## Title
Interpretation of Novel Literary Metaphors by Humans and GPT-4

## Permalink
https://escholarship.org/uc/item/9g40q0bh

## Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

## Authors
Ichien, Nicholas
Stamenkoviƒá, Du≈°an
Holyoak, Keith

## Publication Date
2024

Peer reviewed

# Interpretation of Novel Literary Metaphors by Humans and GPT-4

**Nicholas Ichien (nichien@upenn.sas.edu)**
Department of Psychology, University of Pennsylvania
425 S. University Ave, Philadelphia, PA 19104 USA

**Dušan Stamenković (dusan.stamenkovic@sh.se)**
School of Culture and Education, Södertörn University Stockholm
141 89 Huddinge, Sweden

**Keith J. Holyoak (kholyoak@g.ucla.edu)**
Department of Psychology, University of California, Los Angeles
502 Portola Plaza, Los Angeles, CA 90095

## Abstract

Despite the exceptional performance of large language models (LLMs) on a wide range of tasks involving natural language processing and reasoning, there has been sharp disagreement as to whether their abilities extend to more creative human abilities. A core example is the interpretation of novel metaphors. Given the enormous and non-curated text corpora used to train LLMs, a serious obstacle to designing tests is the need to obtain novel yet high-quality metaphors that are unlikely to have been included in the training data. Here we assessed the ability of GPT-4, a state-of-the-art large language model, to provide natural-language interpretations of novel literary metaphors drawn from Serbian poetry and translated into English. Human judges—blind to the fact that an AI model was involved—rated metaphor interpretations generated by GPT-4 as superior to those provided by a group of college students. In interpreting reversed metaphors, GPT-4, as well as humans, exhibited signs of sensitivity to the Gricean cooperative principle. These results indicate that LLMs such as GPT-4 have acquired an emergent ability to interpret literary metaphors.

**Keywords:** metaphor, large language models, natural language processing

## Introduction

The poet Robert Frost asserted, "The richest accumulation of the ages is the noble metaphors we have rolled up" (Frost, 1931, p. 108). The world's literature (Holyoak, 2019; Lakoff & Turner, 1989; Steen, 1994), as well as everyday speech (Lakoff & Johnson, 1980), is replete with non-literal comparisons of things that are on the face of it unlike each other, e.g., "'Hope' is the thing with feathers —That perches in the soul" (Emily Dickinson). The ability to create and interpret novel metaphors is considered one of the pinnacles of human cognitive abilities, extending literal language and perhaps involving sophisticated analogical reasoning (Bowdle & Gentner, 2005; Holyoak & Stamenković, 2018). If artificial intelligence (AI) aims to ultimately reach or exceed human cognitive abilities, then models of natural language processing and general intelligence will need to acquire the ability to interpret (and perhaps create) novel metaphors.

## Metaphor in Large Language Models

The advent of large language models (LLMs) has triggered intense interest in whether these new AI models are in fact approaching human-level abilities in language understanding (DiStefano et al., 2023; Mahowald et al., 2023; McClelland et al., 2020) and various forms of reasoning (Binz & Schulz, 2023; Chan et al., 2022; Dasgupta et al., 2022; Srivastava et al., 2022; Wei et al., 2022), including analogy (Webb et al., 2023), with vigorous debate between critics (Chomsky et al., 2023) and advocates (Piantadosi, 2023). Given the enormous and non-curated text corpora on which LLMs have been trained, these models have certainly had ample opportunity to mine the metaphors that humans have already formed and planted in texts. More generally, recent reports suggest that LLMs have been exposed to some or all standard AI benchmarks for cognitive tasks, rendering these benchmarks suspect for distinguishing deeper cognitive processing from memorization (Bubeck et al., 2023). It would not be surprising to find that LLMs succeed on tasks involving linguistic expressions contained in their training data, including metaphors.

A serious test of an LLM's ability to deal with novel metaphors requires challenging it with metaphors that are both novel and *apt* (i.e., metaphors in which the source is perceived as providing a unique and accurate description of the target). All conventional metaphors, e.g., "Life is a journey," are certain to have been included in the text corpora used as the training set. Creating novel metaphors that have never been uploaded to the internet is extremely challenging.

Here we test GPT-4, a state-of-the-art LLM, on its ability to generate natural-language interpretations of literary metaphors that passed tests assessing their novelty to the model. Computational analyses have shown that literary metaphors are distinguished by the qualities of high surprisal (a statistical measure of the unexpectedness of words), relative dissimilarity of source and target concepts, the combination of concrete words with relatively complex



In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* 

grammar and high lexical diversity, and extra difficulty (for people) in comprehending the metaphorical meaning (Jacobs & Kinder, 2017, 2018). Studies of individual differences in cognitive abilities have shown that crystalized intelligence (closely linked to verbal ability) impacts comprehension of both conventional and literary metaphors; fluid intelligence (on which analogical reasoning depends heavily) plays a greater role for more complex literary metaphors (especially when presented in isolation without a supportive verbal context) (Stamenković et al., 2020, 2023; Stamenković, Ichien, et al., 2019). Novel literary metaphors thus pose the most challenging test of metaphor comprehension in humans and perhaps AI models. To find literary metaphors unlikely to have been included in the training set for GPT-4, we used English translations of metaphors found in Serbian poetry.

## Interpreting Metaphors from Serbian Poetry

### Method

We had college students and GPT-4 interpret metaphors that originated in Serbian poems and had been translated into English. The original metaphors were rated as highly apt by native Serbian speakers, but were not widely known to them; we assessed and then controlled for the familiarity of the English translations to English-speaking participants and to GPT-4.

#### Selection of Metaphors

The test set included 55 literary metaphors drawn from Serbian poetry and normed on several properties (Milenković et al., under review; Stamenković, Milenković, et al., 2019). The norming studies (primarily following methods used previously; Katz et al., 1988), included metaphors chosen by a literary expert from over 65 nineteenth and twentieth-century poems written by various Serbian poets, including Branko Radičević, Laza Kostić, Vojislav Ilić, Đura Jakšić, Desanka Maksimović, Vladislav Petković Dis, and Branko Miljković. The poems selected for the norming study aimed to represent a wide range of poetic movements and styles across these two centuries. The expert had the task of extracting all metaphorical expressions from these poems. These were then grouped, with all similar/duplicate metaphors counted as one. Subsequently, all metaphors were transformed into *<nominal> is <nominal>* format, resulting in the finalized list of 55 items.

In the first norming study, 235 Serbian-speaking participants rated these 55 metaphors for quality, metaphoricity, aptness, familiarity, comprehensibility, and source-target similarity using a 7-point Likert scale (min = 1, max = 7). The inter-scale correlations for Serbian poetic metaphors were reliable for most dimensions (Stamenković, Milenković, et al., 2019) (ranging from .77 between aptness and quality, and between quality and source-target similarity, to .27 between familiarity and metaphoricity). The only nonsignificant correlation was that between source-target similarity and metaphoricity (.17). These 55 literary metaphors were then translated into English by two translators and the translations were verified by a third. This new list was subjected to two further norming studies (Milenković et al., under review) in which a combined total of 252 (186 for the full set and 66 for a shorter set) English-speaking Serbian participants rated the set for quality, metaphoricity, aptness, and familiarity. These Serbian translations were selected to minimize the likelihood that their English translations appeared in GPT-4's training data.

#### Assessment of Novelty to GPT-4

We performed several tests to evaluate whether GPT-4 was familiar with the metaphors used in our experiment. First, we probed it with questions to assess its knowledge of one paper (Stamenković, Milenković, et al., 2019) and an online supplement to another paper (Stamenković et al., 2023), in which some of these metaphors were discussed. Neither of these publications included the metaphor interpretations that GPT-4 generated in our tests. GPT-4 recognized one paper (Stamenković, Milenković, et al., 2019) and summarized its main points, but was unable to report any of the metaphors in it. When asked for examples, it provided several unrelated metaphorical expressions that did not appear in the study ("hallucinations"). The later paper (Stamenković et al., 2023) was published in March 2023, making it too recent to have been included in GPT-4's training data (OpenAI, 2023); the model did not recognize its online supplement at all.

Second, we considered the possibility that the Serbian-language version of GPT-4 might be familiar with the original Serbian poems from which our materials were derived. We therefore performed an additional test conducted entirely in Serbian using the Serbian-language version of GPT-4. We first provided the model with a list of the 65 poems from which our metaphor stimuli were drawn and adapted, with the query (translated from Serbian): "This is a list of poems and their respective authors. Tell me if you recognize any of them." GPT-4's responses were uniformly vague, so we added a follow-up prompt: "Look at the list of poems I have provided and tell me which of them you could analyze if I asked you to do so." Out of the 65 poems, GPT-4 indicated it could analyze 14. We then assessed the system's knowledge of these 14 poems using a "fill-in-the-gap" exercise, using the query: "Could I provide poems with blanks, where your task would be to fill in these blanks with the missing parts?" In this test, we provided the system with each poem's title and author, along with the poem itself, omitting the lines that contained our selected metaphors (in their original form), and we prompted GPT-4 to fill in the missing lines. Despite being given the above details about each poem, the system was unable to generate the omitted lines for any of the poems. Instead, it generated its own replacements for the missing lines, trying to imitate the style of the poet and the poem, and sometimes explicitly stating that the poem is not available in its current database. We thus found no evidence that GPT-4 had been exposed to the Serbian versions of the metaphors used in our test.

Third, we considered the possibility that at least some of the translated Serbian metaphors had been generated

Table 1: Example of GPT-4's interpretations of metaphors presented in their canonical direction (left columns) and in their reversed direction (right columns).

| Canonical | | Reversed | |
|---|---|---|---|
| Metaphor | GPT-4 interpretation | Metaphor | GPT-4 interpretation |
| A waterfall is a wild, unbridled horse. | A waterfall symbolizes the untamed, powerful energy of nature, much like the spirit of a wild, unbridled horse. | A wild, unbridled horse is a waterfall. | A wild, unbridled horse is a waterfall suggests that unrestrained energy and freedom can possess a raw, captivating beauty. |
| A cemetery is a sad theatre. | A cemetery is a melancholic stage where the final acts of life are commemorated, and loved ones gather to mourn and remember the departed. | A sad theatre is a cemetery. | A sad theatre is a cemetery conveys that a place once filled with emotion and expression can become desolate and lifeless when devoid of its original purpose. |
| Love is radiance. | Love is the glowing warmth that illuminates and uplifts our lives. | Radiance is love. | Radiance is love suggests that the emotion of love has the power to illuminate and transform our lives, shining with warmth and positivity. |

previously in English and uploaded to the internet. In the absence of direct access to GPT-4's training corpus, we quote-searched each metaphor using Google's search engine. Of the 55 metaphors drawn from Serbian poetry, we found a match for 19 of their English translations published online before 2022 (34.5%). For example, we found the expression "Love is radiance" in the poem "Love" by Miklos Zoltai, published online in March 2016: "Love is radiance, lover is perfumed of your self-knowledge, love is the essence of your existence." We coded each translated metaphor as having some online match or not. This variable was included as a predictor in our analyses to assess whether the previous online publication (and hence potential inclusion in GPT-4's training corpus) influenced ratings of GPT-4's metaphor interpretations by human judges. (See interaction term *interpretation source x online match (yes vs. no)* in analyses of data presented in Figure 1 below.)

**Human Experiment and Comparison with GPT-4**

In order to evaluate GPT-4's ability to generate high-quality interpretations of metaphors, we compared its interpretations with that of 39 undergraduate psychology students at the University of California, Los Angeles (UCLA), who completed our task for course credit (approved, including informed consent procedures, by the UCLA Office of the Human Research Protection Program). In order to elicit responses from human participants and GPT-4, we plugged each metaphor into the following prompt: "Please provide an interpretation for the following expression: <SENTENCE>", where the sentence stated the metaphor. Because GPT-4 tended to provide longer interpretations, we prompted it to provide a "short sentence-long interpretation" to elicit responses that were relatively succinct and of similar length to those generated by human participants. Importantly, the experimental task for both human participants and for GPT-4 omitted any mention of the term 'metaphor', opting instead for the more neutral term 'expression' as used above. Thus, no overt cue indicated that the task had anything to do with metaphor.

In order to provide a qualitative assessment of GPT-4's text interpretations, we also assessed the extent that both human- and model-generated text interpretations followed the Gricean cooperative principle, which implies that inapt or poor-quality metaphors may be reinterpreted as expressing a more apt and informative comparison than what they literally say (Chiappe et al., 2003; Grice, 1975). To do so, we elicited interpretations of literary metaphors both in their canonical (apt) form (e.g., "Love is radiance") as well as in their reversed (inapt) form (e.g., "Radiance is love"; see Table 1 for additional examples).

Human participants provided interpretations for 55 metaphor items, each in either its canonical form (e.g., "Love is radiance") or reversed form with source and target switched (e.g., "Radiance is love"). Which form a metaphor took was randomized across items for each participant. Across all unique items (110 total with 55 canonical form and 55 reversed), we collected an average of 19.54 human responses for each item (range = [12, 27]). GPT-4 provided an interpretation of all 110 metaphor items. We used GPT-4's chat interface (in late March 2023) to collect each response after starting a new chat window. This procedure prevented the system from conditioning its response on previously-seen metaphor items, so that its performance was zero-shot.

After eliciting interpretations from both humans and GPT-4, three undergraduate judges who were naïve to presence of nonhuman, model-generated text scored each interpretation on a 0-2-point scale. Judges were instructed to assign interpretations a score of 2 if they attributed properties of the source to the target that described the target aptly. Interpretations were to receive a 1 if they attributed certain properties to the target that seemed less appropriate, especially given the source. Each judge also provided their own interpretation of each metaphor, and all judges scored these judge-generated interpretations. As an additional scoring guide, judges were given each of these judge-generated interpretations, along with the average score each of these interpretations received. This scoring guide is available on this paper's OSF page: https://osf.io/jcg3f/. Finally, interpretations received a 0 if they were blank or failed to express any sort of comparison. For each metaphor, judges were given a set of interpretations that had been shuffled so as to mask any common author of interpretations across items.

## Results

We treated interpretation scores (0, 1, 2) as ordinal data, and assessed inter-rater reliability by computing linear-weighted Cohen's Kappa statistics between each pair of our three judges for human-generated interpretations and for model-generated interpretations separately. Each of these statistics confirmed pairwise inter-rater reliability for human-generated interpretations (judge 1 vs. judge 2: $\kappa = .57$, $z = 32.7$, $p < .001$; judge 1 vs. judge 3: $\kappa = .43$, $z = 26.0$, $p < .001$; judge 2 vs. judge 3: $\kappa = .41$, $z = 25.7$, $p < .001$) and for model-generated interpretations (judge 1 vs. judge 2: $\kappa = .30$, $z = 5.02$, $p < .001$; judge 1 vs. judge 3: $\kappa = .16$, $z = 2.29$, $p = .02$; judge 2 vs. judge 3: $\kappa = .11$, $z = 3.29$, $p < .001$). The reduced $\kappa$ associated with machine-generated interpretations, relative to human-generated interpretations reflects a discrepancy in the number of scores given to model-generated interpretations (110) and those given to human-generated interpretations (> 2000).

To analyze our data, we fit a cumulative link mixed model to trial-level interpretation scores (again, treated as ordinal data), using the *clmm* function from version 2022.11.16 of the ordinal R package (Christiansen, 2022) in R version 4.3.1 (R Core Team, 2021). We defined a full model including *participant*, *judge*, and *metaphor* as random intercept effects, and including two two-way interaction terms as fixed effects: *interpretation source* (*GPT-4* vs. *human*) x *metaphor form* (*canonical* vs. *reversed*) and *interpretation source x online match (yes vs. no)*. Figure 1 shows metaphor score data, broken down according to metaphor form, with GPT-4 coded as one participant.

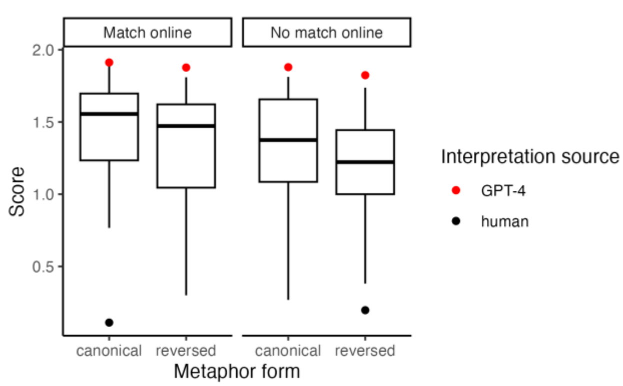


Figure 1: Metaphor interpretation scores (averaged across judges) broken down by metaphor form ("canonical" vs. "reversed") and whether a metaphor's English translation was found online published before 2022 ("Match online" vs. "No match online"). Human data are represented as a boxplot and GPT-4 performance is represented by red points.

Note that these analyses treated each unique rating (i.e., a separate rating from each judge) as a separate datapoint, and we modeled these ratings by included judge as a random intercept effect, which statistically controls for any variability introduced by judge idiosyncrasy.

All statistical tests involve comparing the full model described above with an ablated model that lacks a particular predictor of interest (e.g., coding source) but that is otherwise equivalent to the full model. We then use likelihood ratio tests to evaluate whether or not the ablated model yields increased prediction error compared to the full model: To the extent that it does, the omitted predictor was crucial to model performance and was thus an important predictor of the modeled data.

Removing the *interpretation source x metaphor form* interaction term did not increase model prediction error, $\Delta AIC = 2$, $\chi^2(1) = .01$, $p = .92$, and neither did removing the *interpretation source x online match* interaction term, $\Delta AIC = 2$, $\chi^2(1) = .05$, $p = .81$, nor did jointly removing both interaction terms, $\Delta AIC = 4$, $\chi^2(1) = .07$, $p = .97$. We thus did not detect any performance difference between GPT-4 and human participants that varied as a function of metaphor form or as a function of whether or not a given metaphor was found online pre-2022 (and thus possibly in GPT-4's training corpus). Inspecting the fit model that omitted these interaction terms, we found main effects for both interpretation source ($\beta = .194$, $z = 2.01$, $p = .045$), metaphor form ($\beta = .383$, $z = 7.50$, $p < .001$), and online match ($\beta = .379$, $z = 2.29$, $p = .022$). As is evident from Figure 1, GPT-4 outscored all human participants. The AI system and humans were both affected by the metaphor form, such that interpretations of metaphors in the reversed form received lower scores than those in the canonical form. Both the model and humans were also consistently affected by a metaphor's presence online. A metaphor being published online increases the probability that both GPT-4 and human reasoners may have encountered the expression prior to our experiment, providing a rough indication of its familiarity. Previous work has shown that more familiar metaphors tend to be comprehended more easily than those that are less familiar (Blasko & Connine, 1993; Stamenković, Ichien, et al., 2019).

When the metaphors were presented in the reversed (non-canonical) order of source and target (e.g., canonical: "A man is a butterfly" vs. reversed: "A butterfly is a man"), human participants often gave interpretations that restored the canonical order (50% of responses; "A man goes through many phases of growth in order to reach his full potential."). More generally, we coded the rate at which roles were switched for both canonical and reversed metaphors. As shown in Figure 2, switching was very rare (as would be expected) when the metaphor was presented in canonical form, but was very common when the metaphor was reversed (in which case switching restored the canonical interpretation). Restoring the canonical meaning of a reversed metaphor is consistent with previous findings concerning how people interpret reversed metaphor (Chiappe et al., 2003). This propensity likely reflects the Gricean cooperative principle, according to which people seek effective communication (Grice, 1975). Remarkably, GPT-4 gave interpretations of reversed metaphors that restored their canonical meaning at about the same rate (56%) as did humans (50%). Even more striking, the AI system tended to restore the canonical meaning for *the same specific metaphors* as did people (point-biserial correlation = 0.79, $p < .001$, across the 55 individual reversed metaphors). These findings suggest that in interpreting metaphors, GPT-4 resembles humans in its sensitivity to pragmatic constraints on communication.

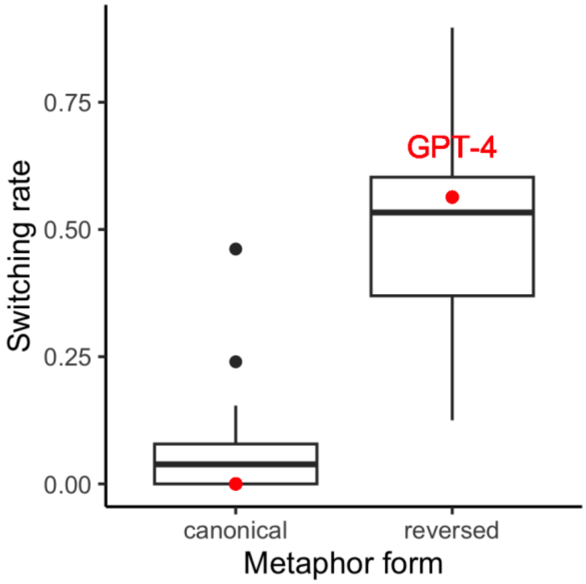


Figure 2: Metaphor switching rates (proportion of times an interpretation described the source of a metaphor, rather than its target) broken down by metaphor form (canonical vs. reversed). Human data are represented as a boxplot and GPT-4 performance is represented by red points.

## Discussion

The language abilities of a state-of-the-art large language model, GPT-4, extend to the interpretation of metaphors, the most prominent form of figurative language. We compared GPT-4 with humans using a challenging set of novel literary metaphors, generated by Serbian poets and translated into English. Regardless of whether one-off versions of the metaphors had made their way online and potentially into GPT-4's training corpus, the AI system produced metaphor interpretations that human judges (blind to the fact that an AI system was involved in the study) rated as superior to those written by *any* of the human participants—college students at a major public university in the United States. GPT-4 also exhibited a human-like propensity to "make sense" of metaphors presented in a non-canonical form (with source and target reversed). On about half the trials, both people and GPT-4 provided interpretations of reversed metaphors that restored their canonical meaning. Moreover, the AI system resembled humans at the level of individual reversed metaphors, reliably reflecting the probability that a human would restore the canonical meaning at the item level. Our findings add to recent evidence that large language models have begun to acquire some aspects of human pragmatic skills (Barattieri di San Pietro et al., 2023).

Although this study has established that GPT-4 can generate very sensible and human-like interpretations of novel literary metaphors, it leaves open the more fundamental question—how does it do it? Achieving scientific understanding of the operation of LLMs such as GPT-4 continues to be impeded by the refusal of their corporate owners to provide either a detailed account of their training data or access to the internal representations the systems have acquired. We used the words "interpret" and "interpretation" in our queries to GPT-4, but its responses to these close associates of "intelligence" and "thinking" certainly do not establish that this AI system "thinks" in the same way humans do.

Specifically with respect to metaphor, the ability to *generate interpretations of novel metaphor*s must not be confused with the ability to *generate novel metaphors*. Besides being able to generate interpretations of metaphors, as shown here, LLMs can certainly generate texts in which metaphors appear. However, to the best of our current knowledge, the metaphors an LLM might generate are limited to those that human writers have already formed and planted into texts, thereby making humanity's store of metaphors available to be mined by LLMs. It remains to be seen whether AI systems will at some point be able to create genuinely novel metaphors, rather than only variations of those we humans have made already. The great writer Jorge Luis Borges thought that truly new metaphors still await discovery, at least by humans. New variations of old metaphors can be very beautiful, he acknowledged, "and only a few critics like myself would take the trouble to say, 'Well, there you have eyes and stars and there you have time and the river over and over again.' The metaphors will strike the imagination. But it may also be given to us—and why not hope for this as well?—it may also be given to us to invent metaphors that do not belong, or that do not yet belong, to accepted patterns" (Borges, 2000, p. 41).

## References

Barattieri di San Pietro, C., Frau, F., Mangiaterra, V., & Bambini, V. (2023). *The pragmatic profile of ChatGPT: Assessing the pragmatic skills of a conversational agent*. https://psyarxiv.com/ckghw/

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Blasko, D. G., & Connine, C. M. (1993). Effects of familiarity and aptness on metaphor processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 295–308. https://doi.org/10.1037/0278-7393.19.2.295

Borges, J. L. (2000). *This craft of verse (the Charles Eliot Norton lectures 1967-1968)*. Harvard University Press.

Bowdle, B., & Gentner, D. (2005). The career of metaphor. *Psychological Review*, *112*(1), 193–216. https://doi.org/10.1037/0033-295X.112.1.193

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv*.

Chan, S. C., Santoro, A., Lampinen, A. K., Wang, J. X., Singh, A. K., Richemond, P. H., McClelland, J., & Hill, F. (2022). Data distributional properties drive emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*.

Chiappe, D., Kennedy, J. M., & Smykowski, T. (2003). Reversibility, aptness, and the conventionality of metaphors and similes. *Metaphor and Symbol*, *18*(2), 85–105. https://doi.org/10.1207/S15327868MS1802_2

Chomsky, N., Roberts, I., & Watumull, J. (2023). *The false promise of ChatGPT*. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Christiansen, R. H. B. (2022). *Ordinal—Regression Models for Ordinal Data*. https://CRAN.R-project.org/package=ordinal

Dasgupta, I., Lampinen, A. K., Chan, S. C., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2022). Language models show human-like content effects on reasoning. *arXiv Preprint*.

DiStefano, P. V., Patterson, J. D., & Beaty, R. (2023). Automatic scoring of metaphor creativity with large language models. *PsyArXiv*. https://doi.org/10.31234/osf.io/6jtxb

Frost, R. (1931). Education by poetry: A meditative monologue. In M. Richardson (Ed.), *The collected prose of Robert Frost*. Belknap Press of Harvard University Press.

Grice, P. (1975). Logic and conversation. In P. Cole & J. J. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). Academic Press.

Holyoak, K. J. (2019). *The spider's thread: Metaphor in mind, brain, and poetry*. MIT Press.

Holyoak, K. J., & Stamenković, D. (2018). Metaphor comprehension: A critical review of theories and evidence. *Psychological Bulletin*, *144*(6), 641–671. https://doi.org/10.1037/bul0000145

Jacobs, A. M., & Kinder, A. (2017). The brain is the prisoner of thought: A machine-learning assisted quantitative narrative analysis of literary metaphors for use in neurocognitive poetics. *Metaphor and Symbol*, *32*, 139–160. https://doi.org/10.1080/10926488.2017.1338015

Jacobs, A. M., & Kinder, A. (2018). What makes a metaphor literary? Answers from two computational studies. *Metaphor and Symbol*, *33*(2), 85–100. https://doi.org/10.1080/10926488.2018.1434943

Katz, A., Paivio, A., Marschark, M., & Clark, J. (1988). Norms for 204 literary and 260 nonliterary metaphors on 10 psychological dimensions. *Metaphor and Symbolic Activity*, *3*(4), 191–214. https://doi.org/10.1207/s15327868ms0304_1

Lakoff, G., & Turner, M. (1989). *More than cool reason: A field guide to poetic metaphor*. University of Chicago Press.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (In press). Dissociating language and thought in large language models: A cognitive perspective. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2024.01.011.

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schütze, H. (2020). Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, *117*(42), 25966–25974. https://doi.org/10.1073/pnas.1910416117

Milenković, K., Stamenković, D., & Tasić, M. (In press). Influence of translation on perceived metaphor features: Quality, aptness, metaphoricity and familiarity. *Linguistics Vanguard*.

OpenAI. (2023). *GPT-4 Technical Report*.

Piantadosi, S. T. (2023). *Modern language models refute Chomsky's approach to language*. https://lingbuzz.net/lingbuzz/007180

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., & others. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv Preprint*, *34*, 1877–1901.

Stamenković, D., Ichien, N., & Holyoak, K. J. (2019). Metaphor comprehension: An individual-differences approach. *Journal of Memory and Language*, *105*, 108–118. https://doi.org/10.1016/j.jml.2018.12.003

Stamenković, D., Ichien, N., & Holyoak, K. J. (2020). Individual differences in comprehension of contextualized metaphors. *Metaphor and Symbol*, *35*(4), 285–301. https://doi.org/10.1080/10926488.2020.1821203

Stamenković, D., Milenković, K., & Dinčić, J. (2019). Studija normiranja književnih i neknjiževnih metafora iz srpskog jezika [A norming study of Serbian literary and nonliterary metaphors]. *Zbornik Matice srpske za filologiju i lingvistiku*, *62*(2), 89–104.

Stamenković, D., Milenković, K., Ichien, N., & Holyoak, K. J. (2023). An individual-differences approach to poetic metaphor: Impact of aptness and familiarity. *Metaphor and Symbol*, *38*(2), 149–161. https://doi.org/10.1080/10926488.2021.2006046

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-023-01659-w

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., & others. (2022). Emergent abilities of large language models. arXiv:2206.07682