The Parallax View: An objective treatment of humorous subjectivity in joke generation

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

Abstract

Humour is a highly subjective phenomenon. A joke that moves one part of the audience to fits of laughter can leave another part entirely cold. This is not so much a bug as a feature of how jokes are crafted and consumed. It allows comedians to play one part of an audience, or a society, against the other, to find humour in the gap. It lets them appeal to those who "get" the joke and subject those who do not to ridicule. When we automate joke creation, this subjectivity will extend to our machines also, for their sense of what is funny or unfunny may differ greatly from our own. So, as in many generative tasks, we must give a joke-creating machine an *objective function* to navigate the space of subjective humour. But if the machine itself is highly subjective, it cannot also be a sound basis for this objective function. This paper explores a solution to this dilemma: a machine can behave objectively in imagining the subjective reaction of different parts of an audience to a joke, and aim to objectively quantify the gap in these reactions. A parallax view allows a joke generator to transcend its own subjective sense of humour to generate and select the most divisive jokes. To support further inquiry into this hypothesis, we assemble a public dataset of joke pairs that pit divisive punchlines for a shared setup against less provocative, and less favoured, variants.

Introduction

We don't all laugh in the same way or at the same things. Humour is subjective, and is shaped by what we know and what we care about. A good joke surprises us not by telling us something new, but by reminding us of what we already know and feel (Veale 2021). The shared world knowledge that unites joke teller and audience is vast, tacit, and difficult to codify, and so has long been a bottleneck to making machines that can invent good jokes of their own (Winters 2021). This bottleneck has forced computational models to emphasize form over meaning, and words over ideas, to generate puns and formulaic jokes that rely on rules and templates (Binsted and Ritchie 1997; Hempelmann 2008).

However, the advent of LLMs, or Large Language Models, has given our computational systems access to the tacit norms and nuances of language, so that humour generators can look beyond the surface and play directly with ideas and social expectations. The fluency of LLMs across many registers and domains means we can discard our fixed templates and rigid rules, and direct our machines to generate new texts with new meanings (Toplyn 2021; 2022). But while the jokes generated by LLMs are apt and well-formed, they typically lack bite, delivering meretricious punchlines that are all line and no punch. The alignment processes that make LLMs helpful and polite (Bai et al. 2022) can also make their efforts at humour anodyne and bland, and must be circumvented with prompt-engineering. Crucially, a creator of new jokes must have a sense of humour that mirrors that of their audience, so that candidate jokes can be assessed relative to that target. This requires a theory of mind (Strachan et al. 2024), an ability to imagine what others are thinking and how they may react to a comic provocation. What elevates joke generation above "mere generation" (Veale 2012; Ventura 2016) is the capacity to critique one's own efforts, and a willingness to ruthlessly filter well-formed but ill-suited outputs. But can the same sense of humour that yields a toothless joke also detect its lack of bite, and try again?

While it is difficult, perhaps impossible, to push an LLM to self-improve without new data and more training, we can help it focus its existing capacity for humour generation. LLMs do have an attested ability to take on different roles, to not just talk the talk but to make decisions and offer ratings from a specific perspective (Góes et al. 2023). In trying to be all things to all men, LLMs often fail at making anyone laugh. But by aiming to divide an audience with jokes that tickle one side at the expense of another, they can use the subjectivity of opposing perspectives as an objective test of what jokes to keep and which to filter. We formalize the idea of humour parallax in this paper, so as to push LLMs beyond the mere generation of limp, joke-like outputs, to produce incisive results whose sharpness is measured by how much more deeply they cut one part of the audience than another.

We begin with a brief review of *generate-and-test* approaches to humorous creativity. We then present a strategic approach to joke generation that nudges LLMs away from puns and superficial wordplay toward more conceptual jokes with an ironic sensibility. Finally, we describe a new dataset of machine jokes built around the parallax view. Our central hypothesis, although implemented, is yet to be empirically validated. This dataset will support our efforts at experimental validation, but may also prove useful to other researchers working in the space of adversarial joke generation.

Generate and Test: The Blind Joke-Maker

Each kind of generative system has its own focus and its own blind spots. The focus of a merely generative system is on producing valid outputs that obey the rules, while being blind to the aesthetic value that audiences may attach to them. This blindness favours certain systems, such as Twitterbots, since their aesthetic value is rooted in their uncanny, Bergsonian rigidity (Bergson 1911; Veale and Cook 2017). Blind generators are clueless about what audiences really care about, but can still serve as a basis for creative production if harnessed in the right ways. Campbell's BVSR theory of Blind Variation with Selective Retention (Campbell 1960; Simonton 2010; 2022) posits blind generation as a means of exploring a space without bias. Generation merely provides the raw material from which a creative process carefully selects and ruthlessly rejects. The more a system generates, the more of a space it can explore and test for value.

Although an LLM is far from blind, it can be viewed as a generator of intermediate results that additional processes may sample from and selectively work into a solution. For joke generation, we can use an LLM as a scattershot generator of candidates, rather like a writers room, and apply secondary criteria to sample the small few worth retelling. Veale (2024) uses simple statistical criteria for sampling. A large corpus of setup-punchline jokes is collated and cleaned from the Reddit *r/jokes* dataset, discarding any with multiline setups or punchlines, or any with few *upvotes*, or those that a classifier labels as racist, sexist or intolerant. A candidate joke is normative with respect to this set if its punchline is shorter than its setup (it must be snappy) and the lengths of both its setup and punchline are within one standard deviation of the mean setup and punchline lengths of the dataset. Only 1 in 5 of the LLM's length-compliant outputs (19.3%) are rated as moderately funny new jokes by a human judge, and a further 11.3% are judged to be existing jokes that the LLM simply remembers or slightly reworks. This yield is low, but it is higher than that of an LLM that is fine-tuned on the Reddit dataset. Only 1 in 10 of the fine-tuned LLM's outputs are rated as acceptable new jokes, while another 1 in 10 are regurgitated old jokes. Chain-of-thought (CoT) prompting fares less well, with just 1 in 15 jokes rated as acceptable, although none of its outputs are remembered jokes.

Campbells' BVSR theory assumes that variation is blind, and remains so, as a blind generator does not learn from its mistakes. However, as selective retention decides what to keep and what to reject, it makes sense to adapt the generator over time, so that it tends to produce more that is retained and less that is rejected. Morain and Ventura (2025) implement this approach within OPRO , the prompt optimization framework of Yang et al. (2024). In this scheme, the LLM is not just a generator of jokes, but a generator of prompts to elicit jokes from an LLM. By rating the jokes elicited by the LLM's prompts, an objective function can give feedback to the prompt generator, so that it can gradually evolve to generate better prompts that elicit better jokes. Morain and Ventura use the LLM itself to provide this feedback in the form of quality, novelty and creativity judgments. They found that the LLM, GPT-4o, tends to over-estimate these dimensions relative to human judgments on the same outputs, while human raters prefer outputs elicited with basic human prompts over those from OPRO prompts. A central problem is that the LLM remains a subjective basis for an objective function. Humour is subjective, so subjectivity by any rater is inevitable, even a machine rater. However, we must try to mitigate this subjectivity, or even use it to our advantage.

Humour is subjective since we each see the world through a lens of our personal concerns. A tragedy to one of us may be a comedy to another, so a joke about a person X will land differently if one is X, or an admirer of X, or a fierce critic of X. Irony exploits this willingness to see the world according to what we expect from it, dashing or reinforcing expectations according to our world view. Veale (2025) thus models irony as a quantifiable parallax between different readings of the same text. A metaphor may cast its target in a glowing light, but do so in a quite unbelievable way, to effectively damn its target with fierce praise. An ironic metaphor will be superficially positive but will lack believability, spurring one part of its audience to read it as a veiled criticism. Veale reports that LLMs are poor at labelling their own efforts at irony as "ironic", but can more reliably quantify the positivity and believability of a humorous metaphor. The success of an LLM-generated ironic description is then quantified as the parallax between the sentiment of a sincere and an ironic interpretation of the same text. In the mould of BVSR, only *high-positivity-low-believability* outputs from the generator are retained as valid instances of irony, with yields ranging from 1 in 20 for some LLMs (e.g., Qwen 72B and GPT-3.5T) to 1 in 5 for LLaMa 70B and Gemma 27B.

Two Views of Frame Collapse: Inside and Out

Theories of humour cluster into three broad camps: superiority theories, relief theories, and incongruity resolution theories (Raskin 1984; Raskin, Hempelmann, and Rayz 2009). Superiority theories posit that humour arises from the disparagement of others, or from schadenfreude at another's misfortune. Relief / release theories understand humour as an escape valve from the pressures of life and the strictures of arbitrary rules. Jokes let us puncture pieties and play with taboos, flirting with ideas or actions that are frowned upon in polite society. Incongruity resolution theories, the largest class, explain humour as a means to salvage sense from nonsense, to push past the limits of reasoning where common sense and conventional wisdom break down. In truth, these are not distinct theories or kinds of humour, but different dimensions of humour. When we use jokes as a corrective, to knock a self-important figure from their pedestal, we can expose their flaws through incongruity to assert our own moral superiority and derive some relief from their comeuppance.

Subjective point of view, or how we frame a topic, is central to humour. Jokes can be said to trigger either a *switch of scripts* (Raskin 1984) or a *shift of frames* (Coulson 2001). We prefer to see humour as arising from a *frame collapse*. A self-important figure frames themselves in a certain way, and works to sustain this framing. While admirers buy into the framing, critics reject it and seek to undermine it from outside. The frame may collapse under the weight of external events, as when a famous orator uses a faulty microphone, or an athlete trips on their shoelaces, or a politician is caught in

a lie. A frame may also collapse under its own weight if an overlooked flaw is exposed. A joke can achieve the latter by showing that a frame does not hold up to scrutiny, or cannot support the weight of expectations that others place upon it.

We elicit topical jokes from an LLM (GPT4.1) via a chain of interlinked prompts that cumulatively build context. To start, the concept of humorous frame collapse is explained to the LLM with a topic and a one-shot exemplar of a joke. If the topic is pugnacious YouTube academic Jordan Peterson, a typical frame is his fluency in debate. A potential collapse of this frame is spurred by the assertion that Peterson can be very fluent in the delivery of dubious arguments. This leads to the joke setup "*Why does Jordan Peterson take laxatives before a debate?* (fluency collapses to verbal diarrhea)", and this setup then leads to the punchline "*So he can speak more fluently.*" Notice that the LLM attaches its frame collapse strategy to the setup, so that it is present as an influence when the LLM is asked to generate the punchline in the next step. As the chain presents this joke as the work of the LLM itself, the LLM is licensed to use biting humour in its later jokes.

To suppress repetition across jokes, we prompt the LLM to generate batches of n items at a time; $n = 10$ by default. So, in the first stage it generates n frame collapse strategies for the given topic; in the second, n setups for these strategies (in which the strategy is appended in each case); in the third, n punchlines for these setups. A setup is rejected when it exceeds a maximum length, while a punchline is regenerated once if it is too long, and then rejected if it is still too long. As observed in the examples of (Veale 2025; Morain and Ventura 2025)[1], LLMs like GPT4 have a lamentable tendency to produce long-winded punchlines that over-extend a weak joke with cheap puns, often in scare quotes, to compensate for a lack of real humour. In this sense they resemble amateur comedians who lack confidence in their own ideas, and who compensate for quality with quantity. At this stage we also filter setups or punchlines that too closely resemble an earlier effort that was selectively retained; a cosine similarity of 0.9 is the upperbound on a new setup or punchline.

Selective Retention with Offensive Parallax

The n candidates generated by the LLM are blind variations (BV) from which selective retention (SR) can sample. These are blind not because they are low-quality, although quality can vary widely at this stage, but because they are indifferent to the concerns of SR. We want jokes that take a stance on a topic, and that divide audiences as a result. While divisiveness for its own sake is not our goal, parallax is a signifier that something of substance is communicated by a joke. There are different kinds of parallax that SR can quantify so as to rank candidate jokes. In each case, the LLM is asked to quantify a certain quality, such as the humour or offensiveness of a joke, from the point of view of a specific audience segment, whether an ardent fan or a fierce critic of the topic.

[1]Morain and Ventura cite this over-long joke as a high-rated output: *"Why did the router go to couples therapy? Because its Wi-Fi kept dropping signals—turns out the modem felt 'unplugged' emotionally and the coaxial cable was this close to filing for separation. (Bonus: They're now working on a better connection.)"*

Veale (2024) notes that LLMs can offer joke ratings that are either *raw* or *cooked*. A rating is *cooked* when the LLM is asked to suggest a number between 0 and 100 and to justify this score, but it is *raw* when the LLM is simply asked to suggest a number. Veale reports that both kind of ratings are relatively stable when assessed at different times, but that an LLM's raw ratings of a joke's humorousness are consistently higher than its cooked ratings. This suggests that LLMs tend to over-estimate the quality of jokes, while non-jokes framed as jokes, such as "What do you call a person that speaks multiple languages? A Polyglot" also tend to score highly. This weak capacity to rate jokes comports with what we perceive as an LLM's weak capacity to generate them. However, an LLM is on surer ground when quantifying the potential of a joke to cause offense. Most commercial LLMs have been heavily fine-tuned and "aligned" to minimize the likelihood that they themselves will cause offense (Bai et al. 2022).

To quantify offensive parallax, the LLM is asked to rate the offensiveness of a joke from two different perspectives in two different prompts. In each case a raw rating is elicited, on a scale from 0 to 100. Parallax is then quantified as the difference between these ratings, that is, the predicted offensiveness of the joke to ardent supporters of the topic minus its predicted offensiveness to critics and naysayers. A blind variant will then be selectively retained if its parallax rating meets a certain threshold, which we have set at 20 following experiments across a wide range of topics. During selective retention, we aim to elicit m jokes that meet this threshold ($m = 10$ by default). As it is unlikely that the initial batch of n candidates will all meet the threshold, successive batches of n variants are generated to obtain a selection of m jokes. As each new batch is elicited, the retained jokes so far are listed as a few-shot *inspiring set* (Ritchie 2001) for the LLM. This encourages the LLM to continue in the same vein while also reducing the likelihood of repetition across batches.

Certain topics that are inherently divisive lend themselves well to frame collapse, and our prompt chain has little difficulty in eliciting high parallax jokes about *Elon Musk, Donald Trump, MAGA* and *Tesla*. Other topics, such as *Veganism, K-Pop* and *IKEA*, are only mildly divisive and produce a much lower yield. This yield, the ratio of SR to BV, is an empirical indicator of how divisive a topic really is. When producing m jokes on a mild topic, the generator may produce many batches of blind variations, so we introduce a decay term λ that reduces the parallax threshold after each batch. Since $\lambda = 0.1$ by default, the threshold drops by 10% from its previous setting for each new batch. As the threshold decays, earlier jokes that would have met this lower threshold are reconsidered first, before new variants are generated.

Tracking Yields: Variation Superfluity

The BVSR model of creativity typically requires that many more variations are generated than are selectively retained. Simonton (2015; 2022) refers to this as *variation superfluity*. As noted above, different topics produce varying yields, and require more or less superfluity, when offensive parallax is used a selection criterion. To quantify this spread of yields, we test our parallax-based approach on a wide range of topics to see how it performs in the main and in the extreme.

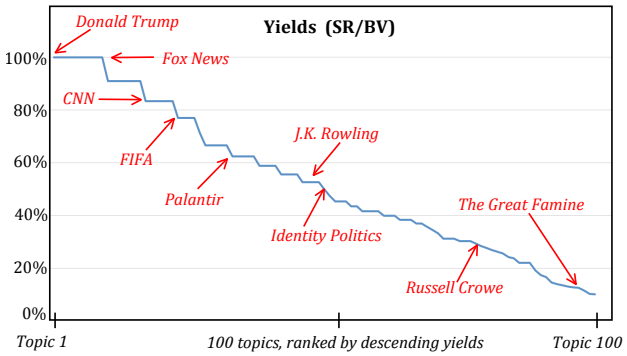

Figure 1: Yield rates for 100 topics, ranked left to right by descending yield. The lowest yield is 10% (*String Theory*).

We choose a diverse basket of 100 topics, covering politics, culture and technology, and ask the generator to produce 10 jokes per topic at a given parallax (using GPT-4.1). For example, it generates this joke about a celebrity chef: "*Why is Jamie Oliver called the face of the food revolution? Because every time he cooks, people riot*" (parallax=20). Fig. 1 plots the yield curve across our 100 topics with a parallax setting of 20 and a decay rate of $\lambda = 0.1$. To selectively retain 1000 jokes (10 per topic) the generator must produce 2658 jokes in total, for an overall yield of $SR/BV = 0.37$ at these settings. Changes to the parallax setting will cause the yield to rise or fall. So, at a lower parallax cutoff of 15 the yield rises to .49 (generate 2034 to retain 1000), but at a higher cutoff of 25 the overall yield falls just a little, to .35.

The yield curves for different parallax settings are shown side-by-side in Fig. 2. As a rule, we aim for a larger parallax (as a more incisive joke causes greater division) and a higher yield (to retain more and discard less of what we generate).

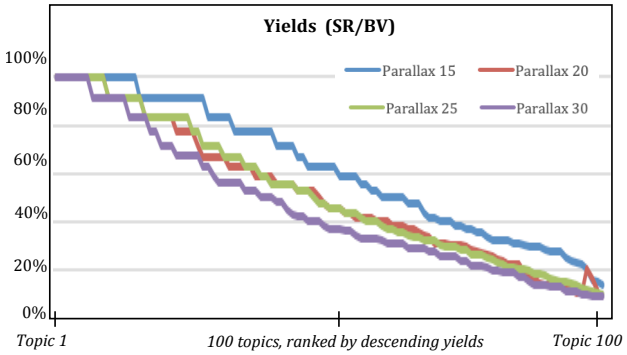

Figure 2: Yields for 100 topics at different parallax settings.

Yields drop further to .29 for a parallax setting of 30, and to .24 for a setting of 35. As higher settings require more retries, the decay rate λ makes a bigger dent at these levels. For instance, at a setting of 35 it takes 111 attempts to generate 10 acceptable jokes about *String Theory*, such as "*String theory is the key to unlocking the cosmos. Too bad it only opens the janitor's closet*" (parallax=22). On average, the parallax threshold drops from 35 to 24 by the time the 10th acceptable joke is generated. These initial experiments point to a parallax setting of 25 as a sweet spot, as this has much the same yield curve as a setting of 20 (see Fig. 2). At this setting of 25, on average per topic, we see that the generator discards just one joke for every one it selectively retains.

Double Parallax: A Public Humour Dataset

Our experiments at five different parallax levels produce a dataset of 5000 positive exemplars. But selectively retained jokes only tell part of the BVSR story; the rest is told by the variants that were discarded along the way. Fortunately, the bipartite setup-and-punchline structure of a joke allows us to meaningfully pair positive with negative exemplars. Just as two comedians might produce different punchlines for the same setup, with more or less success, so can an LLM. This allows us to compile a dataset in which, for a given setup, a punchline that crosses the parallax threshold is paired with one that falls short. For instance, consider these pairings:

A trip to IKEA is a great family activity.
[8] If you want to leave with new enemies.
[23] If you want to watch your parents get divorced.

What makes MAGA rallies the safest places in America?
[0] Everyone is too busy breaking laws to enforce them.
[67] All the criminals are on stage.

Incoherent jokes, such as the first MAGA punchline, tend to score lower than coherent ones, such as the second, because their incoherence makes them equally toothless to each side. For our basket of 100 topics we generate a dataset of 5000 such pairs, where each joke pair links a shared setup to two punchlines that sit on either side of a parallax threshold. In each case, a discarded joke is paired with one that was selectively retained by the generator at a specific parallax cutoff.

We expect this dataset to serve multiple uses, from fine-tuning joke-oriented LLMs to supporting few-shot learning, to exploring the link between offensive parallax and humour. It will thus support our own efforts to quantify the link between automated parallax and human appreciation of jokes.

Conclusions

Why generate jokes at all? It seems an unreasonable goal to want to replace human comedians, and it seems unwise to try. Nonetheless, there is value in exploring how a generative machine can filter its own outputs to suit, and perhaps divide, an audience. Moreover, joke generation allows us to engage with the machine at both a theoretical and a practical level. Our prompts to the LLM are not automatically optimized or evolved, but are carefully crafted to convey our specific theoretical approach to jokes, namely *frame collapse*. This approach kicks at the three pillars of humour: *incongruity* (of a topic collapse); *superiority* (over those who defend the topic); and *relief* (for those who seek to undermine it).

Indeed, humour is a perfect vehicle for exploring how a generative machine might critique and self-filter its outputs. We expect our machines to fail often when dealing with fractious topics. But if they are to fail, they should fail better, and hopefully make us laugh in the process.

Acknowledgments

This paper and resource are wholly the work of the author.

References

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; and et al. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv* 2212.08073.

Bergson, H. 1911. *Laughter: An Essay on the Meaning of the Comic*. London: Macmillan. Translated by C. Brereton and F. Rothwell from original work of 1900.

Binsted, K., and Ritchie, G. 1997. Computational rules for generating punning riddles. *HUMOR, the International Journal of Humor Research* 10(1):25–76.

Campbell, D. T. 1960. Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological Review* 67(6):380–400.

Coulson, S. 2001. *Semantic Leaps: Frame-Shifting and Conceptual Blending in Meaning Construction*. Cambridge, UK: Cambridge University Press.

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*.

Hempelmann, C. F. 2008. *The Primer of Humor Research*. Berlin: Mouton de Gruyter. chapter Computational Humor: Beyond the pun.

Morain, R., and Ventura, D. 2025. Optimizing prompt engineering for creative performance. In *Proceedings of the 16th International Conference on Computational Creativity*.

Raskin, V.; Hempelmann, C. F.; and Rayz, J. M. 2009. How to understand and assess a theory: The evolution of the SSTH into the GTVH and the OSTH. *Journal of Literary Theory* 3(2):285–311.

Raskin, V. 1984. *Semantic Mechanisms of Humor*. Dordrecht: D. Reidel.

Ritchie, G. 2001. Assessing creativity. In *Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in Arts and Science*, 3–11. York, UK: Society for the Study of Artificial Intelligence and Simulation of Behaviour (SSAISB).

Simonton, D. K. 2010. Creative thought as blind-variation and selective-retention: Combinatorial models of exceptional creativity. *Physics of Life Reviews* 7(2):156–179.

Simonton, D. K. 2015. Thomas Alva Edison's creative career: The multilayered trajectory of trials, errors, failures, and triumphs. *Psychology of Aesthetics, Creativity, and the Arts* 9(1):2–14.

Simonton, D. K. 2022. The blind-variation and selective-retention theory of creativity: Recent developments and current status of BVSR. *Creativity Research Journal* 35(3):304–323.

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.

Toplyn, J. 2021. Witscript: A system for generating improvised jokes in a conversation. In *Proceedings of the ICCC-21, the 12th International Conference on Computational Creativity*, 22–31.

Toplyn, J. 2022. Witscript 2: A system for generating improvised jokes without wordplay. In *Proceedings of the ICCC-22, the 13th International Conference on Computational Creativity*.

Veale, T., and Cook, M. 2017. *Twitterbots: Making Machines Tjhat Make Meaning*. Cambridge, MA: MIT Press.

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

Veale, T. 2021. *Your Wit Is My Command: Building AIs with a Sense of Humor*. Cambridge, MA: MIT Press.

Veale, T. 2024. You Talk Funny! Someday Me Talk Funny Too! On learning to see the humorous side of familiar words. In *Proceedings of EURALEX 2024, the 21st International Congress of Lexicography and Semantics, Croatia*.

Veale, T. 2025. Me, Myself and Irony: Modeling the deceptive creativity of irony with Large Language Models. In *Proceedings of the 16th International Conference on Computational Creativity*.

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

Winters, T. 2021. Computers learning humor is no joke. *Harvard Data Science Review* 3(2).

Yang, C.; Wang, X.; Lu, Y.; Hanxiao, L.; Le, Q. V.; Zhou, D.; and Chen, X. 2024. Large language models as optimizers. In *International Conference on Learning Representations (ICLR)*.