UCLA
UCLA Previously Published Works

Title

Large Language Model Displays Emergent Ability to Interpret Novel Literary Metaphors

Permalink

https://escholarship.org/uc/item/1mc135w2

Journal

Metaphor and Symbol, 39(4)

ISSN

1092-6488

Authors

Ichien, Nicholas
Stamenković, Dušan
Holyoak, Keith J

Publication Date

2024-10-01

DOI

10.1080/10926488.2024.2380348

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial License, available at https://creativecommons.org/licenses/by-nc/4.0/

Peer reviewed

eScholarship.org
Powered by the California Digital Library
University of California

Metaphor and Symbol

Routledge
Taylor & Francis Group

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hmet20

Large Language Model Displays Emergent Ability to Interpret Novel Literary Metaphors

Nicholas Ichien, Dušan Stamenković & Keith J. Holyoak

To cite this article: Nicholas Ichien, Dušan Stamenković & Keith J. Holyoak (2024) Large Language Model Displays Emergent Ability to Interpret Novel Literary Metaphors, Metaphor and Symbol, 39:4, 296-309, DOI: 10.1080/10926488.2024.2380348

To link to this article: https://doi.org/10.1080/10926488.2024.2380348

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.

View supplementary material

Published online: 01 Oct 2024.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=hmet20

METAPHOR AND SYMBOL
2024, VOL. 39, NO. 4, 296–309
https://doi.org/10.1080/10926488.2024.2380348

Routledge
Taylor & Francis Group

OPEN ACCESS
Check for updates

Large Language Model Displays Emergent Ability to Interpret Novel Literary Metaphors

Nicholas Ichien[a], Dušan Stamenković[b], and Keith J. Holyoak[c]

[a]University of Pennsylvania; [b]Södertörn University, Stockholm; [c]University of California, Los Angeles

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. Here we assessed the ability of GPT-4, a state-of-the-art large language model, to provide natural-language interpretations of a recent AI benchmark (Fig-QA dataset), novel literary metaphors drawn from Serbian poetry and translated into English, and entire novel English poems. GPT-4 outperformed previous AI models on the Fig-QA dataset. For metaphors drawn from Serbian poetry, 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. In addition, for several novel English poems GPT-4 produced interpretations that were rated as excellent or good by a human literary critic. These results indicate that LLMs such as GPT-4 have acquired an emergent ability to interpret literary metaphors, including those embedded in novel poems.

Introduction

The poet Robert Frost asserted, “The richest accumulation of the ages is the noble metaphors we have rolled up” (Frost, 1931/2007, p. 108). The world’s literature (Holyoak, 2019; Lakoff & Turner, 1989; Rasse, 2022; 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 (particularly those grounded in intense personal and emotional experiences; see Turner & Littlemore, 2023) 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). In artificial intelligence (AI), researchers are exploring the potential for artistic and literary forms of computational creativity (e.g., Veale et al., 2020; Zylinska, 2023). If 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.

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., 2024; Köbis & Mossink, 2021; Mahowald et al., 2024; McClelland et al., 2020) and various forms of

---

CONTACT Dušan Stamenković dusan.stamenkovic@sh.se School of Culture and Education, Södertörn University Stockholm, Alfred Nobels allé 7, Huddinge 141 89, Sweden

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10926488.2024.2380348

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

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). 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, suggesting that successful performance on these benchmarks might merely reflect reproduction of some of the data on which the models were trained, rather than a capacity for deep cognitive processing (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 a model'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). Unfortunately, detailed information about the training data for LLMs is at present unavailable, making it difficult to determine definitively whether any given metaphor is truly novel for the system. 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. Moreover, there is evidence that new metaphors generated by psychologists for experimental purposes are typically weak in aptness (Jones & Estes, 2006). In general, those metaphors that become conventional are those people have considered to be particularly apt.

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 grammar and high lexical diversity, and extra difficulty (for people) in comprehending the metaphorical meaning (Baggio, 2018; Jacobs & Kinder, 2017, 2018). Studies of individual differences in cognitive abilities have shown that crystalized intelligence (closely linked to verbal ability and the accumulation of knowledge and skills over time) impacts comprehension of both conventional and literary metaphors; fluid intelligence (the ability to reason and solve new problems independently of prior knowledge, 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.

We report three studies that examine the ability of GPT-4 to identify or generate interpretations of metaphors. Study 1 provides an extensive comparison between GPT-4 and previous state-of-the-art language models on a two-alternative force-choice metaphor comprehension task based on the Fig-QA dataset developed by Liu et al. (2022). Study 2 goes beyond previous work evaluating large language models on their metaphor comprehension ability (Liu et al., 2022; Prystawski et al., 2023) by testing stimuli drawn from poetry, thus consisting of much more naturalistic literary language. Finally, Study 3 assesses the quality of GPT-4's interpretations of poems written by published poets (but not available for GPT-4's training), as evaluated by a professional literary critic. The metaphors examined in Studies 2–3 contrast with the highly simplistic and formally stereotyped materials designed for use in previous psychology experiments and tests of AI models.

Study 1: interpreting metaphors in fig-QA dataset

Background

Recent work has evaluated a number of LLMs (GPT-2, GPT-neo, GPT-3, BERT, and RoBERTa) on their ability to generate interpretations of metaphors (Liu et al., 2022). In a forced-choice paradigm, models and humans were given a metaphor from the Fig-QA

Table 1. Model and human performance on fig-QA dataset. Bolded row reflects GPT-4's performance, which is original to the current paper. All other data was originally presented in Liu et al. (2022). Note that the "fine-tuned" column corresponds to "fine-tuned (XL)" in Liu et al.'s table 4, presenting model performance after fine-tuning on a large dataset of 8,016 metaphors. "Human (confident)" indicates human performance after excluding any problems for which participants reported lack of confidence in their responses, and "human" indicates overall human performance. This distinction between human (confident) and human was introduced in Liu et al.'s original study.

Model	Zero-shot	Fine-tuned
GPT-2	53.93	62.65
GPT-neo 1.3B	56.89	72.00
GPT-3 Ada	59.08	73.56
GPT-3 Babbage	62.91	77.31
GPT-3 Curie	65.35	81.94
GPT-3 Davinci	68.41	-
BERT	58.14	85.69
RoBERTa	66.18	90.32
GPT-4	**91.27**	-
Human	94.42	-
Human (confident)	95.39	-

dataset (e.g., "The clothing has the smell of a crowded gym") and selected one of two literal rephrases of that metaphor (e.g., "The clothing smells pleasant" vs. "The clothing smells awful"). Humans made their selection directly, whereas models did so implicitly by determining which rephrase yielded a higher value on some metric of sentence plausibility (e.g., log probability for GPT models) when plugged into the following form: "<METAPHOR>, that is to say, <REPHRASE>" (e.g., "The clothing has the smell of a crowded gym, that is to say, the clothing smells awful").

Across the 1,146 metaphors in the test set of Liu et al.'s Fig-QA dataset (https://github.com/nightingal3/Fig-QA; leaderboard can be found at: https://explainaboard.inspiredco.ai/leaderboards?dataset=fig_qa), human participants did very well, achieving 94.42% accuracy (95.39% after excluding any problems for which participants reported lack of confidence in their response). Table 1 depicts performance on the Fig-QA dataset for both human participants and models. Of the models that Liu et al. evaluated, GPT-3 Davinci (Brown et al., 2020; Radford et al., 2019), the largest model (175B parameters, trained on 45TB of text) achieved the best zero-shot performance (i.e., performance achieved without any dedicated training on similar problems) (68.41%) but still fell well-short of human-level responding. RoBERTa came close to human-level responding (90.32%), but only after it was fine-tuned (i.e., tested after pre-training) on 8,016 similar metaphors (Liu et al., 2022).

Method and results

We adapted code from the Fig-QA website (https://github.com/nightingal3/Fig-QA) to test GPT-4 zero-shot by directly prompting the model with natural language text using the following prompt form: "Which of the following is a better interpretation of the following expression: <METAPHOR>. A: <REPHRASE 1> Or B: <REPHRASE 2>?." Such prompts constitute input to GPT-4 and LLM's more generally, in response to which the models produce text output. The model provided an unambiguous response (e.g., "B. The clothing smells awful") most of the time, but on 3.35% of questions, it responded with some variant of "Neither A nor B is a direct interpretation of <METAPHOR>." Counting the latter responses as incorrect, along with those for which the model responded with the wrong interpretation, GPT-4 outperformed both GPT-3 Davinci and the fine-tuned RoBERTa, achieving close to human-level performance (91.27%). Thus

GPT-4 far outperformed other LLMs, and even bested a fine-tuned variant of RoBERTa. Due to budget constraints (similar to those mentioned in Liu et al. (2022), with respect to GPT-3 Davinci), we did not fine-tune GPT-4.

Discussion

A couple of differences between GPT-4 and the models tested in Liu et al. (2022) may explain its superior performance. First, GPT-4 is considerably larger (1.8 trillion parameters, trained on 1 PB of text and images) than GPT-3 Davinci, the largest model Liu et al. tested (OpenAI, 2023). Second, GPT-4 was trained not only on text, but also computer code and images. It also incorporated reinforcement learning with human feedback (RLHF). In this learning procedure, a "reward model" is given a set of text prompts and responses that human judges had independently rated for quality, and is then trained to ultimately predict human-judged quality of text responses. A language model, previously trained to predict the sequence of text in large training corpora, is then trained to produce text responses that maximize their quality, as predicted by the reward model. These additional training steps served to further align the model's text output with that generated by humans and rated by them as highly appropriate.

The metaphors used in the Fig-QA dataset were all crowdsourced from MTurk workers and the dataset was not published until June 23, 2022—later than September 2021, the latest date that the data used to train GPT-4 were collected (OpenAI, 2023). It is therefore unlikely that model performance solely reflects GPT-4's ability to reproduce its memorized training data. However, metaphors generated by MTurk workers are unlikely to be entirely original. Those included in the Fig-QA dataset were clearly easy for humans to reliably interpret, as they achieved nearly ceiling performance. Moreover, though Fig-QA items were labeled by the original authors as metaphors, many might be more accurately characterized as similes (as is evident from a sampling of metaphors in the Fig-QA dataset; e.g., "The clothing has the smell of a crowded gym;" "The math test was as straightforward as building a rocket ship;" "The government plan is as expensive as a hundred mansions"). Accordingly, in Studies 2–3 we examined proper metaphors drawn from literary sources, and that were unlikely to have been included in datasets used to train GPT-4.

Study 2: interpreting metaphors drawn from Serbian poetry

Overview

Study 2 was performed using materials designed to minimize the possibility that GPT-4 had been trained on our test metaphors and could produce interpretations based on memory for these specific examples. Specifically, we selected 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. We used English translations of the original Serbian metaphorical expressions for two main reasons. First, translating them into English added an extra layer of separation from any potential recognition by the model based on its training data. Second, ChatGPT performs significantly better in English compared to Serbian.

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 2 for additional examples). We decided to use both forms to allow for comparisons with previous studies (e.g., Chiappe et al., 2003).

Table 2. 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.

Method

Materials

The test set included 55 literary metaphors drawn from Serbian poetry and normed on several properties (Milenković et al., 2024; 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 (e.g., *A woman is a flower*), 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). 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., 2024) 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 (see Table S1 in Section S1 of Supplemental Online Information).

These Serbian translations were selected to minimize the likelihood that their English translations appeared in GPT-4's training data. 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; 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 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.)

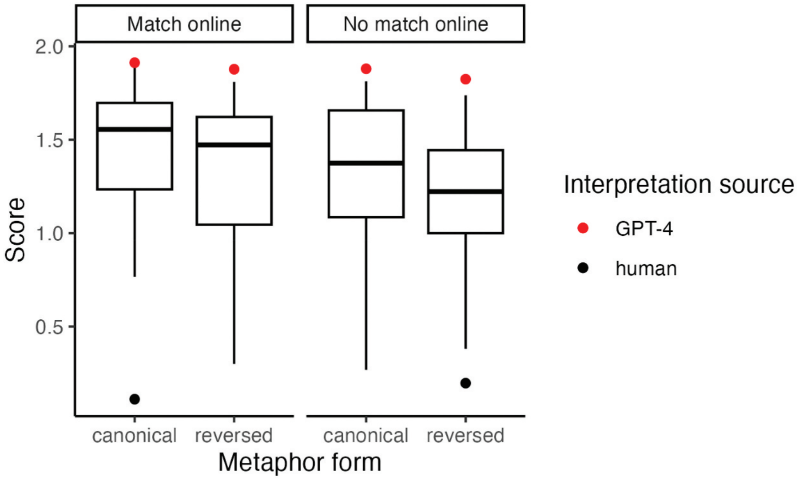

Figure 1. Metaphor interpretation scores (averaged across judges) broken down by metaphor form ("canonical" vs. "reversed") and whether a metaphor's English translation was found online published before 2022 ("match online" vs. "No match online"). Human data are represented as a boxplot and GPT-4 performance is represented by red points. Black points indicate outliers among human participant scores that lie outside of 1.5 * the interquartile range. Boxplots reflect the full distribution of human performance across our sample, whereas the single red point indicates GPT-4's performance (analogous to a single human participant).

Participants

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 ($Mage = 21.31$, $SD_{age} = 5.07$, $range_{age} = [18,47]$; 31 female, 7 male, 1 gender not reported) 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).

Procedure

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 (e.g. "Please provide an interpretation for the following expression: Old age is winter.") 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.

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 (i.e., uninformed by any previous examples).

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. Judges were instructed not to penalize interpretations because they reverse what the sentence appears to describe. 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. Because GPT-4 always provided an interpretation for each metaphor, only about .02 of its received scores were 0, compared to .22 of scores for human participants. For comparison, in Section S3 of Supplementary Online Materials we provide an analysis of interpretation scores after omitting scores of 0.

Results and discussion

Table 2 presents examples of metaphor interpretations generated by GPT-4 and by humans. 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 κ 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 (0, 1, 2 treated as an ordinal variable), 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 (translation published online before 2022 vs. not).* Figure 1 shows metaphor score data, broken down according to metaphor form, with GPT-4 coded as one participant.

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

Removing the *interpretation source x metaphor form* interaction term did not increase model prediction error, $\Delta\mathrm{AIC} = 2$, $\chi^2\ (1) = .01$, $p = .92$, and neither did removing the *interpretation source x online match* interaction term, $\Delta\mathrm{AIC} = 2$, $\chi^2\ (1) = .05$, $p = .81$, nor did jointly removing both interaction terms, $\Delta\mathrm{AIC} = 4$, $\chi^2\ (1) = .07$, $p = .97$. 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 ($\beta = 1.94$, $z = 2.01$, $p = .045$), 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

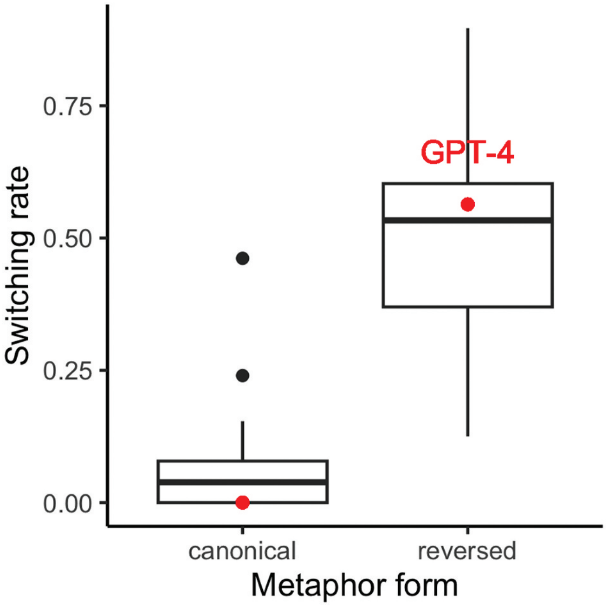

Figure 2. Metaphor switching rates (proportion of times an interpretation described the source of a metaphor, rather than its target) broken down by metaphor form (canonical vs. reversed). Human data are represented as a boxplot and GPT-4 performance is represented by red points. Black points reflect outliers among human participant scores, lying outside of 1.5 * the interquartile range. Boxplots reflect full distribution of human performance across our sample, whereas single red point reflect GPT-4's performance, which is analogous to a single human participant.

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.

Study 3: interpreting novel English poems

Overview

A yet more challenging task is to interpret entire poems, which often include metaphors at multiple levels (e.g., an entire poem may itself be a metaphor), coupled with additional poetic devices involving the interplay of symbolic meaning (e.g., synecdoche, metonymy) and sound (e.g., meter, rhyme, alliteration) (see Rasse, 2022). As an initial qualitative test, we asked GPT-4 to interpret eight short English poems that had not been published at the time the system's training was completed (six have never been published at all). A human literary critic (uninformed about the involvement of an AI system) then evaluated the accuracy and quality of these interpretations. We also asked the critic to evaluate two poems generated by GPT-4.

Method

Materials

We obtained access to eight short poems (10–29 lines in length), six unpublished and two published since 2022. Three poems were written by each of two contemporary authors, Koon Woon (author of two poetry books, one of which won a PEN Oakland/Josephine Miles Literary Award) and Keith Holyoak (author of four poetry books, and an author of this paper), one poem by Amanda Gorman (former National Youth Poet Laureate and author of multiple poetry books), and one by Jay Parini (author of numerous books spanning poetry, novels, biography, screenplays and criticism). The selected poems (see Section 4 of Supplemental Online Materials) vary in content and style, including both free and formal verse; all incorporate some form of metaphor as well as additional literary devices.

Assessments of GPT-4's interpretations by a literary critic

We first gave each poem to GPT-4 without providing the title or author and asked the system to identify it. In each case GPT-4 reported that it was unable to identify the author or source of the poem (as would be expected given that none of the poems were available on the internet when GPT-4 was trained). For each, we then asked GPT-4 to "please give your interpretation of the poem." Each poem was presented in a new chat.

One of the authors of poems that we used (Koon Woon, who also has experience as an editor and publisher of both a poetry journal and an independent literary press) agreed to provide assessments of these interpretations. This literary critic was blind to the origin of the interpretations and was not told that an AI system was involved. The critic first assessed the interpretations provided for the three poems he himself had authored, and then for the remaining poems. We asked him to "take on the role of literary critic or English professor." He was asked to address "whether the interpretation captures the essence of what you see as the meaning of the poem," and to briefly "indicate whatever you think is a major strength or major weakness of the interpretation." Finally, the critic was asked to give each interpretation a letter grade, using the scale: A (excellent), B (good), C (adequate), D (deficient), E (poor). All responses by GPT-4 and by the critic are provided in Section S4 of Supplemental Online Materials. We found it particularly valuable to have the critic evaluate his own poems, as he possessed a uniquely clear understanding of their intended meaning.

Results and discussion

The human critic assigned an overall assessment of "excellent" to the first seven interpretations he evaluated, and "good" to the eighth and final one. For all interpretations, the human critic stated that it accurately expressed the symbolic meaning of the poem (e.g., "The critic of this poem gets it"). (For all poems, GPT-4's interpretations, and assessments by the human critic, see Section S4 of Supplemental Online Materials.) For example, GPT-4 correctly pointed out that the poem "A Skimming Stone" employs the metaphor of a skimming stone to reflect on life, time, and mortality. Moreover, in response to a one-off follow-up query, "Would you say the speaker of the poem is analogous to the child, or to the skimming stone?" GPT-4 reported that "the speaker shifts from being the observer (the child) to being the observed (the stone). This transition signifies the realization of the speaker's own journey through life, mirroring the stone's trajectory." This dynamic metaphoric shift as described by GPT-4 is fully consistent with the poet's intended meaning (as confirmed by Holyoak).

Although the human critic attested to the accuracy of GPT-4's interpretations, he also expressed important reservations related to emotional sensitivity. After finishing all eight assessments, the human critic spontaneously provided an overall characterization of the "critic" being evaluated (i.e., GPT-4): "The critic in all these interpretations zones in on the themes and meanings and their interconnections, but he is weak on evocation portraying just exactly how the reader feels the poem . . . . This suggests to me that the 'critic' relies on a formula and perhaps is an AI program. It just seems to be without any flair at all even though as far as the straightforward features of a poem goes, it is all correct. It is like a meal that looks done right, even beautifully appealing, but without zesty taste."

General discussion

Summary

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. In Study 1 we compared GPT-4 and state-of-the-art language models on a standardized metaphor comprehension task (Liu et al., 2022), and found that GPT-4's performance was substantially more accurate than that achieved by earlier AI models. In Study 2 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). In Study 3, we found that GPT-4 also produced highly accurate interpretations of novel English poems, exhibiting the ability to analyze larger-scale literary metaphors.

Limitations of current study

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 lack of access to 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. The caveat raised by the human literary critic with respect to GPT-4's interpretations of poems – an apparent lack of emotional sensitivity – suggest one possible gap between the system's ability to generate meaningful interpretations and the way a human understands literary texts.

The ability to interpret novel literary metaphors is a remarkable achievement for AI systems. However, it seems premature to conclude that these systems are at present capable of the full range of literary creativity. Specifically with respect to metaphor, the ability to *generate interpretations of novel metaphors* 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 (Holyoak, 2019). 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).

Acknowledgments

A preliminary report of part of this study was presented at the 46th Annual Meeting of the Cognitive Science Society, held in July, 2024, at Rotterdam. We thank Koon Woon for serving as our literary critic in Study 3. We are grateful to both Koon Woon and Jay Parini for providing us with unpublished poems for use in this study. We also thank Rita Debbaneh, Jocelyn Leison, and Kaycee Stiemke for scoring metaphor interpretations in Study 2. Preparation of this paper was supported by a Google Faculty Research Award to K.J.H. Open-access publication was made possible with support from Södertörn University, Stockholm.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Preparation of this paper was supported by a Google Faculty Research Award to Keith J. Holyoak.

Author contributions

Conceptualization, K.J.H.; Methodology, D.S., K.J.H., N.I.; Formal Analysis, N.I.; Investigation, D.S., N.I.; Writing – Original Draft, D.S., K.J.H., N.I.; Writing – Review & Editing, D.S., K.J.H., N.I.; Supervision, K.J.H.; Funding Acquisition, K.J.H.

References

Baggio, G. (2018). *Meaning in the brain*. MIT Press.

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. https://doi.org/10.1073/pnas.2218523120

Blasko, D. G., & Connine, C. M. (1993). Effects of familiarity and aptness on metaphor processing. *Journal of Experimental Psychology: Learning, Memory & 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

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., , and Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

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. https://doi.org/10.48550/arXiv.2303.12712

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*. https://doi.org/10.48550/arXiv.2205.05055

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

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. https://doi.org/10.48550/arXiv.2207.07051

DiStefano, P. V., Patterson, J. D., & Beaty, R. (2024). Automatic scoring of metaphor creativity with large language models. *Creativity Research Journal*. https://doi.org/10.1080/10400419.2024.2326343

Frost, R. (1931/2007). Education by poetry: A meditative monologue. In M. Richardson (Ed.), *The collected prose of Robert Frost* (pp. 102–111). 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*(3), 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

Jones, L. L., & Estes, Z. (2006). Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. *Journal of Memory and Language*, *55*(1), 18–32. https://doi.org/10.1016/j.jml.2006.02.004

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

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, *114*, 106553. https://doi.org/10.1016/j.chb.2020.106553

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

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

Liu, E., Cui, C., Zheng, K., & Neubig, G. (2022). Testing the ability of language models to interpret figurative language. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Hybrid: Seattle, Washington + Online, July 10–15, 2022 (pp. 4437–4452).

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models: A cognitive perspective. *Trends in Cognitive Sciences*, *28*(6), 517–540. 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., Tasić, M., & Stamenković, D. (2024). Influence of translation on perceived metaphor features: Quality, aptness, metaphoricity and familiarity. *Linguistics Vanguard*, *0*. https://doi.org/10.1515/lingvan-2023-0086

OpenAI. (2023). *GPT-4 technical report*.

Prystawski, B., Thibodeau, P., Potts, C., & Goodman, N. D. (2023). Psychologically-informed chain-of-thought prompts for metaphor understanding in large language models. arXiv:2209.08141. https://doi.org/10.48550/arXiv.2209.08141

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Rasse, C. 2022. *Poetic Metaphors: Creativity and interpretation*. Amsterdam: John Benjamins.

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., & Kluska, A. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv. 2206.04615. https://doi.org/10.48550/arXiv.2206.04615

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

Steen, G. (1994). *Understanding metaphor in literature: An empirical approach*. Longman Pub Group.

Turner, S., & Littlemore, J. (2023). Literal or metaphorical? Conventional or creative? Contested metaphoricity in intense emotional experiences. *Metaphor & the Social World*, *13*(1), 37–58. https://doi.org/10.1075/msw.00028.tur

Veale, T., Pérez, Y., & Pérez, R. (2020). Leaps and bounds: An introduction to the field of computational creativity. *New Generation Computing*, *38*(4), 551–563. https://doi.org/10.1007/s00354-020-00116-w

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541. 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., & Chi, E. H. (2022). Emergent abilities of large language models. *arXiv Preprint*.

Zylinska, J. (2023). Art in the age of artificial intelligence. *Science*, *381*, 139–140. https://doi.org/10.1126/science.adh0575