Large Language Models provide support for the parallelogram theory of analogy

Qiawen Ella Liu[1], Raja Marjieh[1], Jian-Qiao Zhu[1,2], Adele E. Goldberg[1], and Thomas L. Griffiths[1]
[1]Princeton University
[2]The University of Hong Kong

Abstract

Four-term word analogies (A:B::C:D) are classically modeled geometrically as parallelograms: adding the vector $B-A+C$ produces D. Recent work suggests that this model poorly captures how humans produce analogies, with simple local-similarity heuristics often providing a better account (Peterson, Chen, & Griffiths, 2020). But does the parallelogram model fail because it is a bad model of analogical relations, or because people are not very good at generating relation-preserving analogies? We compared human and large language model (LLM) analogy completions on the set of problems from Peterson et al. (2020). We find that LLM-generated analogies are reliably judged as better than human-generated ones, and are also more consistent with parallelograms in a distributional embedding space. Crucially, we show that the improvement over human analogies is driven by greater parallelogram alignment and reduced reliance on accessible words rather than enhanced sensitivity to local similarity. Finally, fine-tuning GloVe to better satisfy the parallelogram constraint makes the model's top-ranked candidates more likely to be the completions humans and LLMs actually produced, and improves its prediction of human ratings. Overall, these results provide support for the parallelogram model of word analogies.

Introduction

Analogy has been regarded as a central feature of human creativity and intelligence (Gentner, 1983; Hofstadter, 2001; Holyoak & Thagard, 1995), making investigation of how analogies are formed and why some analogies are better than others a central issue in cognitive science. A simple but canonical paradigm for studying analogy is the four-term word analogy (e.g., *king is to queen as man is to woman*, abbreviated as *king*:*queen*::*man*:*woman*). A classic theory of this type of analogy is the parallelogram model, which posits that concepts exist as points in a geometric mental space where relations are represented as vectors (Rumelhart & Abrahamson, 1973): to find the word that

This study was preregistered on AsPredicted (https://aspredicted.org/2nk8e9.pdf). Correspondence concerning this article should be addressed to Qiawen Ella Liu, Department of Psychology, Princeton University, Princeton, NJ 08540. Email: ella.qiawen.liu@princeton.edu.

completes the analogy A:B::C:?, one applies the vector difference between B and A to the third term, C.

Recent behavioral evidence has, however, called the parallelogram model into question. Peterson et al. (2020) found that for human-generated word analogies the parallelogram model was outperformed by simple local-similarity heuristics that did not take relational similarity into account (e.g., by choosing a word highly similar to the C term). These findings appear to suggest that geometric models fail to capture how people represent and reason about analogies. Yet an alternative possibility is that people are simply not very good at *producing* analogies. Generating a precise analogy may be cognitively demanding due to time pressure, knowledge constraints, and retrieval failures, which may drive people toward more accessible but less relationally aligned responses, even when they are capable of *recognizing* better analogies when presented with them.

The emergence of large language models (LLMs) offers an opportunity to revisit this question. Despite being highly opaque systems with complex attention mechanisms and inaccessible latent representations, LLMs have shown a surprising capacity to simulate human semantic judgments (Piantadosi et al., 2024) and demonstrate emergent relational reasoning manifested in tasks like in-context learning (Brown et al., 2020). Crucially, LLMs may not face the same cognitive constraints that may hinder human analogy production. We therefore ask whether LLM-generated analogies are perceived to be higher quality than human-generated ones and whether any such advantage is captured by simple cognitive models such as the parallelogram model or local similarity heuristics. To address these questions, we prompt six state-of-the-art LLMs with a large set of word analogy problems previously given to humans (Peterson et al., 2020) and collect human judgments of how well LLM and human completions preserve the intended relations.

Our results show that LLM analogies are more predictable than people's by both parallelogram *and* local similarity measures. We also find that people reliably judge LLM analogies to be better than human-generated analogies and that the advantage of LLMs over humans is predicted by the use of lower-frequency words and greater parallelogram alignment in an external distributional embedding space. The LLM advantage is driven not by LLMs producing uniformly superior responses, but by humans generating a long tail of low-quality completions. When the comparison is restricted to only the most frequent (modal) responses, the LLM advantage disappears. Preferred analogical completions in this subset of cases, whether generated by LLMs or people, are still predicted by greater parallelogram-alignment and the use of lower-frequency words. Finally, to test whether the parallelogram structure drives more appropriate analogies rather than the effects being due to the particular embedding space we use, we fine-tune the embedding space to better satisfy the parallel structure. Doing so makes the parallelogram model better at recovering the more highly ranked responses from humans and LLMs. Together, these results suggest that while the parallelogram model may poorly describe how humans *generate* analogies, it nonetheless captures the way people *evaluate* good analogies—and that LLMs are better at producing responses that satisfy this structure.

Background

Geometric models of analogy

Early work by Rumelhart and Abrahamson (1973) proposed that concepts can be represented as points in a multidimensional space and that relations correspond to difference vectors between those points. Using low-dimensional representations derived from multidimensional scaling, they

showed that solving an analogy A:B::C:? (e.g., *rat*:*pig*::*goat*) could be modeled as completing a parallelogram to find a solution at $B - A + C$.

The advent of modern word embeddings has renewed interest in the parallelogram model. Methods such as word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and GloVe (Pennington, Socher, & Manning, 2014) learn vector representations of words from large text corpora, and early demonstrations showed that linear vector arithmetic could recover canonical analogies (e.g., *king* − *man* + *woman* ≈ *queen*). Similar geometric reasoning has also been explored in computer vision, where latent representations support vector-based transformations between visual concepts (Radford, Metz, & Chintala, 2015; Reed, Zhang, Zhang, & Lee, 2015).

The success of vector arithmetic on word analogies, however, is somewhat fragile. The predicted vector $B - A + C$ often lies closest to one of the input words, so the canonical answer surfaces only because those inputs are excluded from the candidate pool (Linzen, 2016). More generally, canonical responses tend to be predictable from the similarity among A, B, and C rather than on the relation per se (Rogers, Drozd, & Li, 2017). Simple heuristics that rely only on local similarities offer potential alternative ways people may generate analogical completions. Sadler and Shoben (1993) suggested a *similarity heuristic* that simply ranks responses by their similarity to C, and another nearest-neighbor (NN) heuristic that uses the relative similarity of A to B versus C to decide whether to retrieve a response near B or near C.

Evaluating the parallelogram rule against both heuristics on a large dataset of human-generated analogies, Peterson et al. (2020) found that while the parallelogram rule tended to capture the top human responses, similarity-based heuristics provided a better account of the full distribution of responses. As a consequence, they concluded that the parallelogram model may not be a good model of human analogy generation.

Analogies by large language models

A growing body of work suggests that LLMs can solve a range of analogy tasks, from four-term word analogies to matrix reasoning and letter-string analogies, and can support plausible open-ended analogical reasoning when prompted to do so (Johnson et al., 2025; Webb, Holyoak, & Lu, 2023; Yasunaga et al., 2023). Webb et al. (2023), for example, reported that LLMs match or even exceed average human accuracy on standardized multiple-choice analogy tasks (Raven's Progressive Matrices, letter-string analogies, four-term verbal analogies) (but see Lewis and Mitchell (2024) for evidence that these abilities may not generalize beyond familiar task formats). More broadly, LLMs are shown to be good at *in-context learning*: given a few examples, LLMs can infer the latent relationship linking inputs to outputs and apply it to novel cases (Brown et al., 2020). Recent theoretical explanations for in-context learning have grounded these capabilities in neural geometry, demonstrating that task-specific relational information can be extracted as "function vectors" from a model's internal activations on a few-shot prompt (Hendel, Geva, & Globerson, 2023; Todd et al., 2023).

Taken together, these findings suggest that LLMs may implement geometric strategies during relational reasoning reminiscent of the parallelogram model proposed by Rumelhart and Abrahamson (1973). Whether such structure resides in the models' own representations, or whether LLMs produce outputs that conform to geometric regularities in the semantic spaces used to evaluate them, has not been investigated. Prior work has also not systematically compared the full distributions of LLM and human responses on a large open-ended generation task, nor evaluated how well existing geometric models account for those distributions. Here, we provide such a comparison and analysis.

Do LLMs Produce Better Analogies than Humans?

We compare LLM and human completions to the same set of analogy problems (*A*:*B*::*C*:?), asking whether LLMs produce *different* analogies and then whether the differences are *improvements*. To do so, we generate completions from six LLMs, characterize how their responses differ from humans' in semantic space, and collect human judgments of analogy quality for both populations.

Methods

Generating LLM analogies

We prompted six LLMs to complete word analogies, including two closed-source models (`GPT-5-mini` (OpenAI, 2026), `o4-mini` (OpenAI, 2025)) and four open-source models (`DeepSeek-V3.1-671B` (DeepSeek-AI, 2024), `Qwen3-235B-A22B-Instruct-2507` (Yang et al., 2025), `Qwen3-32B` (Yang et al., 2025), `Gemma-3-27B` (Gemma Team, 2025)). To minimize artifacts from specific prompt wording, we used four semantically equivalent prompt phrasings for each analogy.[1] Each was repeated 10 times per analogy, yielding 40 responses per model per analogy. For open models, we used decoding parameters that allow the model to sample with substantial variability rather than always returning a single most-likely answer: `temperature = 1.0`, which leaves the model's next-token distribution unchanged, and `top_p = 0.9`, meaning that at each step the model considers only the most probable candidate words whose probabilities together add up to 90%. The two OpenAI reasoning models (`GPT-5-mini`, `o4-mini`) were called using the provider's default decoding configuration. We tested LLMs on all 846 analogy items from Peterson et al. (2020), drawn from Green, Kraemer, Fugelsang, Gray, and Dunbar (2010) (80 analogies), Kmiecik and Morrison (2013) (178 analogies), and Jurgens et al. (2012) (588 analogies annotated with the 20 SemEval relations summarized in Table 1). This yielded 203,040 LLM responses comprising 5,943 distinct completions; the original Peterson et al. (2020) dataset contributes 26,265 human responses (9,136 distinct completions).

Table 1

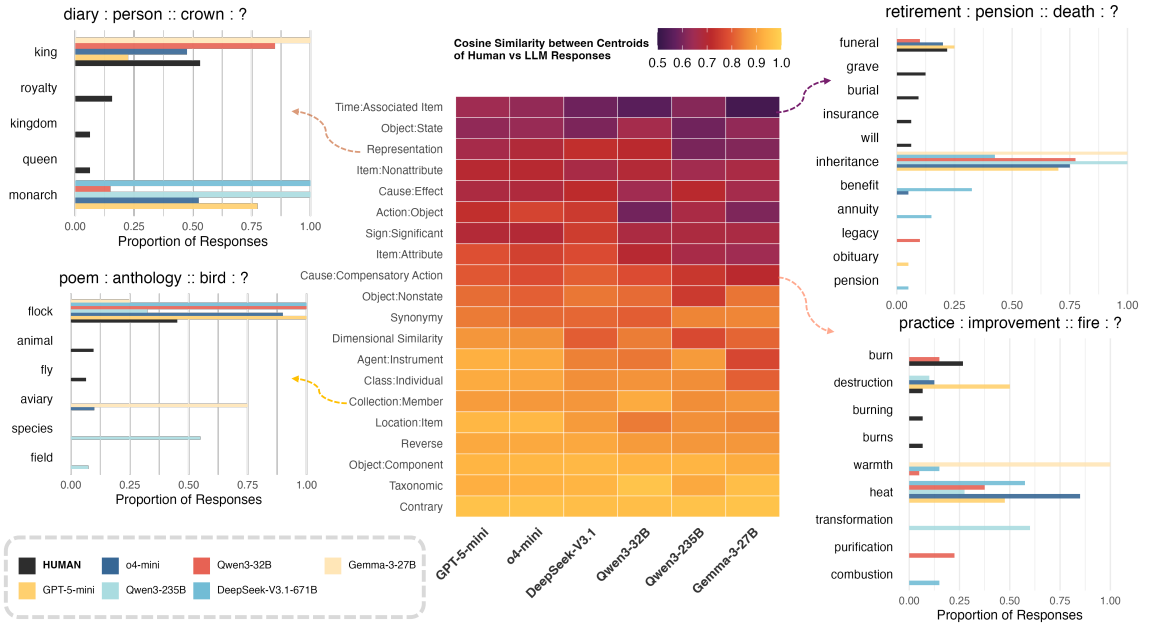
Semantic relation categories included from SemEval-2012 Task 2 (Jurgens et al., 2012).

Category	Subtype 1	Subtype 2
Class-Inclusion	Taxonomic (*flower*:*tulip*)	Class:Individual (*queen*:*Elizabeth*)
Part-Whole	Object:Component (*car*:*engine*)	Collection:Member (*forest*:*tree*)
Similar	Synonymy (*car*:*auto*)	Dimensional Similarity (*simmer*:*boil*)
Contrast	Contrary (*old*:*young*)	Reverse (*attack*:*defend*)
Attribute	Item:Attribute (*beggar*:*poor*)	Object:State (*beggar*:*poverty*)
Non-Attribute	Item:Nonattribute (*harmony*:*discordant*)	Object:Nonstate (*laureate*:*dishonor*)
Case Relations	Agent:Instrument (*farmer*:*tractor*)	Action:Object (*plow*:*earth*)
Cause-Purpose	Cause:Effect (*joke*:*laughter*)	Cause:Compensatory Action (*hunger*:*eat*)
Space-Time	Location:Item (*arsenal*:*weapon*)	Time:Associated Item (*retirement*:*pension*)
Reference	Sign:Significant (*siren*:*danger*)	Representation (*portrait*:*person*)

[1] (1) "Complete this analogy: A is to B as C is to ___"; (2) "What word completes this analogy? A:B::C:?"; (3) "Analogy: A:B::C:? Answer:"; (4) "Fill in the blank: A is to B as C is to _____".

Figure 1

Human–LLM similarity varies across semantic relations. ***Center:*** *cosine similarity between frequency-weighted centroids of human and LLM analogy completions across the 20 SemEval-2012 semantic relation types. Darker colors indicate greater similarity.* ***Insets:*** *example response distributions for humans and LLMs (bar lengths = proportion of responses). Black bars indicate proportion of each response provided by people.*



Collecting human ratings

We recruited 390 participants via Prolific (8–10 raters per analogy), all adult native English speakers from the United States who provided informed consent prior to participation in accordance with a protocol approved by the Princeton University Institutional Review Board (protocol #10859); the study was preregistered on AsPredicted (link). We constructed 4,048 four-term analogies (A:B::C:D) using A:B::C stems from the SemEval-2012 Task 2 dataset (Jurgens et al., 2012) and D terms drawn from human responses in Peterson et al. (2020) and LLM-generated completions to the same items. For both humans and LLMs, we excluded responses provided by only one participant or model, a standard practice in response-generation tasks (Nelson, McEvoy, & Schreiber, 2004). Stimuli were randomly distributed across 42 lists of about 100 analogies each. Participants rated how similar the relationship between C and D was to the relationship between A and B on a 7-point scale (1 = extremely different, 7 = extremely similar). Instructions included examples of similar relationships (reasonable analogies) (*kitten*:*cat*::*chick*:*chicken*) and dissimilar relationships (*chick*:*chicken*::*hen*:*rooster*) (poor analogies). Each list included 5 attention-check trials of obviously dissimilar comparisons that expect ratings ≤ 4 (e.g., *dog*:*puppy*::*happy*:*sad*); 13% of participants who failed more than half of these items were excluded from analyses. The remaining participants ranged in age from 18 to 81 years (M = 44.21, SD = 12.33); 174 identified as female, 157 as male, 3 as non-binary, and 3 preferred not to say.

Results

LLMs and humans produce different analogies

To analyze the degree to which LLM responses resemble human responses, we represented words using 300-dimensional GloVe embeddings trained on the 840B-token Common Crawl corpus (Pennington et al., 2014) and compared where their respective responses fall in semantic space. For each analogy stem, we computed a frequency-weighted centroid of responses in embedding space. Let r index unique responses with embeddings $\mathbf{v}_r$ and relative frequencies f_r. The centroid is $\mathbf{c} = \sum_r f_r \mathbf{v}_r$. We computed separate centroids for humans and each model and measured their cosine similarity; higher similarity indicates that humans and LLMs generated semantically similar responses, even if the exact words differed (e.g., *cat vs cats*). To contextualize the degree of human-LLM convergence, we also computed human–human convergence by randomly splitting the human responses in half for each stem and measuring the cosine similarity between the centroids of the two halves, repeated 1000 times.[2] Across all 846 analogy stems, humans and LLMs produced responses that were broadly similar but reliably distinguishable. The average centroid similarity between randomly split human samples is .898 (SE = .004 across the 846 stems). Every human–LLM convergence fell reliably below this value ($ps < 10^{-6}$), suggesting that no model's completions resemble human completions as closely as human samples resemble themselves. The two closed-source OpenAI models converged most closely with human completions: `GPT-5-mini` (mean human–model centroid similarity = .825) and `o4-mini` (mean = .824). Among open-source models, `DeepSeek-V3.1-671B` aligned most closely with humans (mean = .817), followed by `Qwen3-235B` (.798), `Qwen3-32B` (.797), and `Gemma-3-27B` (.781). The divergence is also different by relation type (Figure 1): some relations (e.g., *Contrary*, *Taxonomic*) elicit highly consistent completions across the two populations, whereas others (e.g., *Time:Associated Item*, *Object:State*, *Representation*) diverge sharply.

The divergence between LLMs and humans also suggests that it is unlikely that LLMs were just copying from their training data. Although these classic four-term analogies recur throughout the web text, a majority (58.9%) of distinct responses generated only by models were never produced by any human given the same stem: e.g., given *beggar*:*poverty*::*child*:?, LLMs most frequently filled in *innocence*, which never appeared in humans' responses (see Supplement D for more examples of responses provided only by LLMs). Moreover, as shown in the next section, many responses provided only by LLMs were not only valid but rated as *better* than human responses.

LLMs produce better-rated analogies

For each analogy stem (A:B::C), we calculated the mean of the relational-similarity ratings, weighting each completion D by how frequently it was produced (equivalent to averaging ratings as if a completion produced 10 times was rated 10 times). We compared humans to each LLM using paired t-tests across all stems.

[2] The centroid is a single-point summary of each response distribution. As a complementary, distribution-level check, we also computed the Jensen-Shannon divergence between the human and LLM distributions per stem (Supplement A). The two analyses agree closely at the relation level (mean pairwise rank correlation across models $\rho = 0.95$): relations that are most convergent under the centroid measure (e.g., *Contrary* and *Taxonomic*) also show the lowest distributional divergence, while the most divergent relations (e.g., *Time:Associated Item*, *Object:State*, and *Representation*) are the same under both measures.

Figure 2

Relational-similarity ratings for LLM versus human analogy completions. Mean rating differences (LLM − Human) by relation type for six LLMs. Dotted lines show average effects across all relations. *** $p < .001$, ** $p < .01$, * $p < .05$, *n.s. = not significant.*

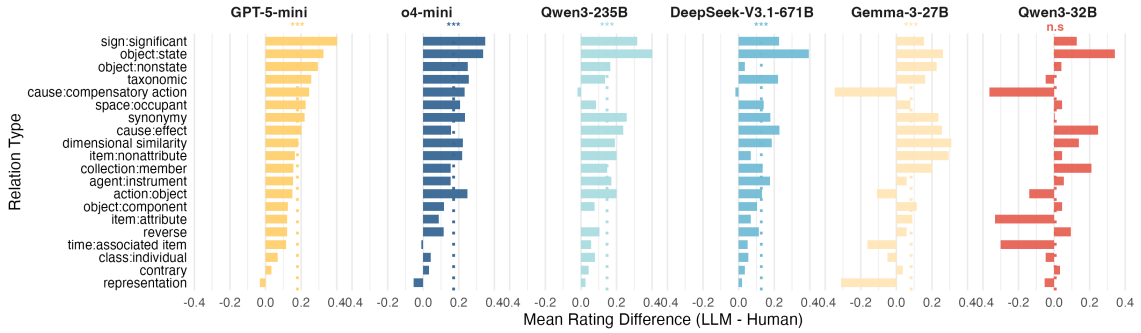


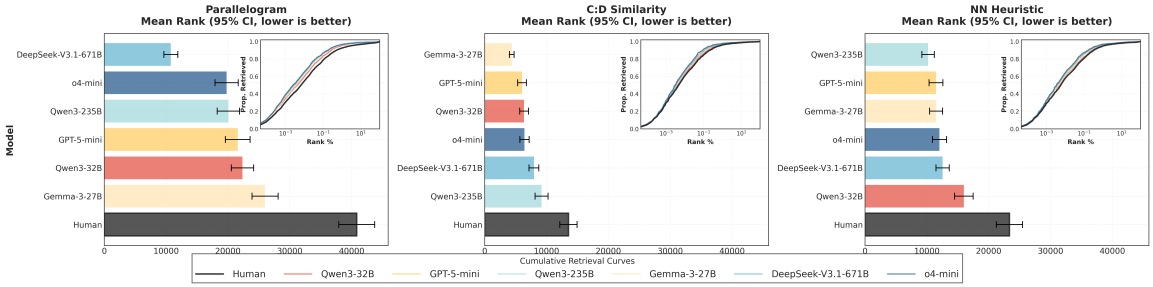
Figure 2 shows the mean rating differences (LLM − Human) for each model, both overall and broken down by the 20 semantic relations. Most LLM analogies were rated statistically significantly higher: `GPT-5-mini` showed the largest difference with humans ($b = .18$, 95% CI $[0.14, 0.22]$, $p < .001$), followed by `o4-mini` ($b = .17$, 95% CI $[0.13, 0.21]$, $p < .001$), `Qwen3-235B` ($b = .15$, 95% CI $[0.11, 0.19]$, $p < .001$), `DeepSeek-V3.1-671B` ($b = .13$, 95% CI $[0.08, 0.20]$, $p < .001$), and `Gemma-3-27B` ($b = .08$, 95% CI $[0.03, 0.13]$, $p < .001$). The only model whose completions were not rated as significantly better than humans' was `Qwen3-32B` ($b = .006$, 95% CI $[-.03, .05]$, $p = .7$).

Beyond receiving higher average ratings, LLMs frequently generated strong completions that no human participant produced. Of the 1,869 LLM-only completions for which we obtained ratings, 217 were rated higher than the *best* human completion given the same stem. For example, given *enigma*:*puzzlement*::*explosion*:?, models produced *shock* (rated 6.38), surpassing the best human response (*chaos*, 5.62); and for *diary*:*person*::*crown*:?, they produced *monarch*, the more abstract, less frequent, but more relation-preserving term, rated above the best human response (*king*, 4.00 vs. 3.50) (see Supplement D for more of such examples).

The magnitude of LLM advantage also varied considerably across relation types. Relations where humans and LLMs already converged showed near-zero rating gaps: Contrary analogies like *happy*:*sad*::*black*:*white* showed near-perfect human–LLM agreement. By contrast, relations with bigger divergence did not always correspond to a bigger rating gap. Some relations (e.g., Object:State) that diverged the most also showed one of the largest rating gaps. For example, given *novice*:*inexperience*::*child*:?, LLMs produced *immaturity* which was significantly better than humans' modal responses *youth* ($b = +.35$, $p < .001$). But some relations e.g., Time:Associated Item ($b = -.05$, $p = .75$) and Representation ($b = -.07$, $p = .41$), which diverged almost to the same extent, showed no significant rating gap. In these cases humans and LLMs drew on different distributions of responses that were nonetheless judged as comparably good, though for different underlying reasons. In Representation, humans and LLMs often read the relation differently, and neither reading was consistently rated as better. The relation between *diary*:*person*, for instance, can be read as a person authors a diary or a diary is about a person. Given *diary*:*person*::*biography*:?,

Figure 3

Performance of humans and six LLMs on three GloVe-based analogy metrics. Bar charts show mean rank with 95% confidence intervals (lower is better). Insets display cumulative proportion of responses retrieved as a function of rank percentage (log scale).



humans answered *author* (who writes it) while LLMs answered *subject* (what it is about). The direction reverses on other stems: e.g., given *backdrop*:*vista*::*documentary*:?, LLMs answered *reality* (what a documentary represents) while humans answered *story* or *film* (what documentary is a kind of), and here the LLM reading was rated higher (5.6 vs. 5.0 and 3.8). Because neither system's reading was rated as consistently better, the rating differences roughly offset across the relation's stems, leaving little net advantage for either side. In TIME:ASSOCIATED ITEM, on the other hand, the two groups drew from a large pool of equally plausible answers. For example, given *infancy*:*cradle*::*childhood*:?, humans answered with the child's sleeping furniture (*bed*, *bunk bed*) while LLMs answered with characteristic play spaces (*playground*, *swing*, *playpen*), which are different objects associated with the same period, all fitting the relation and rated similarly well.

Why are LLM Analogies Better?

To understand why LLM-generated analogies receive higher ratings from humans than human-generated analogies, we revisited the word embedding analysis of Peterson et al. (2020). Specifically, we ask: (1) How well do parallelogram vs. local similarity heuristics capture the responses produced by humans versus LLMs? (2) What predicts the higher human ratings of LLM-generated analogies? (3) Does increased parallelogram alignment reflect how LLMs internally generate analogies, or does it reflect what humans find to be good analogies?

How well do parallelogram vs. local-similarity rules predict human vs. LLM completions?

To quantify how well different geometric rules capture human vs. LLM response distributions, we adopted the cumulative proportion retrieved (CPR) analysis from Peterson et al. (2020), using the same pretrained GloVe embeddings (Pennington et al., 2014). We chose GloVe because it explicitly factorizes a word–context co-occurrence matrix, providing a statistically grounded embedding space where geometric properties have distributional interpretations (Ethayarajh, Duvenaud, & Hirst, 2019; Levy & Goldberg, 2014).

For each analogy stem A:B::C:?, each rule induced a ranking over the full vocabulary V ($|V|$ = 2,196,015). We tested three rules:

1. **Parallelogram model.** Predict the completion (D term) by applying to C the same offset that maps A to B, i.e., $\widehat{D} = B - A + C$. Then, rank all candidates $w \in V$ by cosine similarity $\cos(w, \widehat{D})$.

2. **C:D similarity.** Ignore the A:B relation and rank candidates only by similarity to C, i.e., by $\cos(w, C)$.

3. **Nearest-neighbor (NN) heuristic.** First compute whether A is closer to B or C. Set the target $T = C$ if $\cos(A, B) > \cos(A, C)$, and $T = B$ otherwise. Then, rank by cosine similarity to T, $\cos(w, T)$.

We then measured, at a rank percentile threshold τ, the proportion of observed responses whose predicted rank falls within the top-τ% of candidates. Higher CPR at *lower* τ indicates better prediction.

All three rules captured LLM responses better than human responses

On average, the responses of the LLMs were substantially better captured by all three of the rules ($ps < .001$; Figure 3). For example, at the 0.1% rank threshold, all three rules retrieved a larger proportion of LLM than human responses: parallelogram (86% vs. 78%), C:D similarity (89% vs. 84%), and NN (85% vs. 80%).

Parallelogram underperforms other rules overall, but the gap is smaller for LLMs

Consistent with Peterson et al. (2020), the parallelogram model was relatively good at predicting the top responses, while C:D similarity and NN better captured the full distribution. At the same time, the gap between parallelogram and local-similarity rules was notably smaller for LLMs: parallelogram mean ranks were on average 13,355 positions higher (worse) than C:D similarity for LLMs versus 27,264 for humans, and 7,799 higher than NN for LLMs versus 17,465 for humans. In both cases, the gap between parallelogram and local-similarity heuristic was more than twice as large for humans. As we show in the next section, this apparent advantage of local similarity in capturing *which* completions get produced does not translate into explaining which completions are judged to be *good*: it is parallelogram alignment, not local similarity, that predicts human ratings.

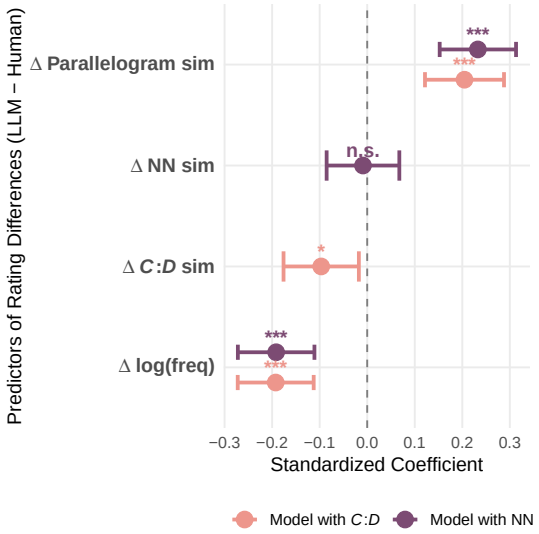
What predicts the higher human ratings of LLM-generated analogies?

To understand what predicts the advantage of LLM-generated analogies, we considered four non-mutually-exclusive predictors: LLMs may receive higher ratings because their responses are (1) more parallelogram-like, (2) more strongly associated with C (i.e., higher local similarity), (3) more aligned with the NN heuristic, or (4) less constrained by lexical accessibility, generating lower-frequency words.

For each completion D, we computed: **parallelogram alignment**, the cosine similarity between $B - A$ and $D - C$; **C:D local similarity**, $\cos(C, D)$; the **NN heuristic** score selecting either $\cos(C, D)$ or $\cos(B, D)$ depending on whether A is closer to B than to C; and **word frequency**, the log frequency of D via the `wordfreq` package (Speer, 2022). We took the weighted average of each metric for human and LLM responses for each stem and computed difference scores (LLM − Human): ΔParallelogram, ΔC:D similarity, ΔNN, and $\Delta \log(\text{freq})$. Because C:D similarity and NN

Figure 4

Standardized regression coefficients predicting rating differences. *** $p < .001$, * $p < .05$, *n.s.* = *not significant.*



are correlated ($r = .66$), these were entered into two multiple regression models (one with C:D and one with NN).

Parallelogram and word frequency explain why LLM analogies are better

As shown in Figure 4, ΔParallelogram strongly predicted the LLM rating advantage (Model with C:D: $\beta = 0.205$, $t = 4.84$, $p < .001$; Model with NN: $\beta = 0.233$, $t = 5.69$, $p < .001$): when LLM completions better satisfied the parallelogram constraint in GloVe, they were also rated as better analogies. Local similarity did not: ΔNN was not a significant predictor ($\beta = -0.009$, $p = .82$) and ΔC:D similarity was a significant *negative* predictor ($\beta = -0.097$, $p = .017$). To the extent that LLM completions were more similar to C compared to human completions, the LLM advantage actually shrank. Finally, Δ log(freq) was a negative predictor (Model with C:D: $\beta = -0.192$, $p < .001$; Model with NN: $\beta = -0.191$, $p < .001$), indicating that LLMs outperformed humans more when their completions were less frequent. We also verified that the parallelogram alignment advantage is not a downstream effect for LLMs providing more low-frequency responses, given that the advantage exists at different levels of word frequencies (see Supplement B for details).

Do LLMs' internal representations show the same parallelogram–rating relationship?

While LLM completions show stronger parallelogram alignment in GloVe space and this alignment predicts human ratings, two interpretations are possible: (1) LLMs excel because their internal representations are more geometrically structured, or (2) parallelogram structure in distributional spaces like GloVe captures what humans find compelling, and LLMs produce better analogies without relying on directly observable geometric representations. To distinguish these, we extracted word representations from the residual stream of each open-source model (`DeepSeek-V3.1`, `Qwen3-235B`, `Qwen3-32B`, `Gemma-3-27B`), sampling every fourth layer, and computed parallelogram

alignment using either isolated word embeddings (e.g., "king" presented alone) or contextual embeddings (e.g., "king" within *king is to queen as man is to woman*). We then tested how well alignment at each layer predicted human ratings.

Parallelogram alignment in GloVe predicts human judgments better than LLM internal representations

Parallelogram alignment in GloVe space strongly predicted human ratings ($\beta = 0.193$, $p < .001$), consistently outperforming all LLM internal representations. Aggregating across layers, LLM effects using isolated embeddings were substantially weaker: `Gemma-3-27B` ($\beta = 0.094$), `Qwen3-32B` ($\beta = 0.041$), `DeepSeek-V3.1` ($\beta = 0.040$), and `Qwen3-235B` ($\beta = 0.033$); contextual embeddings yielded even weaker predictions (see Supplement E for details). These results support the second interpretation: LLMs' advantage does not necessarily stem from more parallelogram-like internal representations but from generating analogies that better satisfy relational constraints captured in embedding spaces optimized for semantic representations.

The LLM advantage disappears for modal responses, but what makes a good analogy does not

While LLMs produce higher-rated analogies than humans on average, this advantage could reflect either that LLMs' responses are overall superior to humans' responses or that LLMs produce fewer low-quality responses. We examined the distributions of responses by humans and LLMs and found human response distributions were considerably more dispersed than LLM distributions: human modal responses accounted for 64% of total responses, compared to 85% for LLMs. In other words, humans produced a long tail of less common completions, even though unique completions were removed. LLMs, on the other hand, tended to concentrate their outputs on a few dominant answers. An entropy analysis of the response distributions further shows that human responses were reliably more dispersed than those of every LLM (Supplement F). If LLMs' advantage stems primarily from humans' long tail of weak responses rather than from LLMs producing better answers across the board, then restricting the comparison to only modal (most frequent) responses should eliminate the LLM advantage.

We tested this prediction by repeating the human–LLM comparison using only modal completions (frequency-weighted, with ties preserved). We found the LLM advantage over humans largely disappeared: no model showed a statistically significant overall advantage over humans ($ps > .05$), while `Qwen3-32B` and `Gemma-3-27B` were rated *lower* than human modal responses ($b = -0.10$, $p = .003$; $b = -0.08$, $p = .02$, respectively). All other models' modal responses were statistically indistinguishable from humans' ($ps > .35$). Paralleling this convergence in ratings, most LLMs' modal responses were also not significantly more parallelogram-aligned than humans' in GloVe space (DeepSeek-V3.1-671B: $p < .001$; Qwen3-235B: $p = .03$; all others $ps > .05$).

Crucially, the factors that predicted the LLM advantage in the full dataset remained significant when the analysis was restricted to modal responses. Although the two populations no longer differ on average in parallelogram alignment, alignment still varies from stem to stem, and that variation continues to track which completion is rated better. In our two regression models predicting the rating difference between LLM and human modal completions, ΔParallelogram was a strong positive predictor (Model with $C{:}D$: $\beta = 0.189$, $p < .001$; Model with NN: $\beta = 0.231$, $p < .001$), and $\Delta \log(\text{freq})$ remained a significant negative predictor (Model with $C{:}D$: $\beta = -0.173$, $p < .001$; Model with NN: $\beta = -0.161$, $p < .001$). $\Delta C{:}D$ similarity was still a significant negative

predictor ($\beta = -0.141$, $p < .001$), and ΔNN was a weaker negative predictor ($\beta = -0.086$, $p = .03$) (see Supplement F for additional visualizations of the modal-response analyses). These results indicate that the mechanisms driving rating differences between human and LLM completions, i.e., parallelogram alignment and lexical accessibility, operate consistently regardless of whether one examines the full distribution or only the dominant responses.

Optimizing an embedding space for parallelogram structure further improves prediction of both human and LLM responses

The GloVe embeddings we used were trained only to capture word–context co-occurrences, they were never designed to encode parallelogram structure. If parallelogram structure is what makes an analogy good, then sharpening the space's parallelogram geometry should make it a better account of the completions people and LLMs actually produce. We tested this by fine-tuning the GloVe embeddings to better satisfy the parallelogram constraint and asking whether the resulting space predicts human and LLM completions more precisely than raw GloVe.

Approach

We built a corpus of analogy quadruples (A:B::C:D) from pairs of words that share a semantic relation. We drew relations from Bejar, Chaffin, and Embretson (1991), who organize them into 81 fine-grained types (e.g., *taxonomic*, *contradictory*, *cause:effect*). For each type, we hand-collected about fifteen word pairs and prompted a language model (`Claude-Opus-4.8` (Anthropic, 2026)) to suggest roughly another forty pairs. Hand-picked examples include *flower*:*tulip* and *bird*:*robin* (taxonomic), *alive*:*dead* and *remember*:*forget* (contradictory), and *joke*:*laughter* and *tragedy*:*grief* (cause:effect); the model extended these with pairs such as *spice*:*cinnamon* (taxonomic), *win*:*lose* (contradictory), and *earthquake*:*destruction* (cause:effect).

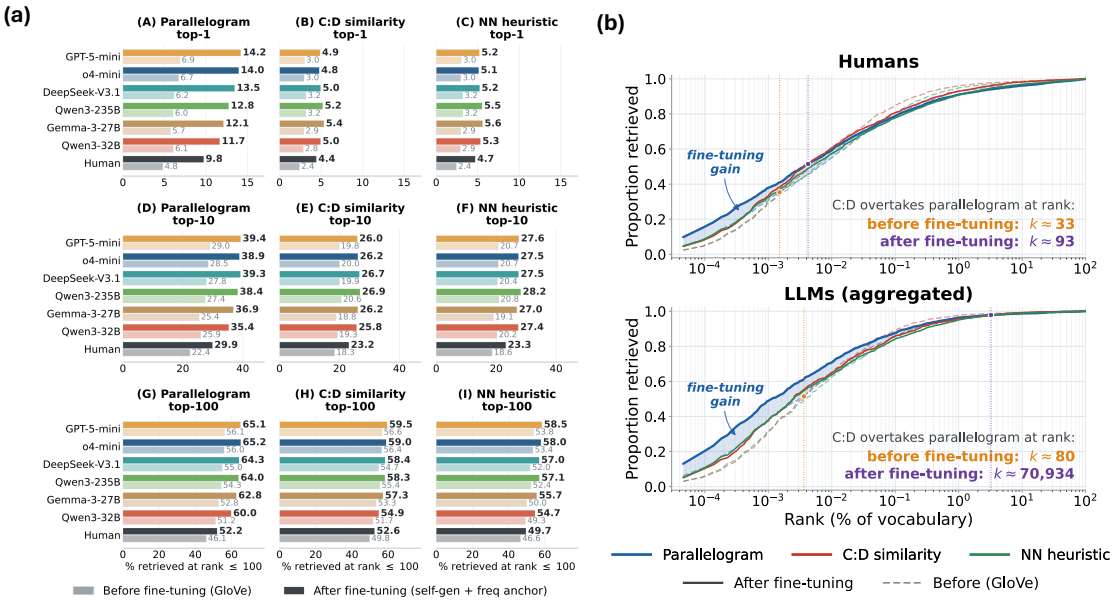
We then combine these pairs within the same type of relations into four-term analogies: from the taxonomic pairs *flower*:*tulip*, *bird*:*robin*, and *tree*:*oak*, for instance, we form *flower*:*tulip*::*bird*:*robin*, *flower*:*tulip*::*tree*:*oak*, and *bird*:*robin*::*tree*:*oak*. To prevent the model from memorizing the items we evaluate on, we removed any word pair that appeared in a quadruple (A:B::C:D) produced by humans or LLMs in our data before recombining. The resulting corpus contains 4,422 unique word pairs, recombined into 123,454 analogies. We then trained a small adapter that learns a single transformation applied to every GloVe word vector (Pennington et al., 2014), so that the transformed embeddings better satisfy the parallelogram relationship $B-A+C \approx D$. The details of adapter training are given in Supplement C.

Results

Strengthening the parallelogram structure of a GloVe space made it markedly better at describing the responses people and models give (see Figure 5): the top predictions now capture a substantially larger proportion of responses. For human completions, the fine-tuned parallelogram model returned the produced word as its top-1 guess 9.8% of the time (vs. 4.8% before) and within its top ten 29.9% of the time (vs. 22.4% before); for the LLMs, on average, top-1 rose from 6.3% to 13.0% and top-10 from 27.3% to 38.0% (see Figure 5a for effects broken down by LLMs). C:D similarity still outperforms the parallelogram on average rank (mean rank for capturing human responses is 17,800 for C:D similarity vs. 39,269 for the parallelogram; LLM mean rank for C:D similarity is 9,911 vs. 12,442 for parallelogram). But fine-tuning sharply extended the range

Figure 5

Sharpening the parallelogram improves retrieval of human and LLM completions. ***(a)*** *Percentage of held-out responses retrieved within the top 1/10/100 neighbors under the parallelogram, C:D similarity, and a nearest-neighbor heuristic, before (faded) and after (saturated) fine-tuning.* ***(b)*** *Cumulative proportion retrieved vs. rank for humans (top) and LLMs (bottom, aggregated across six models); solid = after, dashed = before, shading = fine-tuning's gain. Vertical dashed lines mark the rank at which C:D similarity overtakes the parallelogram; fine-tuning pushes this crossover much further down the list, extending the range over which the parallelogram remains the best predictor.*



over which the parallelogram is the best predictor: for humans, $C{:}D$ similarity did not overtake parallelogram until rank 93 (vs. rank 33 before), and for the LLMs the parallelogram led all the way out to rank 70,900 before $C{:}D$ similarity caught up (vs. rank 80 before fine-tuning; Figure 5b). The fine-tuned embeddings also predicted human ratings of relational similarity better than the raw embeddings: the correlation between the parallelogram prediction $B - A + C$ and the produced word D rose from Pearson $r = 0.16$ to 0.24 (Supplement C).[3]

General Discussion

Whether simple analogies can be captured via parallelograms in semantic vector spaces is a classic question in cognitive science. By directly comparing LLMs' completions of classic four-term word analogies with human completions, we were able to explore whether today's LLMs solve analogies like people do and, if not, what the differences reveal about the models that describe

[3] This offset-based score is one of two ways to measure the relationship between parallelogram alignment and similarity ratings. The other, used in Section "Why are LLM Analogies Better", followed Peterson et al., 2020 and measured whether the two relation vectors are parallel, i.e., the cosine between $B - A$ and $D - C$. Because the current adapter is trained to satisfy the offset relation $B - A + C \approx D$ and not the parallel direction score, fine-tuning improves the latter to a lesser degree, from $r = 0.19$ to 0.22.

each system. We found: (1) People judge LLM-generated analogies to be better than human-generated ones, on average. (2) LLM analogies are better predicted by the parallelogram model than human responses, although, much like people's, they are better captured by local similarity heuristics overall. (3) LLM-generated analogies are considered better when they more cleanly instantiate the parallelogram relation and involve less accessible (lower-frequency) words, despite little evidence that LLMs internally represent the parallelogram. (4) This advantage is driven not by LLMs producing superior responses across the board, but by humans generating a long tail of weaker completions. When only modal (most frequent) responses are compared, the LLM advantage disappears, though rating differences between LLM and human modal completions continue to be predicted by greater parallelogram alignment and lower word frequency. (5) Improving the parallelogram structure of a model improves the extent to which parallelogram can capture human and LLMs responses.

Overall, our results suggest that previous failure of the parallelogram model may reflect limitations of human analogy production rather than limitations of the model itself. With LLMs more consistently producing relation-preserving responses, we found the parallelogram model provides a robust account of what makes a word analogy good, even if it may not be a complete account of how either humans or LLMs generate them. In the remainder of the paper, we consider what these results imply for the theories of analogy, why local similarity continues to outperform the parallelogram model in predicting the full distribution, limitations, and directions for future research.

Parallelogram captures the perceived quality of analogies

Compared to humans, LLMs show closer alignment with the parallelogram model's predictions, and this alignment reliably predicts higher human quality ratings. In contrast, local similarity measures such as semantic similarity between C and D do not explain LLMs' advantage and, if anything, predict worse ratings. While Peterson et al. (2020) suggested that the parallelogram model may simply fail to capture human-like analogies, our findings suggest an alternative interpretation: people may tend to fail to *produce* strong relational analogies, but they nonetheless *prefer* analogies well captured by parallelograms.

One possible interpretation is that LLMs outperform humans because they internally represent analogical relations more geometrically. We found little evidence for this account. The LLMs' own internal representations predict human judgments substantially less well than the parallelogram models' predictions generated in an external embedding space optimized to represent lexical semantics. This suggests that the model that best characterizes analogies need not reflect the process by which those analogies are generated.

We also observe a robust effect of lexical accessibility on relation ratings. LLM completions tend to be lower-frequency than human completions, and people tended to assign lower-frequency completions higher ratings. Crucially, however, the effect of parallelogram alignment remains stable after controlling for accessibility, suggesting that LLM responses are rated better not only because they use less common or more "clever-sounding" words, but also because they better satisfy a relational constraint captured by the parallelogram model. Finetuning GloVe to be more parallelogram-aligned provides the strongest evidence that the underlying judgments track relational geometry per se, not GloVe specifically. Together, these findings suggest that the parallelogram model may characterize the structure of good analogies, without implying that either humans or LLMs explicitly encode analogies as parallelograms.

Local similarity heuristics and the parallelogram model

We also replicate and extend a key result from prior work: local similarity heuristics (especially C:D similarity) outperformed the parallelogram model as the best overall predictor for both humans and LLMs. Peterson et al. (2020) attributed this pattern in humans partially to cognitive ease: when rushed or confused, people may be drawn to local similarity judgments rather than engage in full relational reasoning. However, the fact that LLMs exhibit the same preference for local similarity heuristics challenges this account, as LLMs face neither the time pressure nor the cognitive load that constrains human responding. Alternatively, we think local similarity may reliably predict completions because most strong analogies exhibit both relational and semantic similarity between source and target domains. For instance, in *man*:*woman*::*king*:*queen*, *queen* is both relationally appropriate (gender) and semantically similar to *king* (royalty). Even when a clean relational offset is hard to infer due to ambiguity of word meaning or weak relational signal in A:B, retrieving a word close to C remains a robust fallback for the task.

At the same time, the advantage of local similarity under mean rank should also be interpreted with caution, as it can be disproportionately influenced by a relatively small number of responses that a rule fails to capture at all. Once an observed completion falls far down the ranked list, the difference between 5,000th and 50,000th has little practical consequence for predicting analogy completions, as both are misses. However, the larger value weighs ten times as heavily on the mean. In our case, just 1% of completions accounts for 50% of the parallelogram's mean rank for humans, and 80% for LLMs. Mean rank, therefore, reflects mainly how badly a rule fails on the completions it misses, rather than how well it predicts the completions humans and LLMs are actually likely to produce. Among the highest-ranked candidates, the parallelogram model retrieves a larger proportion of the completions humans and LLMs produced; its disadvantage in mean rank arises farther down the list, where C:D similarity places the completions both rules miss less catastrophically. This explains why the gap between parallelogram and C:D similarity's mean rank was more than twice as large for humans, as their long tail supplies most of these misses. Our fine-tuning results further support this interpretation, where strengthening the parallelogram structure primarily improved the model's ability to push observed responses toward the top of the ranked list, while having much less impact on how it ordered the large number of unlikely candidates farther down the list.

Limitations and Future Directions

Several limitations of the present work point toward directions for future research. First, the dispersion of LLM responses partly reflects our parameter settings during sampling. With `top_p` = 0.9, the model considers only the most probable candidate words whose probabilities together add up to 90%, which cuts off some low-probability tail that humans produce. Our modal-response analyses, which are insensitive to this truncation, suggest the core conclusions do not depend on it, but systematically varying the decoding parameters to match human and LLM dispersion would provide a cleaner test.

Second, the human analogy completions were collected by Peterson et al. (2020) several years before our study, and a small number of stems have since become outdated, e.g., *everest*:*mountain*::*charles*:*?*, elicited *prince* from human participants at the time of collection, a response that should be updated to *king* now.

A deeper limitation is that human similarity ratings, our primary measure of analogy quality, are themselves shaped by the same accessibility biases that affect production. People tend to rate

completions higher when they are stuck on the most salient reading of a relation, even when a less obvious reading is equally or more valid. The relation *diary*:*person*, for example, most naturally calls to mind authorship, i.e., a person writes a diary, so *biography*:*author* gets rated as a better completion than *biography*:*subject*, even though the latter captures the intended representational dimension of the relation just as legitimately. The same tension appears in cases like *elizabeth*:*queen*::*ireland*:*republic*, where *republic* is arguably the more precise answer but loses to more familiar alternatives (*country*) in human ratings; and given *sing*:*dirge*::*cook*:*?*, *feast* preserves the structural parallel more cleanly than *meal* (because a dirge is a genre of song, a feast is a genre of meal), yet raters preferred the simpler answer. In short, our quality measure inherits some of the same biases we are trying to study, which means we likely underestimate how often LLMs are "right" in a deeper sense.

This points to another interesting open question. Humans and LLMs do not just differ in how well they complete analogies—they seem to differ in how they interpret the relations underlying a given pair A:B in the beginning. When a relation supports multiple valid interpretations, humans and LLMs seem to diverge in reliable ways. Whether this reflects differences in how the two systems represent relations, or simply differences in what gets reinforced during training versus development, is not yet clear. Understanding what the qualitative difference reveals about the underlying representations on each side is a rich direction for future work.

Conclusion

We revisit and lend renewed support to a classic cognitive model of analogy. While the parallelogram model doesn't fully explain how humans or LLMs generate analogies, it reliably predicts what people judge to be good analogies. Moreover, when the geometric structure of an embedding space is explicitly optimized to better satisfy the parallelogram relation, the resulting model becomes better at recovering the completions that humans and LLMs produced and better at predicting human similarity judgments.

Our results also clarify the source of the apparent LLM advantage. LLMs do not outperform humans because they consistently produce superior responses. When only the most frequent responses are compared, the advantage largely disappears. Instead, the difference is driven by humans producing a broader distribution with a long tail of weaker completions, whereas LLMs concentrate their responses on a set of stronger, more relation-preserving completions. This pattern suggests that earlier evidence against the parallelogram model may have partially reflected the variability of human analogy production rather than the inadequacy of the parallelogram model.

More broadly, our results illustrate how LLMs can serve as useful tools for evaluating cognitive theories. Comparing human and LLM behavior allows theories to be tested against multiple intelligent systems rather than against human behavior alone. In the current case, doing so suggests that evidence previously interpreted as a failure of the parallelogram model may instead reflect limitations of human analogy production. Viewed this way, LLMs provide not just another benchmark for analogy performance, but a new source of evidence for distinguishing weaknesses of a cognitive theory from weaknesses of the human behavior traditionally used to evaluate it.

Data Availability

All data and analysis code are available on the Open Science Framework at https://osf.io/hqtw2/overview?view_only=cbb155437f03419cb47404574654d90d.

Acknowledgments

This research was supported by Toyota Motor North America, Inc. We also thank Dedre Gentner for thoughtful feedback.

References

Anthropic. (2026). *System card: Claude Opus 4.8.* Retrieved from https://www-cdn.anthropic.com/0b4915911bb0d19eca5b5ee635c80fef830a37ea.pdf (May 28, 2026)

Bejar, I. I., Chaffin, R., & Embretson, S. (1991). Theories of memory representation and analogical reasoning. In *Cognitive and psychometric analysis of analogical problem solving* (pp. 7–54). Springer.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems 33*.

DeepSeek-AI. (2024). *Deepseek-v3 technical report.* Retrieved from https://arxiv.org/abs/2412.19437

Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019). Towards understanding linear word analogies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3253–3262).

Gemma Team. (2025). *Gemma 3 technical report.* Retrieved from https://arxiv.org/abs/2503.19786

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170.

Green, A. E., Kraemer, D. J., Fugelsang, J. A., Gray, J. R., & Dunbar, K. N. (2010). Connecting long distance: semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cerebral Cortex*, *20*(1), 70–76.

Hendel, R., Geva, M., & Globerson, A. (2023). In-context learning creates task vectors. *arXiv preprint arXiv:2310.15916*.

Hofstadter, D. R. (2001). Analogy as the core of cognition. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 499–538). Cambridge, MA: The MIT Press.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. MIT Press.

Johnson, T., ter Veen, M., Choenni, R., van der Maas, H., Shutova, E., & Stevenson, C. E. (2025). Do large language models solve verbal analogies like children do? In *Proceedings of the 29th Conference on Computational Natural Language Learning* (pp. 627–639).

Jurgens, D., Turney, P. D., Mohammad, S. M., Holyoak, K. J., Rumshisky, A., Bigham, J., & Dyer, C. (2012). Semeval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)* (pp. 356–364).

Kmiecik, M. J., & Morrison, R. G. (2013). Semantic distance modulates the n400 event-related potential in verbal analogical reasoning. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society.*

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Advances in Neural Information Processing Systems 27*.

Lewis, M., & Mitchell, M. (2024). Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *arXiv preprint arXiv:2402.08955*.

Linzen, T. (2016). Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP* (pp. 13–18).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (Vol. 26).

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407.

OpenAI. (2025). *OpenAI o3 and o4-mini system card.* Retrieved from https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf (April 16, 2025)

OpenAI. (2026). *OpenAI GPT-5 system card.* Retrieved from https://arxiv.org/abs/2601.03267 (System card originally released August 13, 2025)

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).

Peterson, J. C., Chen, D., & Griffiths, T. L. (2020). Parallelograms revisited: Exploring the limitations of vector space models for simple analogies. *Cognition*, *205*, 104440.

Piantadosi, S. T., Muller, D. C., Rule, J. S., Kaushik, K., Gorenstein, M., Leib, E. R., & Sanford, E. (2024). Why concepts are (probably) vectors. *Trends in Cognitive Sciences*, *28*(9), 844–856.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Reed, S. E., Zhang, Y., Zhang, Y., & Lee, H. (2015). Deep visual analogy-making. *Advances in Neural Information Processing Systems 28*.

Rogers, A., Drozd, A., & Li, B. (2017). Too many problems with analogy-based evaluation of word vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 135–148).

Rumelhart, D. E., & Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology*, *5*(1), 1–28.

Sadler, D. D., & Shoben, E. J. (1993). Context effects on semantic domains as seen in analogy solution. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(1), 128.

Speer, R. (2022, September). *rspeer/wordfreq: v3.0.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.7199437 (Version 3.0.2) doi: 10.5281/zenodo.7199437

Todd, E., Li, M. L., Sharma, A. S., Mueller, A., Wallace, B. C., & Bau, D. (2023). Function vectors in large language models. *arXiv preprint arXiv:2310.15213*.

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

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., ... Qiu, Z. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yasunaga, M., Chen, X., Li, Y., Pasupat, P., Leskovec, J., Liang, P., ... Zhou, D. (2023). Large language models as analogical reasoners. *arXiv preprint arXiv:2310.01714*.