frontiers in Language Sciences

TYPE Brief Research Report
PUBLISHED 04 June 2026
DOI 10.3389/flang.2026.1756514

OPEN ACCESS

EDITED BY
Seana Coulson,
University of California, San Diego, La Jolla, CA, United States

REVIEWED BY
Ross Deans Kristensen-McLachlan,
Aarhus University, Denmark
Chiara Battaglini,
University Institute of Higher Studies in Pavia, Italy

*CORRESPONDENCE
Laura Pissani
✉ laura.pissani@uni-saarland.de

RECEIVED 28 November 2025
REVISED 02 April 2026
ACCEPTED 27 April 2026
PUBLISHED 04 June 2026

CITATION
Pissani L, Jobanputra M and Demberg V (2026) LLMs replicate metaphor norms based on word co-occurrence but struggle with topic-vehicle mappings. *Front. Lang. Sci.* 5:1756514. doi: 10.3389/flang.2026.1756514

LLMs replicate metaphor norms based on word co-occurrence but struggle with topic-vehicle mappings

Laura Pissani[1]*, Mayank Jobanputra[2] and Vera Demberg[1,2]

[1]Department of Language Science and Technology, Saarland University, Saarbrücken, Germany, [2]Department of Computer Science, Saarland University, Saarbrücken, Germany

Normed linguistic stimuli are fundamental in psycholinguistics because they capture lexical and semantic properties that influence comprehension. However, generating these norms at scale is challenging, often leading researchers to rely on *ad hoc* norms collected from small samples, which can introduce inconsistencies and limit cross-study comparisons. In the present study, we investigated how large language models (LLMs) can support psycholinguistic research by prompting eight current LLMs to norm 300 English two-word metaphor combinations, such as *sharp mind*. We selected the dimensions of familiarity, aptness, concreteness, metaphoricity, and constituency, as these tap distinct cognitive processes and may provide insight into which aspects LLMs capture accurately and which they do not. We varied stimulus presentation (in context vs. in isolation) and response format (categorical vs. numerical) to examine which manipulation yields norms most closely aligned with human ratings. We then assessed the reliability and validity of model responses and used them to replicate existing analyses of metaphor comprehension. Overall, LLM-generated norms aligned best with familiarity and metaphoricity, which rely on word co-occurrence. In contrast, aptness, concreteness, and constituency—which require reasoning about the relationship between the topic (e.g., *mind*) and the vehicle (e.g., *sharp*)—proved more challenging for LLMs.

KEYWORDS
aptness, concreteness, constituency, familiarity, large language models, metaphoricity, metaphors, psycholinguistic norms

1 Introduction

Research in psycholinguistics relies heavily on normed linguistic stimuli. Such norms often involve lexical properties (e.g., frequency, length, constituency) as well as human judgments about semantic properties (e.g., familiarity, aptness, concreteness) known to affect comprehension. These datasets are therefore essential for norming experimental materials and for identifying and controlling factors that may influence participants' responses.

However, building such normed datasets at scale is time-consuming and costly, often leading researchers to create their own *ad hoc* datasets with a smaller number of participants. This approach can result in inconsistencies, as different judges may apply different standards, and comparisons across studies become difficult when each researcher relies on an independently normed dataset.

To address these limitations, we investigated how large language models (LLMs) can be used to norm stimuli for psycholinguistic research. We focused on a recently released metaphor dataset (Pissani and de Almeida, 2026) comprising 300 two-word English metaphor combinations rated on familiarity, aptness, concreteness, metaphoricity, and constituency, and further validated our results on a second dataset (Cardillo et al., 2010).

Metaphors provide an ideal test case for evaluating LLMs' capabilities, as understanding metaphors requires not only retrieval of linguistic information but also the integration of pragmatic and contextual knowledge (Levorato and Cacciari, 2002). Metaphors reflect mappings between two seemingly unrelated domains, in which a topic is described in terms of a source domain, also called vehicle. For instance, the source domain of colors can convey warning signals (e.g., *red flag, green flag*); temperature can describe personality traits (e.g., *cold heart, hot temper*); animals can characterize human behavior (e.g., *night owl, busy bee*); and light can convey mental states (e.g., *bright idea, dark thoughts*).

Metaphor has been a topic of scholarly interest since Aristotle's *Rhetoric*, and research on it continues to grow in the fields of linguistics, psychology, and philosophy (e.g., Grice, 1975; Clark and Lucy, 1975; Searle, 1979; Lakoff and Johnson, 1980; see Genovesi, 2020, 2023 for a review). More recently, two-word metaphors have been employed to investigate the time course and mechanisms of metaphor processing (Al-Azary et al., 2021; Forgács et al., 2015; Park et al., 2021; Pissani and de Almeida, 2022, 2023), properties of the expressions that influence comprehension (Gagné, 2002; Pissani et al., 2026a), individual differences in comprehension (Pissani et al., 2026b), the neural correlates of metaphor comprehension (Arzouan et al., 2007; Goldstein et al., 2012), and hemispheric contributions to processing (Briner et al., 2018; Kavé et al., 2014).

2 Related work

The recent integration of LLMs into psycholinguistic research has introduced a paradigm shift in how linguistic norms are collected and validated. Addressing the scalability issues inherent in human data collection, Trott (2024) demonstrated that models such as GPT-4 can generate norms for semantic dimensions like concreteness, valence, and arousal that correlate strongly ($r > .80$) with human "gold standards," effectively acting as synthetic subjects for lexical norming. This finding is supported by Dillion et al. (2023), who proposed the "number needed to beat" framework, suggesting that while LLMs can serve as cost-effective proxies for human participants, their utility varies significantly by task complexity and requires rigorous validation against human baselines. Further extending this methodology to multi-word expressions, recent work has shown that LLMs can capture the nuances of contextualized meaning (Martínez et al., 2025), a critical capability for metaphor research where the intended meaning is highly dependent on the surrounding phrase.

In the specific domain of figurative language, LLMs have demonstrated emergent capabilities that go beyond simple lexical retrieval. Ichien et al. (2024) found that GPT-4 could interpret novel literary metaphors with a level of sophistication that human judges rated as superior to that of university students, exhibiting sensitivity to Gricean pragmatic principles such as the "restorative" interpretation of reversed metaphors. Complementing this interpretive capacity, DiStefano et al. (2025) showed that fine-tuned LLMs (e.g., RoBERTa, GPT-2) could predict human creativity ratings for metaphors with high reliability ($r \approx 0.72$), significantly outperforming traditional semantic distance metrics. Similarly, Hu et al. (2023) found that larger LLMs (e.g., Flan-T5 (XL) and text-davinci-002) performed near human level when interpreting expressions requiring the integration of pragmatic knowledge, with metaphor interpretation showing the closest alignment with human judgments, suggesting that, like humans, LLMs are sensitive to contextual information that favors figurative over literal interpretation.

However, the use of LLMs as autonomous evaluators remains contentious. Bavaresco et al. (2025), in their JUDGE-BENCH study across 20 NLP tasks, reported substantial variance in model reliability and a tendency for LLMs to align more closely with non-expert crowd workers than with domain experts. This underscores the necessity of the current study's approach: leveraging LLMs not as replacement participants, but as sources of normative data requiring careful psycholinguistic validation.

3 The current study

This study examines the integration of large language models (LLMs) into psycholinguistic research by prompting eight current LLMs to rate 300 two-word metaphors on familiarity, aptness, concreteness, metaphoricity, and constituency. We extend previous work by manipulating (1) the presentation of stimuli (in context vs. in isolation) to examine whether LLMs benefit from context when evaluating metaphors; (2) the type of output generated (categorical vs. numerical) to examine whether LLMs perform better when assigning categories rather than ratings; and (3) the inclusion of metaphor-specific dimensions to assess how models generate ratings for dimensions based on frequency of use (e.g., familiarity) vs. dimensions that rely more on relational mapping (e.g., aptness), sensory or physical experience (e.g., concreteness), contextual information (e.g., metaphoricity), or structural analysis (e.g., constituency).

The study has three main objectives. First, to determine whether and when LLM-generated norms are reliable and align with human norms. Second, to assess whether they replicate previously observed effects in studies of metaphor comprehension and can therefore be used in psycholinguistic research. Third, to provide best-practice recommendations for generating and using LLM-generated norms. To this end, we measured the internal consistency of model responses across random seeds and computed correlations with human ratings for all dimensions. We then conducted three substitution analyses to examine whether LLM-generated ratings replicate patterns observed in the metaphor dataset, reproduce effects reported in online studies using a subset of items from the dataset, and generalize to other types of metaphor stimuli.

4 Methods

4.1 The metaphor dataset

We used the metaphor dataset from a recent norming study (Pissani and de Almeida, 2026), which provides semantic and lexical properties for 300 two-word metaphors, judged by 164 native English speakers. The metaphors are of the form *YX* and include two grammatical combinations: adjective + noun (e.g., *early bird*) and noun + noun (e.g., *night owl*). Each expression involves a "vehicle," the word carrying most of the metaphorical content (e.g., *sharp*), and a "topic," the constituent that is predicated on (e.g., *mind*). Most expressions were drawn from everyday conversations, streaming services, and social media to capture widely used metaphors (e.g., *bright side*, *cold feet*). Others were more technical (e.g., *pie chart*, *cheese writing*), and some were adapted from poetry (e.g., *nodding leaves*, *unkempt afternoon*). The dataset was designed to cover a broad range of metaphors that vary in these properties to facilitate the investigation of how features of the expressions or individual differences among participants inform the interpretation and semantic composition of two-word metaphors.

This metaphor dataset is particularly valuable because it includes five dimensions that differ in the cognitive processes required to assign their ratings. Familiarity involves recalling prior exposure and frequency of use, sometimes approximated using frequency counts or Google searches (e.g., Roncero and de Almeida, 2014). Aptness requires identifying the topic and vehicle of a metaphor, and whether the vehicle captures important features that can be predicated on the topic (Roncero and de Almeida, 2015). Concreteness requires assessing the degree to which the referent of the vehicle is perceptible through human sensory or physical experience. Metaphoricity involves recognizing that some expressions can be interpreted literally or figuratively, depending on context. Finally, constituency is unique to two-word metaphors and requires identifying whether both words are metaphorical or whether one word is metaphorical and the other remains literal. Examining these dimensions may provide insight into which cognitive processes LLMs perform well and which they perform less successfully.

In the following sections, we describe the five dimensions. All were rated on a seven-point scale, with seven indicating higher values, except for constituency, which used a transformed scale from —3 to 3 to make the ratings more intuitive for participants.

4.1.1 Familiarity

Familiarity ratings reflect the extent to which participants had previously encountered the expressions, either by hearing or reading them. More familiar metaphors are often easier to retrieve from memory, and processed faster than less familiar expressions (Bowdle and Gentner, 2005; Giora, 1997; Blasko and Connine, 1993). Familiarity was collected both in context (e.g., *She was left with a BROKEN HEART after the split with her partner*) and in isolation (e.g., *broken heart*). In the metaphor dataset, expressions presented in context received higher familiarity ratings.

4.1.2 Aptness

Aptness ratings reflect the quality of the metaphorical mapping, specifically whether the vehicle of the metaphor (e.g., *butter* in *butter fingers*) effectively conveys salient features of the topic (e.g., *fingers* that are slippery). Aptness has been shown to modulate the retrieval of metaphorical content (Blasko and Connine, 1993), to influence whether a mapping is preferred in metaphor or simile form (Chiappe et al., 2003; Chiappe and Kennedy, 1999), and to facilitate metaphor interpretation in clinical populations (Roncero and de Almeida, 2014). Aptness ratings were also collected both in context and in isolation. However, unlike familiarity, these ratings did not show a significant effect of context in the metaphor dataset.

4.1.3 Concreteness

Concreteness ratings reflect the extent to which the meaning conveyed by a metaphorical vehicle can be perceived through the senses or actions. For example, *cloud nine* conveys a state of extreme happiness, which is difficult to perceive and received a low rating (2.00). Concreteness has been shown to influence metaphor processing in neuroimaging studies (Lai et al., 2009; Canal et al., 2022; Forgács et al., 2015). For instance, Lai et al. (2009) found that action verbs used metaphorically (e.g., *bending a rule*) elicited larger N400 amplitudes, associated with increased semantic processing difficulty, than the same verbs used literally (e.g., *bending a rod*).

4.1.4 Metaphoricity

Metaphoricity ratings reflect the extent to which an expression is perceived as metaphorical rather than literal. For example, *fan brush* received a low rating (1.48), as it can be interpreted either literally (a brush used to fan) or metaphorically (a brush shaped like a fan but not used for fanning). In contrast, *couch potato* received a high rating (6.32), as it is consistently interpreted metaphorically, regardless of context. In the metaphor dataset, metaphoricity was influenced by other dimensions, with higher familiarity increasing metaphoricity ratings and higher aptness and concreteness decreasing them.

4.1.5 Constituency

Constituency ratings reflect how the metaphorical content is distributed across the two words of an expression. This dimension was measured numerically and categorically. On the numerical scale (–3 to 3), ratings closer to –3 indicate that the metaphorical content is primarily in the first word (e.g., *brilliant student*), ratings near 0 indicate that it is equally distributed across both words (e.g., *smoking gun*), and ratings closer to 3 indicate that it is primarily in the second word (e.g., *early bird*). The categorical version of this dimension (first, both, or second) was annotated by the researchers and labeled *vehicle position*. In the metaphor dataset, expressions with metaphorical content in either the first or second word were

perceived as more concrete, whereas expressions with content distributed across both words were perceived as less concrete.

4.2 Models

We evaluated eight contemporary LLMs that differ in size, pretraining data, and post-training methods (see Table 1). Medium-sized models include Apriel-1.5-15B-Thinker (Radhakrishna et al., 2025), Mistral-Small-3.2-24B-Instruct-2506 (Mistral AI Team, 2025), and Qwen3-32B (Qwen Team, 2025), all ranging between 15–32B parameters. Larger models include Apertus-70B-Instruct-2509 (Hernández-Cano et al., 2025), Llama-3.3-70B-Instruct (Grattafiori et al., 2024), Qwen3-Next-80B-A3B-Thinking (Qwen Team, 2025), gpt-oss-120B (OpenAI Team, 2025), and DeepSeek-V3.2-Exp (DeepSeek-AI et al., 2024), all ranging between 70–685B parameters. All models are either instruction-tuned or "thinking" variants designed to follow natural language prompts and generate step-by-step justifications. Thinking models were chosen because the metaphor dimensions evaluated here often involve mapping relations from a source domain to a target domain—a process that relies on analogical reasoning (Gentner, 1983; Holyoak, 2013). Tasks of this type may particularly benefit from Chain-of-Thought prompting, as it encourages explicit multi-step reasoning that can facilitate cross-domain mappings. We accommodated this by allowing models to output reasoning traces within specific blocks before providing the final rating. For elaborate explanation on the "thinking" paradigm, refer to DeepSeek-AI et al. (2024). By spanning a wide parameter range and including multiple model families, our selection allows us to examine how model size and architecture affect alignment with human metaphor norms.

We hosted all the models locally (except for DeepSeek-V3.2-Exp) using vLLM and queried them via API without any additional fine-tuning. Each item was submitted independently to obtain the final ratings without appending past responses, eliminating the possibility of anchoring effects from prior items in the dataset. For each model, we utilized the default hyperparameters, including temperature and top-p, as recommended by the providers on their Hugging Face pages (see Table 6 in Appendix for exact values).[1] All LLMs received the same few-shot prompts, presenting items exactly as they were shown to humans with the target expression in uppercase to ensure fidelity to the human task (e.g., *Learning a new language requires BABY STEPS at first.*). The prompts differed only in the experimental manipulations described above (in context vs. in isolation and numerical vs. categorical output), requesting numbers for ratings and words for categorical judgments. To ensure valid data extraction, we instructed models to wrap their final answer in XML-style tags (e.g., `<constituency>4</constituency>` or `<aptness>high</aptness>`), which we parsed programmatically. To assess robustness, we sampled three independent outputs for each item by varying the random seed, and averaged these responses to obtain a single rating per item and model. This design enables a direct comparison of models both to one another and to human judgments, while also allowing us to quantify internal consistency across generations.

5 Results

We first evaluated the reliability and validity of LLM outputs in replicating human judgments by assessing their internal consistency and alignment with human ratings across several dimensions: familiarity in context (FAMc) and in isolation (FAMi); aptness in context (APTc) and in isolation (APTi); concreteness (CON); metaphoricity (MET); constituency (CONS); and vehicle position (vehicle). These analyses were conducted for both the full set of 300 metaphors and a subset of 24 metaphors, which was used in an online study of metaphor comprehension, to evaluate whether substitution analyses based on the subset would produce statistically reliable and valid results.

Next, we conducted substitution analyses to assess whether LLM-generated norms replicated the findings reported in the metaphor norming study (Pissani and de Almeida, 2026) and the online comprehension study (Pissani et al., 2026a).

Finally, to evaluate the external validity of the LLM-generated norms, we applied the prompts used to generate familiarity ratings to a separate set of metaphors from an offline study (Stamenković et al., 2019), which normed their materials using familiarity ratings from Cardillo et al. (2010).

5.1 Reliability

For numerical data, we assessed the reliability of model ratings using the intraclass correlation coefficient (ICC), computed with the `psych` package in R (Revelle, 2025) for both the full set and the subset.

For human data, the ICC reflects interrater reliability (i.e., variability among raters), whereas for model data it reflects intrarater reliability (i.e., variability in outputs generated by the same model across three random seeds). We report ICCs based on mean ratings, consistency, and two-way mixed-effects models (ICC 3, *k*) in Table 2. In the norming study, ICCs were computed from mean ratings aggregated across approximately 25 participants per dimension and 10 participants for constituency, and are reported here as a human baseline. Higher ICCs indicate greater internal consistency and, consequently, greater stability of the ratings. Following the guidelines of Koo and Li (2016), values above 0.50 indicate moderate reliability, above 0.75 good reliability, and above 0.90 excellent reliability.

For the full set, all models achieved good to excellent reliability in familiarity, aptness, concreteness, metaphoricity, and constituency, except for Apertus-70B-Instruct-2509, which showed

1 While varying temperatures can introduce variability, recent research (Pavlovic and Poesio, 2024) demonstrates that temperatures above zero are actually advantageous for subjective language tasks because they better approximate diverse human opinion distributions. As shown in Table 6 in Appendix, we also use a temperature > 0, as suggested by Pavlovic and Poesio (2024). Furthermore, our robust assessment of intraclass correlation coefficients across three random seeds confirms that these settings maintain strong internal consistency without compromising the overall stability of the outputs.

TABLE 1 Models tested in our studies.

Model	# parameters	Pretraining data	# Post-training datapoints
Apriel-1.5-15B-Thinker	15B	Pixtral-12B	> 2M text SFT samples
Mistral-Small-3.2-24B-Instruct-2506	24B	∼10T	NA
Qwen3-32B	32B	36T	∼300*K* CoT examples (Stage 1)
Apertus-70B-Instruct-2509	70B	15T (1811 langs)	≈3.8M SFT examples
Llama-3.3-70B-Instruct	70B	15T	>25M synthetic examples
Qwen3-Next-80B-A3B-Thinking	80B	15T (subset of 36T)	NA
gpt-oss-120b	120B	∼15T	NA
DeepSeek-V3.2-Exp (thinking)	685B	14.8T + 943B	NA

moderate reliability (ICC = 0.51) in the constituency dimension. In particular, gpt-oss-120b and DeepSeek-V3.2-Exp demonstrated excellent reliability across all dimensions, exceeding that of human raters. For the subset, all models exhibited good to excellent reliability on familiarity and metaphoricity, while substantial variability was observed in the other dimensions. Notably, gpt-oss-120b, DeepSeek-V3.2-Exp, and Mistral-Small-3.2-24B-Instruct-2506 consistently showed good to excellent reliability across all dimensions. For the dimension of aptness, ICC was reported as zero for Apriel-1.5-15B-Thinker, Llama-3.3-70B-Instruct, and Qwen3-Next-80B-A3B-Thinking, as these models assigned the same value to all items in the subset, resulting in insufficient variance for reliability estimation.

For categorical data, reliability was measured using Krippendorff's alpha. Table 2 reports alpha values comparing variability in responses produced by the same model across three seeds for vehicle position (i.e., whether the metaphorical content is in the first, second, or both words of the expression). Following the interpretation guidelines of Marzi et al. (2024), values near 0 indicate no agreement, values below 0.67 indicate poor agreement, values between 0.67 and 0.79 indicate moderate agreement, values above 0.80 indicate good agreement, and values approaching 1 indicate nearly perfect agreement. Overall, reliability values were variable and inconsistent, with no systematic relationship to model size. Roughly half of the models showed poor reliability, while the remainder showed moderate reliability. Only Apertus-70B-Instruct-2509 and DeepSeek-V3.2-Exp (Thinking) reached the threshold for good internal consistency for both the full set and the subset, while Mistral-Small-3.2-24B-Instruct-2506 and Apertus 70B Instruct 2509 achieved perfect reliability for the subset.

5.2 Validity

For numerical data, we assessed validity by computing Spearman rank correlations between model and human ratings across all dimensions (Table 3). Human ratings were averaged per item as reported in the norming study Pissani and de Almeida (2026), whereas model ratings were averaged per item across the three seeds, separately for the full set (300 items) and the subset (24 items). For the full set, correlations were moderate to strong for familiarity and metaphoricity, moderate for concreteness, and weak to moderate for aptness, with substantial variability for constituency. For the subset, correlations remained moderate to strong for metaphoricity, but for familiarity, only half of the models reached significance. None of the correlations for aptness were significant. For concreteness, only Apriel-1.5-15B-Thinker and gpt-oss-120b showed a significant correlation, while correlations for constituency remained variable. Overall, there was no consistent relationship between model size and alignment with human ratings.

For categorical responses, validity was assessed using Cohen's kappa, computed with the `irr` package in R (Gamer et al., 2012). We compared model responses from the first seed with responses annotated by human raters for vehicle position reported in the norming study. Following the interpretation guidelines of McHugh (2012), values above 0.40 indicate weak agreement, values above 0.60 indicate moderate agreement, values above 0.80 indicate strong agreement, and values above 0.90 indicate almost perfect agreement. Only Apriel-1.5-15B-Thinker reached moderate agreement for both the full set and the subset. Qwen3-32B and gpt-oss-120b showed weak agreement for the full set. All other models showed little to no agreement with human annotations in either set.

5.3 Substitution analysis

We first performed a substitution analysis on the full set of metaphors to determine whether the relationships among dimensions followed the same patterns observed in the metaphor dataset.[2] We then performed a second substitution analysis on a subset of expressions used in an online study of metaphor processing to examine whether LLM-generated ratings replicated the observed effects. Finally, we performed a third substitution analysis on a separate set of metaphors from an offline study of metaphor comprehension to evaluate the external validity of our results.

5.3.1 Findings from the norming study

We examined whether the full set of LLM-generated ratings replicated the relationships among dimensions

2 The categorical variable *vehicle position* showed poor reliability (see Table 2) and was therefore excluded from substitution analyses.

TABLE 2 Reliability values across three seeds for each model.

	Full set								Subset							
Model	FAMc	FAMi	APTc	APTi	CON	MET	CONS	Vehicle	FAMc	FAMi	APTc	APTi	CON	MET	CONS	Vehicle
Metaphor dataset	0.88	0.88	0.78	0.70	0.87	0.94	0.93	—	0.74	0.75	0.65	0.50	0.80	0.93	0.95	—
Apriel-1.5-15B-Thinker	0.94	0.94	0.84	0.89	0.87	0.84	0.96	0.73	0.91	0.95	0.00	0.00	0.71	0.79	0.99	0.55
Mistral-Small-3.2-24B-Instruct-2506	0.99	0.99	0.91	0.99	0.97	0.96	0.87	0.00	0.98	0.98	0.85	0.99	0.97	0.92	0.75	1.00
Qwen3-32B	0.91	0.92	0.78	0.83	0.82	0.93	0.84	0.66	0.91	0.91	0.53	0.38	0.59	0.82	0.79	0.36
Apertus-70B-Instruct-2509	0.93	0.94	0.92	0.91	0.82	0.95	0.51	0.76	0.87	0.95	0.89	0.96	0.84	0.94	0.68	1.00
Llama-3.3-70B-Instruct	0.96	0.95	0.96	0.88	0.92	0.94	0.79	0.73	0.92	0.92	0.90	0.00	0.94	0.89	0.70	0.49
Qwen3-Next-80B-A3B-Thinking	0.97	0.97	0.89	0.92	0.88	0.92	0.98	0.61	0.97	0.96	0.00	0.00	0.08	0.89	0.93	0.65
gpt-oss-120b	0.97	0.96	0.92	0.94	0.91	0.90	0.96	0.82	0.97	0.97	0.85	0.86	0.84	0.81	0.95	0.64
DeepSeek-V3.2-Exp (Thinking)	0.99	0.99	0.92	0.92	0.94	0.97	0.95	0.69	0.85	0.95	0.89	0.78	0.90	0.95	0.94	0.68

FAMc: familiarity in context; FAMi: familiarity in isolation; APTc: aptness in context; APTi: aptness in isolation; CON: concreteness; MET: metaphoricity; CONS: constituency; vehicle: vehicle position.
Intraclass correlation coefficient (FAMc–CONS): values below 0.50 are shown in gray. Krippendorff's alpha (Vehicle): values below 0.67 are shown in gray.

TABLE 3 Alignment between model and human responses.

	Full set								Subset							
Model	FAMc	FAMi	APTc	APTi	CON	MET	CONS	Vehicle	FAMc	FAMi	APTc	APTi	CON	MET	CONS	Vehicle
Apriel-1.5-15B-Thinker	0.70*	0.57*	0.36*	0.29*	0.58*	0.61*	0.80*	0.71	0.58·	0.36	0.01	-0.32	0.51·	0.30	0.80*	0.67
Mistral-Small-3.2-24B-Instruct-2506	0.65*	0.55*	0.24*	0.33*	0.48*	0.56*	0.07	0.00	0.32	0.18	−0.09	−0.16	0.06	0.60*	0.00	0.00
Qwen3-32B	0.77*	0.67*	0.41*	0.39*	0.59*	0.67*	0.69*	0.51	0.55*	0.38	0.13	0.24	0.24	0.59*	0.47·	0.10
Apertus-70B-Instruct-2509	0.58*	0.47*	0.29*	0.28*	0.44*	0.45*	0.05	−0.01	0.35	0.15	0.07	0.00	0.29	0.55*	0.06	−0.08
Llama-3.3-70B-Instruct	0.71*	0.61*	0.36*	0.47*	0.49*	0.55*	0.45*	0.38	0.10	0.12	0.01	0.08	0.17	0.59*	0.47·	0.23
Qwen3-Next-80B-A3B-Thinking	0.76*	0.66*	0.46*	0.47*	0.51*	0.61*	0.81*	0.07	0.65*	0.24	0.29	0.04	0.08	0.72*	0.78*	−0.01
gpt-oss-120b	0.74*	0.55*	0.44*	0.46*	0.60*	0.60*	0.79*	0.52	0.63*	0.36	0.01	−0.03	0.59*	0.56*	0.63*	0.31
DeepSeek-V3.2-Exp (Thinking)	0.70*	0.64*	0.41*	0.46*	0.45*	0.64*	0.78*	0.29	0.36	0.44·	0.03	−0.09	0.14	0.45·	0.65*	0.14

Spearman rank correlations (FAMc–CONS): asterisks (*) indicate $p < 0.01$, dots (·) indicate $p < 0.05$, and nonsignificant values are shown in gray. Cohen's kappa (Vehicle): values below 0.40 are shown in gray.

10.3389/flang.2026.1756514

observed in the metaphor dataset. First, in the norming study (Pissani and de Almeida, 2026), familiarity ratings[3] and COCA frequency scores were moderately positively correlated ($r = 0.43$; Finding 1). Second, familiarity and aptness ratings were highly positively correlated ($r = 0.75$; Finding 2). Third, metaphoricity and concreteness ratings were moderately negatively correlated ($r = -0.40$; Finding 3).

As shown in Table 4, all models replicated the positive relationships between familiarity and frequency (Finding 1), as well as between familiarity and aptness (Finding 2). Likewise, all models except DeepSeek-V3.2-Exp (Thinking) replicated the negative relationship between metaphoricity and concreteness (Finding 3). Figure 1 presents the plots for Findings 1–3 for the metaphor data set (first row) and the best-performing models (Mistral-Small-3.2-24B-Instruct-2506, Qwen3-32B, and Qwen3-Next-80B-A3B-Thinking). Although all dimensions exhibit the same correlation patterns observed in human data, the LLMs appear to be more categorical and dispersed, favoring extreme values at both ends—particularly Qwen3-Next-80B-A3B-Thinking—unlike humans, whose ratings are more evenly distributed.

5.3.2 Findings from the online study

We examined whether the subset of LLM-generated ratings replicated results from an online study of metaphor comprehension that used 24 metaphors from the metaphor dataset. In a maze task, participants read metaphors embedded in sentences (e.g., *John is an early bird so he can...*) and were slower and less likely to select the target continuation (e.g., *attend*) when it was paired with a distractor related to the literal meaning of the metaphor (e.g., *fly*) than with an unrelated distractor (e.g., *cry*). These findings suggest that the literal meaning of metaphors is accessed during comprehension and can be "awakened" by a subsequent literal cue (Pissani and de Almeida, 2022, 2023; Pissani et al., 2026b).

In a follow-up study, the magnitude of the *awakening* effect was found to vary with aptness and concreteness (Pissani et al., 2026a).[4] Aptness increased accuracy [$\beta = 0.86$, 95% CRI (0.17, 1.56); Finding 4] and decreased response time in the related-distractor condition [$\beta = -0.06$, 95% CRI (–0.12, –0.01); Finding 5], suggesting a reduced activation of the literal meaning. By contrast, concreteness increased response time in the related-distractor condition [$\beta = 0.05$, 95% CRI (0.01, 0.09); Finding 6], suggesting that the literal meaning is enhanced by concreteness during sentence comprehension.

Table 4 shows that the effect of aptness on accuracy (Finding 4) was replicated only by Apertus-70B-Instruct-2509. The effect of aptness on response time (Finding 5) was observed in Apriel-1.5-15B-Thinker, Qwen3-32B, and Qwen3-Next-80B-A3B-Thinking. The effect of concreteness on response time (Finding 6) was not replicated by any of the models. Given the weak alignment between model ratings and human norms for aptness and concreteness in the subset, the limited replication of these effects is not unexpected.

5.3.3 Findings from the offline study

We assessed the external validity of LLM-generated ratings for the most reliable dimension, familiarity. Given the limited empirical data available for two-word metaphor combinations like those normed in the metaphor dataset, we extended our analysis to predicate metaphors. Predicate metaphors occur in sentence form, with the metaphorical content primarily conveyed by the verb (e.g., *The violent image RATTLED in her head*). Although predicate metaphors differ in structure and complexity from two-word metaphors, the dimension of familiarity is defined consistently—that is, the extent to which an expression has been read or used previously. We therefore expected that models demonstrating reliability and validity in familiarity ratings for two-word metaphors would also yield robust familiarity ratings for predicate metaphors.

In the offline study, Stamenković et al. (2019)[5] examined whether familiarity influenced comprehension difficulty in a multiple-choice task, in which participants selected the best metaphorical interpretation of predicate metaphors from three options. Results showed that participants were more likely to select the correct choice for high-familiarity expressions than for low-familiarity expressions [$\beta = 0.29$, 95% CRI (0.03, 0.56)]. Although their study also included individual differences as predictors of metaphor comprehension, those analyses are beyond the scope of the present work.

Each model generated familiarity ratings for the 24 predicate metaphors used in Stamenković et al. (2019) study. For the substitution analysis, we selected six LLMs based on reliability (ICC ≥ 0.75) and validity (Spearman's $\rho \geq 0.40$, $p < 0.05$). The selected models were Mistral-Small-3.2-24B-Instruct-2506, Qwen3-32B, Apertus-70B-Instruct-2509, Llama-3.3-70B-Instruct, Qwen3-Next-80B-A3B-Thinking, and DeepSeek-V3.2-Exp (Thinking). Models excluded due to low or nonsignificant correlations were Apriel-1.5-15B-Thinker and gpt-oss-120b (see Table 5).

Results indicated that only Qwen3-32B [$\beta = 0.43$, 95% CRI (–0.09, 0.94)] and Qwen3-Next-80B-A3B-Thinking [$\beta = 0.15$, 95% CRI (–0.15, 0.45)] showed familiarity effects similar in direction to those observed with Cardillo et al. (2010) ratings (see Figure 2). Although the 95% credible intervals for both models include zero, the posterior probabilities of a positive effect were relatively high [Qwen3-32B: $p(\hat{\beta} > 0) = 0.95$; Qwen3-Next-80B-A3B-Thinking: $p(\hat{\beta} > 0) = 0.84$], suggesting that familiarity trends in these models broadly align with human judgments.

3 Since familiarity in context (FAMc) and in isolation (FAMi), as well as aptness in context (APTc) and in isolation (APTi), are highly correlated, we used the versions in context (FAMc, APTc) for this and all subsequent analyses involving familiarity or aptness.

4 Data for this project, including datasets, analysis scripts, and figures are available at https://osf.io/d2pna.

5 In Stamenković et al. (2019), 24 predicate metaphors were selected from Cardillo et al. (2010), with 12 low familiarity items and 12 high familiarity items, treating familiarity as a categorical predictor. In our analyses, we use numerical familiarity ratings, so this variable is treated continuously. Because of that, we re-ran their analysis using numerical familiarity and replicated the reported effects, which we compare with results from LLM-generated familiarity ratings.

TABLE 4 Summary of substitution analyses.

	The norming study			The online study		
Model	Finding 1	Finding 2	Finding 3	Finding 4	Finding 5	Finding 6
Metaphor dataset	0.43*	0.75*	−0.40*	0.86 [0.17, 1.56]	−0.06 [−0.12, −0.01]	0.05 [0.01, 0.09]
Apriel	0.43*	0.52*	−0.15·	0.73 [−0.83, 2.27]	−0.17 [−0.31, −0.04]	0.00 [−0.01, 0.02]
Mistral	0.36*	0.58*	−0.23*	−0.62 [−1.54, 0.31]	−0.04 [−0.11, 0.03]	−0.02 [−0.05, 0.00]
Qwen3	0.47*	0.55*	−0.22*	0.10 [−1.32, 1.52]	−0.11 [−0.20, −0.01]	−0.04 [−0.06, −0.01]
Apertus	0.34*	0.56*	−0.56*	0.95 [0.24, 1.68]	−0.04 [−0.08, 0.00]	0.00 [−0.02, 0.02]
Llama3	0.36*	0.53*	−0.13·	1.31 [−0.23, 2.86]	−0.07 [−0.20, 0.06]	−0.01 [−0.03, 0.00]
Qwen3–Next	0.36*	0.61*	−0.39*	0.56 [−1.23, 2.35]	−0.20 [−0.37, −0.03]	−0.05 [−0.10, 0.00]
gpt–oss	0.37*	0.69*	−0.13·	0.11 [−0.90, 1.14]	−0.04 [−0.12, 0.04]	0.01 [−0.01, 0.02]
DeepSeek	0.30*	0.66*	0.07	−0.26 [−0.81, 0.31]	−0.01 [−0.04, 0.02]	−0.01 [−0.03, 0.01]

Findings 1-3: Pearson correlations marked with asterisks (*) indicate $p < 0.01$, dots (·) indicate $p < 0.05$. Findings 4-6: CRIs including zero are shown in gray.

6 Discussion

We evaluated whether LLM-generated norms can serve as reliable and valid resources for psycholinguistic research, with a particular focus on metaphor comprehension. To this end, eight current LLMs produced ratings for 300 two-word metaphors on five dimensions—familiarity, aptness, concreteness, metaphoricity, and constituency—from a recently released metaphor dataset. These dimensions were selected because they tap distinct cognitive processes, allowing us to assess which aspects of metaphor comprehension are captured accurately by LLMs and which are not.

Reliability was assessed by measuring internal consistency across three random seeds for each model and dimension. Validity was assessed by comparing LLM-generated ratings against human "gold standard" norms from the metaphor dataset. Finally, to evaluate the utility of LLM-generated ratings as substitutes for human norms in psycholinguistic research, we replaced them into existing analyses of metaphor comprehension and examined whether they reproduced previously observed effects.

Norms collected for single words, whose meanings can vary substantially depending on context (e.g., *wooden table* vs. *data table*), may benefit from contextualized presentation (Trott, 2024). In contrast, norms collected for two-word metaphors appear to vary minimally as a function of context (Pissani and de Almeida, 2026). This may be because many metaphors can only be interpreted figuratively (e.g., *bright student*). Even when a literal interpretation is possible, the figurative sense tends to be more salient (e.g., *cold turkey*); whereas for expressions in which literal and figurative interpretations are equally salient (e.g., *hot water*), task instructions may further constrain figurative interpretation. We manipulated stimulus presentation to identify conditions in which model responses align most closely with human judgments. However, LLM-generated ratings of familiarity and aptness did not vary systematically with context presentation. Although familiarity ratings in context had a numerically higher correlation coefficient than ratings in isolation for the full set, both remained within the same correlation threshold. For the subset, only four models (Apriel-1.5-15B-Thinker, Qwen3-32B, Qwen3-Next-80B-A3B-Thinking, and gpt-oss-120b) showed a substantial improvement in familiarity alignment when metaphors were presented in context rather than in isolation—all other differences were small and inconsistent.

Response format was evaluated by measuring the same construct using numerical responses (constituency) and categorical responses (vehicle position). For numerical responses, reliability ranged from moderate to excellent in both the full set and the subset, whereas alignment with human norms was moderate to strong for most models. This relatively high alignment may reflect the use of mean ratings averaged across random seeds, which likely yields more precise estimates than individual seed outputs. Two models—Mistral-Small-3.2-24B-Instruct-2506 and Apertus-70B-Instruct-2509—were exceptions, showing very weak alignment. For categorical responses, roughly half of the models showed moderate to strong reliability, but of these, only Apriel-1.5-15B-Thinker and gpt-oss-120b were aligned with human judgments, indicating that some models were consistent yet inaccurate. Given the consistent reliability on the numerical scale, models appear to be more reliable when using numerical rather than categorical responses for the same construct. However, because alignment for the numerical scale was inconsistent, conclusions about the categorical scale remain limited, as poor performance may reflect the complexity of the dimension rather than the response format itself. Future work may examine the effect of response format using more reliable dimensions, such as familiarity and metaphoricity.

6.1 Are LLM-generated norms reliable and aligned with human responses?

Reliability metrics were consistently high across most dimensions and models, with the exception of vehicle position. This pattern indicates that, when presented with the same prompts, LLMs generally produce stable and reproducible outputs, supporting their reliability for generating comparable responses across studies.

Alignment with human ratings was moderate to high for most dimensions and models, with the exception of constituency and

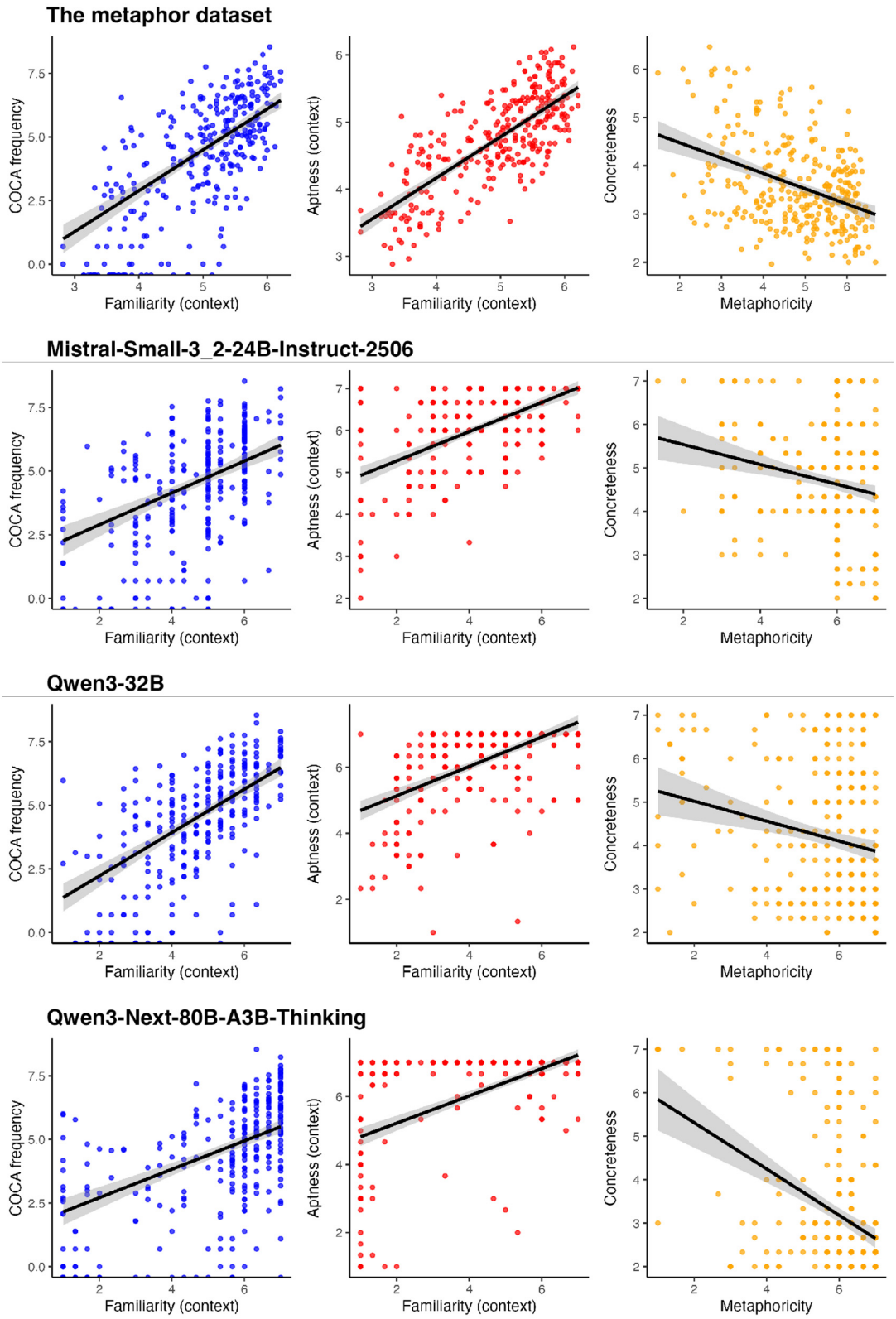

FIGURE 1
Correlation plots among dimensions for the metaphor dataset and the best-performing LLMs.

vehicle position. As expected, familiarity showed the strongest alignment, likely because it reflects frequency, which is a property that LLMs can readily estimate. Metaphoricity showed the next-highest alignment, suggesting that LLMs can distinguish expressions that are literally or metaphorically interpretable depending on context. In contrast, aptness requires evaluating the quality of the metaphorical mapping—that is, whether the vehicle conveys important features of the topic. For example, LLMs may struggle to assess whether the vehicle *sharp* effectively conveys the mental acuity of the topic *mind*. Concreteness involves relating the meaning of an expression to physical experience. In this case, LLMs may find it difficult to determine whether the 'mental acuity' conveyed by *sharp* can be experienced through the senses. Constituency and vehicle position reflect how metaphorical content is distributed across an expression, whether it is conveyed by the first word, the second word, or both. For example, in the expression *broken heart*, LLMs may differ in whether they attribute the metaphorical content to *broken*, *heart*, or their combination, referring to a love gone awry. While familiarity and

TABLE 5 Reliability and alignment with human ratings for familiarity ratings of predicate metaphors.

Model	ICC	ρ
Apriel-1.5-15B-Thinker	0.77	0.24
Mistral-Small-3.2-24B-Instruct-2506	0.95	0.64*
Qwen3-32B	0.81	0.75*
Apertus-70B-Instruct-2509	0.83	0.43·
Llama-3.3-70B-Instruct	0.93	0.81*
Qwen3-Next-80B-A3B-Thinking	0.73	0.42·
gpt-oss-120b	0.82	0.18
DeepSeek-V3.2-Exp (Thinking)	0.83	0.45·

Asterisks (*) indicate $p < 0.01$, dots (·) indicate $p < 0.05$.

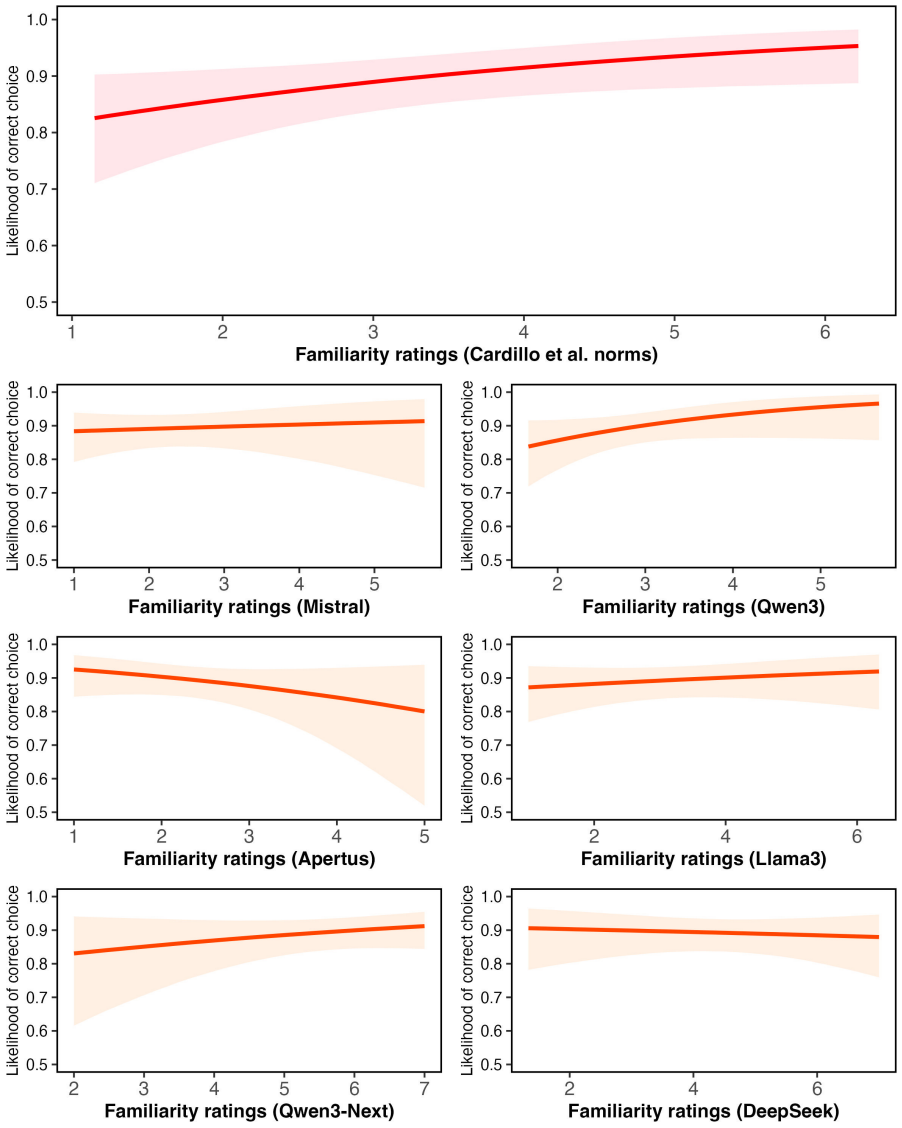

FIGURE 2
Posterior predictions of accuracy using familiarity ratings from Cardillo et al. (2010) and the best-performing LLMs.

metaphoricity rely primarily on judgments about the expression as a whole, aptness, concreteness, and constituency require identifying the topic and vehicle of the metaphor and evaluating their relationship—processes that LLMs may find particularly challenging because they involve integrating both semantic and syntactic information. As a result, these dimensions showed only moderate alignment with human ratings in the full dataset and largely no alignment in the subset, suggesting that analyses based on them should be interpreted with caution. Consistent with these findings, Mangiaterra et al. (2025) reported that GPT performs better on ratings of familiarity and comprehensibility of metaphors than on more complex dimensions such as imageability (i.e., the ease of generating a mental image form an expression), which, like concreteness, depends on sensory experience.

6.2 Do LLM-generated norms replicate observed effects and could they support psycholinguistic research?

Whether LLMs can be used to norm linguistic stimuli appears to depend on the dimension. Well-established and relatively straightforward dimensions, such as familiarity and metaphoricity, are predicted more accurately by LLMs, showing promise for psycholinguistic research. In contrast, more complex dimensions that require parsing the expression and understanding the source–target mapping of a metaphor, as well as sensitivity to sensory or physical experience, remain less reliable and therefore less suitable for research use.

All models replicated the three relationships between dimensions observed in the metaphor dataset: a positive correlation between frequency and familiarity, a positive correlation between familiarity and aptness, and a negative correlation between concreteness and metaphoricity. These results show that LLM-generated norms can capture the same patterns seen in human data, suggesting they may be a useful tool for studying relationships between dimensions—at least when a sufficient number of items is available.

Nevertheless, even when LLM-generated norms were highly reliable and aligned with human ratings, they did not consistently reproduce effects observed in metaphor comprehension studies. Substituting LLM-generated ratings of aptness and concreteness to predict response times and accuracy in the online study yielded mixed results, possibly due to the complexity of these dimensions or features of this particular subset. Similarly, in the offline comprehension study, only two of the six models that met reliability and alignment criteria showed familiarity effects in the same direction as those reported for the predicate metaphor dataset.

6.3 Best-practice recommendations for generating and using LLM-generated norms in psycholinguistic research

Based on our empirical results, we propose several best practices for prompting LLMs in psycholinguistic research. First, to ensure robustness and reliability, researchers should not rely on a single generation but rather employ *multiple seed variations*. Our methodology sampled three independent outputs for each item by varying the random seed to obtain a single average rating. This approach allows for the quantification of internal consistency and numerical precision issues across generations.

Second, we recommend utilizing models or prompts that encourage explicit reasoning or *Chain-of-Thought* processing. Our study included "thinking" variants designed to generate step-by-step justifications, and results indicated that the Qwen3-Next-80B-A3B-Thinking, gpt-oss-120b, and DeepSeek-V3.2-Exp (Thinking) achieved excellent reliability across most dimensions, even surpassing human raters in some instances.

Third, prompts should combine *explicit structured constraints* with task instructions that remain as close as possible to those given to human participants. In our study, all models received the same prompts under controlled experimental manipulations, yet the dimension of constituency and its categorical equivalent (vehicle position) yielded low reliability and validity. To mitigate such failures in complex structural tasks, we advise augmenting standard instructions with *few-shot prompting* examples to better guide the model toward the desired output format.

Fourth, to support robust evaluation, researchers should use sufficiently large datasets. In our study, alignment with human ratings was moderate to strong for the full set of 300 items but decreased substantially in the subset of 24 items, which indicates that smaller datasets may provide insufficient variance for stable assessment.

Finally, prompting LLMs across different numerical scales or categorical labels is essential for understanding their behavior and measuring consistency, as our comparison of numerical vs. categorical formats demonstrated that performance and alignment with human judgments can vary depending on the required output type.

7 Conclusion

In this study, we examined the extent to which large language models align with human judgments across multiple dimensions of two-word metaphor combinations, including familiarity, aptness, metaphoricity, concreteness, and constituency. Overall, our results indicate that LLMs show moderate to high alignment with human ratings for dimensions that depend on word co-occurrence, such as familiarity and metaphoricity, but substantially weaker alignment for dimensions that require parsing the expression and making judgments based on the relationship between the topic an vehicle. To our knowledge, this is the first study to compare the performance of eight LLMs, including "thinking" variants, in the norming of metaphors.

7.1 Limitations

While our study provides insight into how LLMs align with human judgments of metaphor, several limitations should be considered. First, our dataset consists exclusively of English metaphors, which may limit the generalizability of

our findings to other languages—particularly given that English is the best-represented language in most LLM training data (Zhang et al., 2023). Second, although all models are open weights—allowing anyone with sufficient compute to run them locally—they are not fully open source, limiting transparency regarding their architecture, training data, and optimization (Ivanova, 2025). While the metaphor dataset was published *after* LLM training, earlier versions of these norms may have been included in the training data. Third, LLM-generated norms were obtained via prompting, which may produce less reliable estimates than direct probability-based measurements (Hu and Levy, 2023).

Data availability statement

The original contributions presented in the study are publicly available. The data, including prompts, datasets, and analysis scripts, can be found at the OSF repository at: https://osf.io/bmnyw.

Ethics statement

This study involved secondary analysis of previously collected human data from a prior study approved by the Concordia University Human Research Ethics Committee (Approval No. 10000023). No new participant recruitment or data collection was conducted. Data use complied with the original ethical approval and applicable institutional guidelines.

Author contributions

LP: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. MJ: Data curation, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. VD: Funding acquisition, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This research was funded by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 948878). MJ was supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project ID 389792660-TRR 248 (Foundations of Perspicuous Software Systems).

Acknowledgments

We are grateful to Dušan Stamenković for providing the data and materials from the offline metaphor study, which allowed external validation of our findings. The authors gratefully acknowledge the computing time made available to them on the high-performance computer at the NHR Center of TU Dresden. This center is jointly supported by the Federal Ministry of Research, Technology and Space of Germany and the state governments participating in the NHR (https://www.nhr-verein.de/unsere-partner).

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/flang.2026.1756514/full#supplementary-material

References

Al-Azary, H., Gagné, C. L., and Spalding, T. L. (2021). Flute birds and creamy skies: the metaphor interference effect in modifier-noun phrases. *Can. J. Exp. Psychol.* 75, 175–181. doi: 10.1037/cep0000251

Arzouan, Y., Goldstein, A., and Faust, M. (2007). Brainwaves are stethoscopes: ERP correlates of novel metaphor comprehension. *Brain Res.* 1160, 69–81. doi: 10.1016/j.brainres.2007.05.034

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., et al. (2025). "LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks," in *Proceedings of the 63rd annual meeting of the association for computational linguistics (Volume 2: Short Papers)* (Vienna: Association for Computational Linguistics), 238–255. doi: 10.18653/v1/2025.acl-short.20

Blasko, D. G., and Connine, C. M. (1993). Effects of familiarity and aptness on metaphor processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 19:295. doi: 10.1037/0278-7393.19.2.295

Bowdle, B. F., and Gentner, D. (2005). The career of metaphor. *Psychol. Rev.* 112:193. doi: 10.1037/0033-295X.112.1.193

Briner, S. W., Schutzenhofer, M. C., and Virtue, S. M. (2018). Hemispheric processing in conventional metaphor comprehension: the role of general knowledge. *Neuropsychologia* 114, 101–109. doi: 10.1016/j.neuropsychologia.2018.03.040

Canal, P., Bischetti, L., Bertini, C., Ricci, I., Lecce, S., Bambini, V., et al. (2022). N400 differences between physical and mental metaphors: the role of theories of mind. *Brain Cogn.* 161:105879. doi: 10.1016/j.bandc.2022.105879

Cardillo, E. R., Schmidt, G. L., Kranjec, A., and Chatterjee, A. (2010). Stimulus design is an obstacle course: 560 matched literal and metaphorical sentences for testing neural hypotheses about metaphor. *Behav. Res. Methods* 42, 651–664. doi: 10.3758/BRM.42.3.651

Chiappe, D., Kennedy, J. M., and Chiappe, P. (2003). Aptness is more important than comprehensibility in preference for metaphors and similes. *Poetics* 31, 51–68. doi: 10.1016/S0304-422X(03)00003-2

Chiappe, D. L., and Kennedy, J. M. (1999). Aptness predicts preference for metaphors or similes, as well as recall bias. *Psychon. Bull. Rev.* 6, 668–676. doi: 10.3758/BF03212977

Clark, H. H., and Lucy, P. (1975). Understanding what is meant from what is said: a study in conversationally conveyed requests. *J. Verbal Learning Verbal Behav.* 14, 56–72. doi: 10.1016/S0022-5371(75)80006-5

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., et al. (2024). DeepSeek-V3 technical report. *arXiv [preprint]*. arXiv:2412.19437. Covers the DeepSeek-V3 series, including DeepSeek-V3.2-Exp (Thinking; 685B).

Dillion, D., Tandon, N., Gu, Y., and Gray, K. (2023). Can AI language models replace human participants? *Trends Cogn. Sci.* 27, 597–600. doi: 10.1016/j.tics.2023.04.008

DiStefano, P. V., Patterson, J. D., and Beaty, R. E. (2025). Automatic scoring of metaphor creativity with large language models. *Creat. Res. J.* 37, 555–569. doi: 10.1080/10400419.2024.2326343

Forgács, B., Bardolph, M. D., Amsel, B. D., DeLong, K. A., and Kutas, M. (2015). Metaphors are physical and abstract: ERPs to metaphorically modified nouns resemble ERPs to abstract language. *Front. Hum. Neurosci.* 9:28. doi: 10.3389/fnhum.2015.00028

Gagné, C. L. (2002). Metaphoric interpretations of comparison-based combinations. *Metaphor Symb.* 17, 161–178. doi: 10.1207/S15327868MS1703_1

Gamer, M., Lemon, J., Gamer, M. M., Robinson, A., and Kendall's, W. (2012). Package 'irr'. *Various Coefficients of Interrater Reliability and Agreement,* 22, 1–32.

Genovesi, C. (2020). Metaphor and what is meant: metaphorical content, what is said, and contextualism. *J. Pragmat.* 157, 17–38. doi: 10.1016/j.pragma.2019.11.002

Genovesi, C. (2023). Grice's café: coffee, cream, and metaphor comprehension. *Front. Commun.* 8:1175587. doi: 10.3389/fcomm.2023.1175587

Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. *Cogn. Sci.* 7, 155–170. doi: 10.1016/S0364-0213(83)80009-3

Giora, R. (1997). Understanding figurative and literal language: the graded salience hypothesis. *Cogn. Linguist.* 7, 183–206. doi: 10.1515/cogl.1997.8.3.183

Goldstein, A., Arzouan, Y., and Faust, M. (2012). Killing a novel metaphor and reviving a dead one: ERP correlates of metaphor conventionalization. *Brain Lang.* 123, 137–142. doi: 10.1016/j.bandl.2012.09.008

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al. (2024). The Llama 3 herd of models. *arXiv [preprint]*. arXiv:2407.21783. The technical report covering the Llama 3 family, including Llama-3.3-70B-Instruct.

Grice, H. P. (1975). *Logic and Conversation, Volume 3*. Berlin: BRILL. doi: 10.1163/9789004368811_003

Hernández-Cano, A., Hägele, A., Huang, A. H., Romanou, A., Solergibert, A.-J., et al. (2025). Apertus: democratizing open and compliant LLMs for global language environments. *arXiv [preprint]*. arXiv:2509.14233. Covers the Apertus-70B-Instruct-2509 model.

Holyoak, K. J. (2013). "Analogy and relational reasoning," in *The Oxford Handbook of Thinking and Reasoning*, eds. K. J. Holyoak, R. G. Morrison (Oxford: Oxford University Press), 234–259. doi: 10.1093/oxfordhb/9780199734689.013.0013

Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., and Gibson, E. (2023). "A fine-grained comparison of pragmatic language understanding in humans and language models," in *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, eds. A. Rogers, J. Boyd-Graber, and N. Okazaki (Association for Computational Linguistics), 4194–4213. doi: 10.18653/v1/2023.acl-long.230

Hu, J., and Levy, R. (2023). Prompting is not a substitute for probability measurements in large language models. *arXiv [preprint]*. arXiv:2305.13264. doi: 10.18653/v1/2023.emnlp-main.306

Ichien, N., Stamenković, D., and Holyoak, K. J. (2024). Large language model displays emergent ability to interpret novel literary metaphors. *Metaphor Symb.* 39, 296–309. doi: 10.1080/10926488.2024.2380348

Ivanova, A. A. (2025). How to evaluate the cognitive abilities of llms. *Nat. Hum. Behav.* 9, 230–233. doi: 10.1038/s41562-024-02096-z

Kavé, G., Gavrieli, R., and Mashal, N. (2014). Stronger left-hemisphere lateralization in older versus younger adults while processing conventional metaphors. *Laterality* 19, 705–717. doi: 10.1080/1357650X.2014.905584

Koo, T. K., and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* 15, 155–163. doi: 10.1016/j.jcm.2016.02.012

Lai, V. T., Curran, T., and Menn, L. (2009). Comprehending conventional and novel metaphors: an ERP study. *Brain Res.* 1284, 145–155. doi: 10.1016/j.brainres.2009.05.088

Lakoff, G., and Johnson, M. (1980). *Metaphors We Live by*. Chicago, IL: University of Chicago Press.

Levorato, M. C., and Cacciari, C. (2002). The creation of new figurative expressions: psycholinguistic evidence in Italian children, adolescents and adults. *J. Child Lang.* 29, 127–150. doi: 10.1017/S0305000901004950

Mangiaterra, V., Al-Azary, H., Pietro, C. B. S., Canal, P., and Bambini, V. (2025). Can GPT replace human raters? Validity and reliability of machine-generated norms for metaphors. *arXiv [preprint]*. arXiv:2512.12444.

Martínez, G., Molero, J. D., González, S., Conde, J., Brysbaert, M., and Reviriego, P. (2025). Using large language models to estimate features of multi-word expressions: concreteness, valence, arousal. *Behav. Res. Methods* 57:5. doi: 10.3758/s13428-024-02515-z

Marzi, G., Balzano, M., and Marchiori, D. (2024). K-alpha calculator-krippendorff's alpha calculator: a user-friendly tool for computing krippendorff's alpha inter-rater reliability coefficient. *MethodsX*, 12:102545. doi: 10.1016/j.mex.2023.102545

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochem. Med.* 22, 276–282. doi: 10.11613/BM.2012.031

Mistral AI Team (2025). *Mistral Small 3.2Model Card and API Documentation. Official Documentation. A Minor Update to the Mistral-Small Series with 24B Parameters.*

OpenAI Team (2025). *Technical Report: Performance and Baseline Evaluations of GPT-OSS-Safeguard-120b and GPT-OSS-Safeguard-20b. OpenAI Technical Report*. Covers the GPT-OSS series, including the 120B parameter model.

Park, J., Sana, F. Gagné, C. L., and Spalding, T. L. (2021). Factors that influence the processing of noun-noun metaphors. *Metaphor Symb.* 36, 20–44. doi: 10.1080/10926488.2020.1843970

Pavlovic, M., and Poesio, M. (2024). Understanding the effect of temperature on alignment with human opinions. *arXiv [preprint]*. arXiv:2411.10080.

Pissani, L., and de Almeida, R. G. (2022). Can you mend a broken heart? Awakening conventional metaphors in the maze. *Psychon. Bull. Rev.* 29, 253–261. doi: 10.3758/s13423-021-01985-y

Pissani, L., and de Almeida, R. G. (2023). Early birds can fly: awakening the literal meaning of conventional metaphors further downstream. *Metaphor Symb.* 38, 346–362. doi: 10.1080/10926488.2023.2225561

Pissani, L., and de Almeida, R. G. (2026). Metaphors in context and in isolation: familiarity, aptness, concreteness, metaphoricity, and structure norms for 300 two-word expressions. *Behav. Res. Methods* 58:31. doi: 10.3758/s13428-025-02902-0

Pissani, L., Genovesi, C., and de Almeida Roberto, G (2026a). *Aptness Reduces, But Concreteness Enhances Access to the Literal Meaning of a Metaphor During Sentence Comprehension*. Available online at: https://osf.io/d2pna (Accessed November 24, 2025).

Pissani, L., Scholman, M. C., and Demberg, V. (2026b). Working memory and vocabulary knowledge sustain activation of the literal meaning of metaphors during sentence comprehension. *PsyArXiv* [preprint]. Available online at: https://osf.io/preprints/psyarxiv/dqanh_v1 (Accessed February 4, 2026).

Qwen Team (2025). Qwen3 technical report. *arXiv [preprint]*. arXiv:2505.09388. Covers the Qwen3 series, including Qwen3-32B.

Radhakrishna, S., Tiwari, A., Shukla, A., Hashemi, M., Maheshwary, R., Malay, S. K. R., et al. (2025). Apriel-1.5-15B-thinker: mid-training is all you need. *arXiv [preprint]*. arXiv:2510.01141. Model with 15B parameters, focusing on reasoning capacity.

Revelle, W. (2025). *Psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 2.5.3. Evanston, IL: Northwestern University.

Roncero, C., and de Almeida, R. G. (2014). The importance of being APT: metaphor comprehension in Alzheimer's disease. *Front. Hum. Neurosci.* 8:973. doi: 10.3389/fnhum.2014.00973

Roncero, C., and de Almeida, R. G. (2015). Semantic properties, aptness, familiarity, conventionality, and interpretive diversity scores for 84 metaphors and similes. *Behav. Res. Methods* 47, 800–812. doi: 10.3758/s13428-014-0502-y

Searle, J. R. (Ed.) (1979). "Metaphor," in *Expression and Meaning. Studies in the Theory of Speech Acts* (Cambridge: Cambridge University Press), 76–116. doi: 10.1017/CBO9780511609213.006

Stamenković, D., Ichien, N., and Holyoak, K. J. (2019). Metaphor comprehension: an individual-differences approach. *J. Mem. Lang.* 105, 108–118. doi: 10.1016/j.jml.2018.12.003

Trott, S. (2024). Can large language models help augment English psycholinguistic datasets? *Behav. Res. Methods* 56, 6082–6100. doi: 10.3758/s13428-024-02337-z

Zhang, X., Li, S., Hauer, B., Shi, N., and Kondrak, G. (2023). Don't trust chatgpt when your question is not in English: a study of multilingual abilities and types of LLMs. *arXiv [preprint]*. arXiv:2305.16339. doi: 10.48550.arXiv:2305.16339