. 1984. On the pretense theory of irony. *Journal of Experimental Psychology: General* 113(1):121–126.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: the final frontier? ECAI'12, 21–26. NLD: IOS Press.

DeepSeek-AI; Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; and et al., X. Z. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv* 2501.12948.

Evans, J. 2003. In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences* 7(10):454–459.

Garmendia, J. 2019. *Irony. Key Topics in Semantics and Pragmatics*. Cambridge, UK: Cambridge University Press.

Ghosh, A., and Veale, T. 2017. Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Glucksberg, S., and Keysar, B. 1990. Understanding metaphorical comparisons: Beyond similarity. *Psychological Review* 97(1):3–18.

Glucksberg, S. 1998. Understanding metaphors. *Current Directions in Psychological Science* 7:39–43.

Glucksberg, S. 2008. How metaphor creates categories – quickly! In Raymond W. Gibbs, J., ed., *The Cambridge Handbook of Metaphor and Thought*. Cambridge, UK: Cambridge University Press.

Grice, P. 1978. Logic and conversation. In Cole, P., and Morgan, J., eds., *Syntax and Semantics 3: Speech Acts*. New York: Academic Press. 41–58.

Guilford, J. 1950. Creativity. *American Psychologist* 5:444–454.

Góes, L.; Sawicki, P.; Grzes, M.; Brown, D.; and Volpe, M. 2023. Is GPT-4 good enough to evaluate jokes? In *Proceedings of the 14th International Conference on Computational Creativity*.

Hao, Y., and Veale, T. 2010. An ironic fist in a velvet glove: Creative mis-representation in the construction of ironic similes. *Minds and Machines* 20(4):483–88.

Jentzsch, S., and Kersting, K. 2023. ChatGPT is fun, but it is not funny! Humor is still challenging Large Language Models. *arXiv* 2306.04563.

Kahneman, D. 2011. *Thinking, fast and slow*. London, UK: Penguin Books.

Kreuz, R. J., and Glucksberg, S. 1989. How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General* 118(4):374–386.

Kumon-Nakamura, S.; Glucksberg, S.; and Brown, M. 1995. How about another piece of pie: The allusional pretense theory of discourse irony. *Journal of Experimental Psychology: General,* 124(1):3–21.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; and et al., C. L. W. 2022. Training language models to follow instructions with human feedback. In *Proceedings of NeurIPS-22, the 36th Conference on Neural Information Processing Systems*.

Picasso, P. 1923. Picasso speaks: A statement by the artist. *The Arts: An Illustrated Monthly Magazine Covering All Phases of Ancient and Modern Art* 3(5).

Reyes, A.; Rosso, P.; and Veale, T. 2013. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation* 47(1):1–30.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17:76–99.

Runco, M. 2010. Divergent thinking, creativity, and ideation. In Kaufman, J. C., and Sternberg, R. J., eds., *The Cambridge handbook of creativity*, 413–446. Cambridge, UK: Cambridge University Press.

Sperber, D., and Wilson, D. 1981. Irony and the use-mention distinction. In Cole, P., ed., *Radical pragmatics*. New York, NY: Academic Press. 295–318.

Sperber, D., and Wilson, D. 1992. On verbal irony. *Lingua* 87(1-2):53–76.

Strachan, J. W. A.; Albergo, D.; Borghini, G.; Pansardi, O.; Scaliti, E.; Gupta, S.; Saxena, K.; Rufo, A.; Panzeri, S.; and et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour* 8(7):1285=95.

Valitutti, A., and Veale, T. 2015. Inducing an ironic effect in automated tweets. In *Proceedings of the 6th International Conference on affective computing and intelligent interaction (ACII2015)*. Xi'an, China: IEEE Computer Society. 153–159.

Veale, T. 2012. *Exploding the Creativity Myth: The Computational Foundations of Linguistic Creativity*. London, UK: Bloomsbury.

Veale, T. 2018. The 'default' in our stars: Signposting non-defaultness in ironic discourse. *Metaphor and Symbol* 33(3):175–184.

Ventura, D. 2016. Mere Generation: Essential barometer or dated concept? In *Proceedings of the 7th International Conference on Computational Creativity, Paris, France*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of NeurIPS-22, the 36th Conference on Neural Information Processing Systems*.