A Cognitively Motivated Multidimensional Framework for Evaluating Metaphor Explanations

Ana Naveriani[1], Jakob Suchan[1], Stefano Zoia[2], Mehul Bhatt[3], Antonio Lieto[4], Gian Luca Pozzato[2],

[1]Constructor University, Bremen, Germany, [2]University of Turin, Italy,
[3]Örebro University, Sweden, [4]University of Salerno, Italy

Correspondence: jsuchan@constructor.university

Abstract

Current evaluation of metaphor explanations relies mainly on holistic quality ratings, revealing little about how explanation quality is structured or where human judgments agree and diverge. We introduce a cognitively motivated framework that decomposes metaphor explanation quality into six theoretically grounded dimensions. In a dense annotation study (11,200 ratings), we find that: **(i)** explanation quality is genuinely multidimensional; **(ii)** annotator disagreement is systematic rather than random; and **(iii)** the six dimensions collapse into a shared cluster and two independent axes of judgment. An exploratory feasibility study further shows that a standard automatic evaluation pipeline can recover parts of this structure, predicting the most discriminative dimensions well while its errors correlate human (dis)agreement. Together, these results suggest that multidimensional evaluation offers richer diagnostic insight than holistic ratings, and that automatic evaluators for open-ended generation tasks should be judged on how well they preserve the structure of human judgment.

1 Introduction

Metaphors are a pervasive feature of human language and thought (Lakoff and Johnson, 1980, 1999), enabling people to understand abstract concepts through more familiar and concrete experiences. For instance, consider the following passage:

"Life is a journey where time is money and the mind works like a computer. But an Achilles' heel may prove difficult to overcome, and a heart of stone can keep us from what truly matters."

While most speakers intuitively understand these expressions, explaining *why* they convey meaning is considerably more difficult (Glucksberg, 2001). Effective explanations must reveal underlying conceptual mappings, account for emergent meaning, connect figurative language to familiar experiences, and provide sufficient context for interpretation. Consequently, metaphor explanation is not merely a linguistic task, but also a cognitive and communicative one. Research in cognitive linguistics has studied metaphors as a cognitive mechanism through which abstract concepts are understood via mappings from more concrete domains (Lakoff and Johnson, 1980, 1999). Computational research has focused primarily on metaphor identification, interpretation, and generation (Shutova et al., 2010; Tsvetkov et al., 2014), while recent large language models have substantially improved the quality of ex-post generated metaphor explanations (Ichien et al., 2024). However, comparatively little work has focused on their systematic evaluation. As such, evaluation remains a central challenge in Natural Language Generation (NLG). Traditional automatic metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2020) often correlate poorly with human judgments for open-ended generation tasks (Schmidtova et al., 2024). Consequently, recent work has increasingly emphasized human-centered and multidimensional evaluation frameworks (Howcroft et al., 2020; Celikyilmaz et al., 2021; van der Lee et al., 2021), while LLM-based evaluators have emerged as promising alternatives for automatic assessment (Zheng et al., 2023; Liu et al., 2023; Huang et al., 2025). Furthermore, human disagreement is increasingly recognized as informative factor for evaluation (Uma et al., 2021).

Existing work on metaphor generation and explanation typically relies on holistic quality ratings or a small number of task-specific criteria (Chakrabarty et al., 2021; Stowe et al., 2021). Such evaluations provide limited insight into which aspects of an explanation influence human judgments or where evaluators systematically disagree. We address this gap by investigating how humans apply different dimensions of metaphor explanation. Our

Proceedings of the 19th International Natural Language Generation Conference: Short Papers, pages 108–118
October 17-21, 2026 ©2026 Association for Computational Linguistics

CRITERION	DEFINITION	THEORETICAL BASIS	EXAMPLE
NON-CIRCULARITY	The explanation avoids restating the metaphor in equivalent figurative terms.	Principle of Experiential Grounding (Lakoff and Johnson, 1980)	Explaining *Time is Money* as "time is valuable because it is like money" is circular.
ORIGIN AND CULTURAL ACCURACY	The explanation captures relevant cultural, historical, or contextual background.	Cognitive Grammar (Langacker, 2008)	*Achilles' Heel* requires knowledge of its origin in Greek mythology.
MAPPING	The explanation identifies the conceptual relation between source and target domains.	Conceptual Metaphor Theory (Lakoff and Johnson, 1980)	*Life is a Journey* maps paths, obstacles, and destinations onto stages, challenges, and goals in life.
EMERGENCE	The explanation captures the non-literal meaning that arises from the metaphor.	Blended Space Theory (Fauconnier and Turner, 2002)	*Heart of Stone* conveys emotional coldness and lack of empathy, meanings that go beyond the literal properties of stone.
ACCESSIBILITY	The explanation is understandable to the intended audience.	Grice's Maxim of Manner (Grice, 1975)	Explaining *Life is a Journey* using familiar experiences is more accessible than relying on technical linguistic terminology.
PURPOSE	The explanation clarifies why the metaphor is communicatively effective or useful.	Grice's Maxim of Relevance (Grice, 1975)	*The Mind is a Computer* frames cognition as information processing, making complex mental processes easier to understand.

Table 1: Proposed evaluation dimensions for metaphor explanations and their theoretical grounding.

contributions are: **(1)** introducing a cognitively motivated multidimensional framework for evaluating metaphor explanations; **(2)** analysing these dimensions in terms of reliability, redundancy, and disagreement; and **(3)** conducting an exploratory automatic evaluation study investigating how standard language encoders can recover these aspects.

2 A Multidimensional Framework for Evaluating Metaphor Explanations

Building on insights from Conceptual Metaphor Theory (Lakoff and Johnson, 1980, 1999), analogical reasoning (Gentner, 1983), psycholinguistic models of metaphor comprehension (Glucksberg, 2001; Bowdle and Gentner, 2005), and recent cognitively inspired computational approaches (Lieto et al., 2025; Cappa et al., 2025), we propose a multidimensional framework for evaluating metaphor explanations. The framework comprises six dimensions, capturing complementary aspects considered when assessing whether an explanation is useful, meaningful, and informative (Table 1): *Non-Circularity* evaluates informational value, *Origin & Cultural Accuracy* contextual grounding, *Mapping* conceptual correspondence, *Emergence* derived figurative meaning, 'Accessibility' communicative clarity, and 'Purpose' communicative function. In essence, the framework combines complementary constructs from cognitive linguistics and pragmatics into an operational annotation scheme rather than directly operationalizing any single theoretical account.

3 Analysis of Human Judgments

To examine whether the proposed dimensions correspond to consistent and structured aspects of human judgment, we conducted a dense annotation study in which metaphor explanations were independently rated across all six dimensions and an additional holistic overall quality score. Analysis focuses on investigating the reliability, the structure, and the (dis)agreement between participants.

3.1 Corpus and Annotation

The corpus consists of 100 metaphor explanations, each independently rated by 16 annotators on all six dimensions and an overall quality score, using a five-point Likert scale, yielding a dense corpus of $11,200$ individual human quality ratings (16 participants $\times\ 6 + 1$ dimensions $\times\ 100$ metaphor explanations).

Metaphor Selection & Explanation Generation. The corpus was curated from linguistic databases, idiom collections, and natural language sources, covering diverse conceptual domains including physical, social, cultural, and abstract metaphors. For each metaphor, one explanation was generated. The benchmark was deliberately engineered to provide variation across all evaluation dimensions, comprising 15 reference explanations satisfying all six evaluation dimensions, 60 targeted failures violating exactly one dimension (10 per dimension), and 25 mixed failures violating multiple dimensions. While these design profiles specify which criterion an explanation was intended to violate, they do not provide a target score and are not used as ground-truth labels or supervision in the analysis. The controlled construction instead serves to expose annotators and models to both successful and unsuccessful realizations of each criterion. Out of the 100 metaphors 15 have source domains traceable to a specific cultural, historical, or institutional tradition. These include examples grounded in Greek and Roman mythology (*a Her-*

Dimension	α	ICC(2,1)	ICC(2,k)
Non-Circularity	0.370	0.377	0.906
Origin and Cultural Accuracy	0.423	0.433	0.924
Mapping	0.423	0.427	0.922
Emergence	0.579	0.583	0.957
Accessibility	0.456	0.460	0.932
Purpose	0.490	0.494	0.940
Overall Quality	0.525	0.528	0.947

Table 2: Inter-rater reliability by dimension (100 metaphors, 16 raters).

	NC	OR	MP	EM	AC	PU
NC	1.00	−0.07	0.12	0.13	0.00	0.17
OR	−0.07	1.00	0.00	0.01	0.21*	0.02
MP	0.12	0.00	1.00	0.76**	0.47**	0.78**
EM	0.13	0.01	0.76**	1.00	0.49**	0.89**
AC	0.00	0.21*	0.47**	0.49**	1.00	0.54**
PU	0.17	0.02	0.78**	0.89**	0.54**	1.00

Table 3: Pairwise Spearman correlations ($^{*}p < .05$; $^{**}p < .0001$).

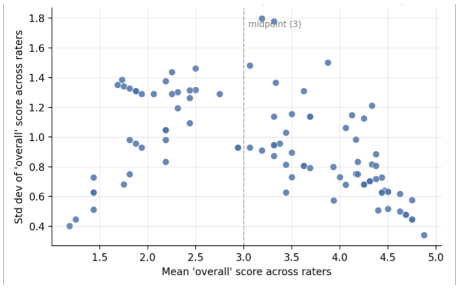


Figure 1: Mean score and rater deviation.

Predictor	Std. β	p
Non-Circularity	0.062	<.001
Origin and Cultural Accuracy	−0.070	<.001
Mapping	0.061	.004
Emergence	0.499	<.001
Accessibility	0.014	.420
Purpose	0.290	<.001

Table 4: Standardized partial regression coefficients predicting Overall quality from the six criteria.

culean task, Midas touch, Pandora's box), European history (*crossing the Rubicon, meeting their Waterloo*), Judeo-Christian and Hindu traditions, as well as culturally situated games, food, and drink. For the remaining 85 metaphors, the criterion is not applicable.

Participants. Each metaphor-explanation pair was independently evaluated according to detailed evaluation guidelines by 16 participants across the six proposed evaluation dimensions alongside an additional holistic overall quality score on a five-point Likert scale. Participants self-reported fluent-to-B2 English, spanning nine native languages, three reported linguistics/literature training, and ages range 18–73 (median = 21). All were recruited via personal contact.

3.2 Reliability, Structure, and Disagreement

In the following we present analysis of inter-rater reliability, the correlational structure among dimensions, and the relationship between disagreement and explanation quality. Key results are:

Agreement varies systematically across dimensions. We assess inter-rater reliability using Krippendorff's α, a chance-corrected measure of inter-annotator agreement, and ICC, with $\mathrm{ICC}(2,1)$ measuring the reliability of an individual rating and $\mathrm{ICC}(2,k)$ the reliability of the mean rating across the $k = 16$ participants. The results for each dimension (Table 2) shows that agreement varied substantially across dimensions, with Emergence showing the highest agreement ($\alpha = 0.579$), while Non-Circularity shows the lowest ($\alpha = 0.370$). No dimension reaches the conventional threshold for "reliable" agreement ($\alpha \geq 0.8$), and most support only tentative conclusions. This suggests that annotators can apply these dimensions consistently enough to support the structural analyses, but not consistently enough to treat any single dimension's mean rating as a precise ground-truth signal in isolation. ICC(2,1) tracks α for every dimension, providing a cross-check that these estimates are not artifacts of the chosen reliability measure.

The evaluation dimensions form two distinct groups. Pairwise Spearman correlations between the six dimensions (Table 3) show that two dimensions (Non-Circularity and Origin & Cultural Accuracy) are near-independent of every other dimension ($|\rho| \leq 0.21$). In contrast, Mapping, Emergence, and Purpose form a strongly correlated cluster ($\rho = 0.76$–0.89, $p < 10^{-25}$), and Accessibility correlates moderately with this cluster ($\rho \approx 0.47$–0.54). This structure suggests that the six dimensions do not function as six equally distinct criteria of judgment. Mapping, Emergence, and Purpose capture closely related aspects of explanation quality, while Non-Circularity and Origin & Cultural Accuracy behave as rather independent criteria.

Emergence and Purpose dominate the prediction of overall quality. As several dimensions are highly correlated, pairwise correlations may overestimate their independent contribution to the overall quality. Therefore we fit a linear model predicting Overall quality from all six standardized dimensions simultaneously (Table 4). Emer-

CRITERION	MAE	R^2	Weight
Non-Circularity	0.208	0.194	0.020
Origin and Cultural Accuracy	0.263	0.129	0.044
Mapping	0.354	0.509	0.046
Emergence	0.343	0.755	0.746
Accessibility	0.260	0.336	0.045
Purpose	0.350	0.684	0.099
OVERALL QUALITY	0.689	0.373	–

Table 5: Criterion-level automatic evaluation. (GroupK-Fold split, 80 train / 20 test metaphors).

gence ($\beta = 0.499$, $p < .001$) remained the strongest independent predictor, followed by Purpose ($\beta = 0.290, p < .001$). The remaining dimensions contributed only comparatively small effects, with Accessibility showing no significant independent association. These results suggest that Emergence and Purpose account for most of the independent predictive signal, while Mapping, Emergence, and Purpose remain partially redundant rather than representing fully independent criteria of judgment.

Disagreement is structured. Rating disagreement followed a clear U-shaped relationship with explanation quality (Figure 1). Annotators agreed most on explanations receiving consistently high or low ratings and disagreed most on borderline cases. For example, *"Love is a Fine Wine"* received near-unanimous ratings, whereas *"A Dead End"* produced the largest variation across annotators, suggesting that disagreement reflects ambiguity in explanation quality.

4 Preliminary Automatic Evaluation

In the following we explore the question of whether a standard automatic evaluation pipeline can recover any of this structure.[1] For this, we use a comparatively lightweight supervised BERT encoder to test whether the proposed criterion-level judgments can be predicted directly from explanation text and whether their relative importance reflects the structure observed in the human analysis.

Pipeline. We fine-tuned a BERT-base encoder to predict the six dimensions from the metaphor and its explanation. Predicted scores were combined into an overall quality estimate using a gradient-boosted judge model. We evaluate using grouped k-fold cross-validation at the metaphor level (no metaphor appears in both training and test folds within a fold).

[1]This study serves as a feasibility check. We do not claim competitive automatic evaluation.

METRIC	VALUE
Pearson r (σ vs. Error)	0.397 ($p = .083$)
Spearman ρ (σ vs. Error)	0.203 ($p = .392$)
Within Noise Band (Error $\leq \sigma$)	13/20 (65.0%)
Mean / Median Gold-Adjusted Error	0.814 / 0.835

Table 6: Relationship between human disagreement and model error on 20 held-out test metaphors.

4.1 Results

Preliminary findings from the feasibility study are:

Prediction performance reflects the multidimensional structure. Table 5 reports per-dimension mean absolute error (MAE) on the five-point rating scale, the proportion of variance in the human ratings explained by the model (R^2), and the relative feature importance assigned by the gradient-boosting model to each predicted dimension when estimating Overall quality. The pipeline predicts Emergence most accurately ($R^2 = 0.76$), followed by Purpose ($R^2 = 0.68$) and Mapping ($R^2 = 0.51$). Prediction performance is lower for Accessibility ($R^2 = 0.34$), Non-Circularity ($R^2 = 0.19$), and Origin & Cultural Accuracy ($R^2 = 0.13$). Overall quality is predicted with MAE$= 0.69$, $R^2 = 0.37$.

Learned structure partially mirrors the human structure. The judge model assigns most weight to Emergence and Purpose, the same two dimensions identified as the strongest independent predictors of Overall quality in the human analysis. Given that Emergence and Purpose are themselves strongly correlated ($\rho = 0.89$) and are also the two best-predicted dimensions ($R^2 = 0.76$ and 0.68, respectively), this suggests that the automatic evaluator captures the dominant structure underlying human judgments while assigning comparatively little weight to the remaining dimensions.

Model error shows a non-significant association with human disagreement. Table 6 compares model error against human rating variability on the 20 held-out test metaphors. Model error correlates positively with human disagreement in the expected direction (Pearson $r = 0.397$, $p = .083$; Spearman $\rho = 0.203$, $p = .392$), but neither correlation reaches significance, and only 13 of 20 predictions (65%) fall within one human standard deviation of the mean rating. The direction of the association is consistent with that hypothesis, however, it should be confirmed with a larger test set.

5 Discussion and Outlook

We have presented a cognitively motivated multidimensional framework for evaluating metaphor explanations. Through an empirical study, we also show how the dimensions, i.e., the six complementary cognitive constructs, exhibit distinct patterns of reliability, redundancy, and disagreement. Furthermore, an exploratory automatic evaluation pipeline partially recovered this structure. The diagnostic insight provided by this multidimensional evaluation, together with the finding that explanation quality is organized around a small number of dominant underlying dimensions, provides a basis for developing future automatic evaluation systems (e.g., utilizing LLM-as-a-judge (Huang et al., 2025)) that explicitly assess these dimensions before aggregating them into an overall evaluation, thereby producing judgments that are both more interpretable and more closely aligned with the multidimensional structure of human judgments. Beyond evaluation, future work will explore how the proposed dimensions can inform the *generation* of metaphor explanations, building on declarative approaches to semantically grounded explanation and natural language externalisation (Suchan et al., 2018a, 2015), and abductive approaches for constructing explanatory interpretations (Suchan et al., 2018b, 2021).

The strong correlations among Mapping, Emergence, and Purpose do not necessarily imply that these dimensions should be collapsed. They represent theoretically distinct properties and provide different diagnostic information: an explanation may, for example, correctly identify a mapping without adequately conveying why the metaphor is used. Because our benchmark deliberately includes criterion-specific failures, future work should validate the framework on naturally occurring explanations produced by humans and contemporary language models, and investigate whether the identified core dimensions generalize beyond metaphor explanation to other open-ended explanation tasks.

Limitations

Our findings should be interpreted in light of several limitations. The proposed framework is theory-informed but has only been evaluated on a deliberately constructed benchmark of metaphor explanations, rather than naturally occurring explanations produced by humans or contemporary language models. Moreover, while the annotation study provides evidence that the proposed dimensions capture meaningful aspects of human evaluation, it does not establish that they constitute an exhaustive or universal characterization of explanation quality. Finally, the automatic evaluation study is exploratory, using a relatively small dataset to investigate feasibility rather than to establish a state-of-the-art evaluation system.

Ethical Statement

This work studies the evaluation of metaphor explanations and does not involve the deployment of language models in real-world decision-making. Participants provided informed consent and annotated metaphor explanations using anonymous ratings. The benchmark consists of linguistic examples and intentionally constructed explanation failures, none of which contain personal or sensitive information.

References

Brian Bowdle and Dedre Gentner. 2005. The career of metaphor. *Psychological Review*, 112(1):193–216.

Silvia Cappa, Anna Sofia Lippolis, and Stefano Zoia. 2025. Meanings are like onions: A layered approach to metaphor processing. In *Proceedings of the Joint Ontology Workshops (JOWO), Catania, Italy, September 8-12, 2025*, volume 4176 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. Evaluation of text generation: A survey. *Transactions of the Association for Computational Linguistics*, 9:995–1022.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250–4261, Online. Association for Computational Linguistics.

Gilles Fauconnier and Mark Turner. 2002. *The Way We Think*. Basic Books.

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

Sam Glucksberg. 2001. *Understanding Figurative Language: From Metaphor to Idioms*. Oxford University Press.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Vol. 3: Speech Acts*, pages 41–58. Academic Press.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2025. An empirical study of LLM-as-a-judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5880–5895, Vienna, Austria. Association for Computational Linguistics.

Nicholas Ichien, Dušan Stamenković, and Keith J. Holyoak. 2024. Large language model displays emergent ability to interpret novel literary metaphors. *Metaphor and Symbol*, 39(4):296–309.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

George Lakoff and Mark Johnson. 1999. *Philosophy in the Flesh*. Basic Books.

Ronald W. Langacker. 2008. *Cognitive Grammar: A Basic Introduction*. Oxford University Press.

Antonio Lieto, Gian Luca Pozzato, and Stefano Zoia. 2025. The delta of thought: Channeling rivers of commonsense knowledge in the sea of metaphorical interpretations. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 10316–10324. International Joint Conferences on Artificial Intelligence Organization. Human-Centred AI.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Patricia Schmidtova, Saad Mahamood, Simone Balloccu, Ondrej Dusek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondrej Platek, and Adarsa Sivaprasad. 2024. Automatic metrics in natural language generation: A survey of current evaluation practices. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 557–583, Tokyo, Japan. Association for Computational Linguistics.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1002–1010, Beijing, China. Coling 2010 Organizing Committee.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6724–6736, Online. Association for Computational Linguistics.

Jakob Suchan, Mehul Bhatt, and Harshita Jhavar. 2015. Talking about the Moving Image: A Declarative Model for Image Schema Based Embodied Perception Grounding and Language Generation. *CoRR*, abs/1508.03276.

Jakob Suchan, Mehul Bhatt, and Srikrishna Varadarajan. 2021. Commonsense visual sensemaking for autonomous driving - on generalised neurosymbolic online abduction integrating vision and semantics. *Artificial Intelligence*, 299:103522.

Jakob Suchan, Mehul Bhatt, Srikrishna Vardarajan, Seyed Ali Amirshahi, and Stella Yu. 2018a. Semantic Analysis of (Reflectional) Visual Symmetry: A Human-Centred Computational Model for Declarative Explainability. *Advances in Cognitive Systems*, 6:65–84.

Jakob Suchan, Mehul Bhatt, Przemyslaw Andrzej Walega, and Carl P. L. Schultz. 2018b. Visual explanation by high-level abduction: On answer-set programming driven reasoning about moving objects. In *32nd AAAI Conference on Artificial Intelligence (AAAI-18), USA*, pages 1965–1972. AAAI Press.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silvana Vida-Robles, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1365–1402.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Supplementary Material

The appendix provides supplementary analyses supporting the results presented in the main paper. These include (A) additional analyses of annotator disagreement, complete regression diagnostics, (B) further evaluation of the exploratory automatic-evaluation pipeline, (C) aggregate participant characteristics, and (D) reproducibility details.

A Additional Human Analyses

A.1 Annotator Disagreement Across Quality Levels

Figure 2 summarizes annotator disagreement across buckets of mean Overall quality. Disagreement is operationalized as the standard deviation of the Overall ratings assigned to each metaphor. The resulting pattern is approximately U-shaped: annotators show greater agreement when an explanation is judged clearly poor or clearly successful, whereas disagreement is highest for explanations receiving intermediate scores.

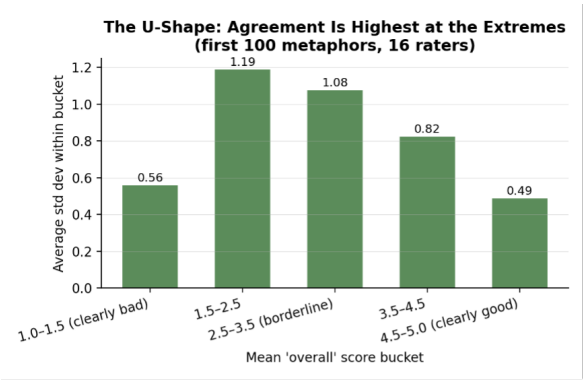


Figure 2: Average annotator disagreement across buckets of mean Overall quality.

A.2 Disagreement Across Evaluation Dimensions

Figure 3 compares average annotator disagreement across the six evaluation dimensions and Overall quality. Purpose exhibits the greatest average variability, followed by Mapping. Origin exhibits the lowest variability, suggesting that annotators converged more strongly when evaluating the source or cultural basis of a metaphor than when evaluating its communicative purpose or conceptual mapping.

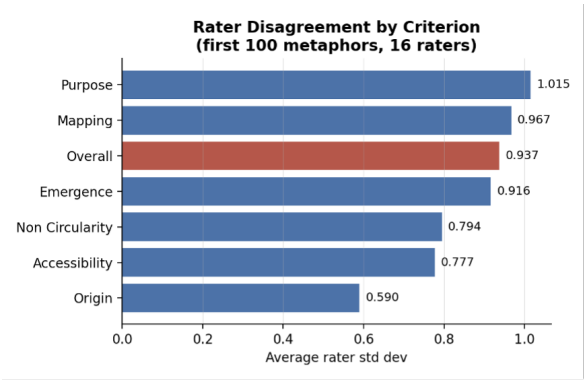


Figure 3: Average annotator disagreement by evaluation dimension. Bars represent the mean within-item standard deviation across metaphors.

A.3 Examples of High and Low Annotator Agreement

Tables 7 and 8 present the five metaphors with the lowest and highest variability in Overall ratings, respectively. High agreement occurs at both extremes of the quality scale: annotators converge on explanations that are clearly successful as well as explanations that are clearly unsuccessful. The highest-disagreement cases generally receive intermediate mean scores, consistent with the aggregate pattern in Figure 2.

A.4 Full Regression Results

Table 9 reports complete simultaneous regression predicting Overall quality from the six standardized evaluation dimensions. Emergence provides the strongest independent predictive signal, followed by Purpose. The remaining coefficients are comparatively small, and Accessibility does not exhibit a significant independent association after controlling for the other dimensions. The small negative coefficient for Origin should be interpreted cautiously. Its sign may reflect shared variance or a suppression effect among correlated predictors

Metaphor	Evaluated explanation	Overall	SD
An apology is the super glue of life.	A sincere expression of regret repairs damaged relationships so effectively that it restores trust and reconnects people who had emotionally separated.	4.88	0.34
The elephant in the room.	The metaphore describing a large trunked mammal that has somehow entered a domestic living space or office building.	1.19	0.40
She is an open book.	This means comparing a female human being to a printed collection of bound pages that is not currently closed.	1.25	0.45
His bruised ego will take time to recover.	Describes the painful blow to someone's self-esteem after a failure, requiring a period of emotional healing to regain confidence.	4.75	0.45
A Herculean task.	According to the Twelve Labors of Hercules in Greek myth, it describes a task requiring superhuman strength or extraordinary effort.	4.75	0.45

Table 7: Metaphors with the highest annotator agreement (lowest Overall-rating standard deviation).

Metaphor	Evaluated explanation	Overall	SD
A dead end.	This refers to a stratified cryospheric accumulation where frozen atmospheric water creates a fibrous textile covering over the terrestrial surface.	3.19	1.80
Hitting a brick wall.	This is a statement describing a person who experiences a sudden physical impact against a solid, rectangular masonry structure.	3.31	1.78
Life is a roller coaster.	This phrase frames human existence as a sequence of events using the logic of an amusement park.	3.88	1.50
Journalism is literature in hurry.	The rapid production of written narratives about current events, created under tight deadlines without the luxury of time.	3.06	1.48
Frozen with fear.	Frozen with fear means feeling fear so intense that a person becomes frozen, unable to move or respond.	2.50	1.46

Table 8: Metaphors with the lowest annotator agreement (highest Overall-rating standard deviation).

Pred.	β	SE	t	p	90% CI
NC	0.062	0.016	3.91	$< .001$	[0.031, 0.094]
OR	-0.070	0.016	-4.42	$< .001$	$[-0.102, -0.039]$
MP	0.061	0.021	2.90	.004	[0.020, 0.103]
EM	0.499	0.023	21.98	$< .001$	[0.454, 0.543]
AC	0.014	0.017	0.81	.420	$[-0.020, 0.047]$
PU	0.290	0.023	12.51	$< .001$	[0.245, 0.335]

Table 9: Full standardized regression output predicting Overall quality from all six evaluation dimensions simultaneously. Coefficients represent associations with Overall quality while holding the remaining dimensions constant.

rather than a substantively negative relationship between origin accuracy and explanation quality.

B Additional Analyses of the Automatic-Evaluation

B.1 Criterion-Level Generalization

Table 10 compares mean absolute error (MAE) on the training and held-out test sets for each evaluation dimension. Most dimensions exhibit higher error on the held-out metaphors, indicating a degree of overfitting consistent with the modest size of the available dataset. The increase is largest for Purpose and Mapping, suggesting that these dimensions are comparatively more difficult to generalize. By contrast, Non-Circularity shows slightly lower error on the test set, although this difference is small and should not be overinterpreted given the limited number of held-out examples.

Dimension	Train MAE	Test MAE	Δ
NC	0.261	0.208	-0.053
OR	0.243	0.263	0.020
MP	0.252	0.354	0.102
EM	0.251	0.343	0.092
AC	0.210	0.260	0.049
PU	0.239	0.350	0.110

Table 10: Training and held-out-test MAE for the six criterion-level predictors. Positive values of Δ indicate higher error on the held-out metaphors.

B.2 Association Between Human Variability and Model Error

The relationship between human disagreement and automatic-evaluation error was positive but uncertain. Pearson correlation produced $r = .397$ ($p = .083$), while Spearman correlation produced $\rho = .203$ ($p = .392$). Thus, although the observed direction is consistent with the possibility that examples that are difficult for humans are also difficult for the automatic evaluator, the evidence does not establish a reliable association.

B.3 High- and Low-Disagreement Groups

As an additional exploratory analysis, the held-out metaphors were divided using a median split of their Overall-rating standard deviation ($\sigma = 0.935$).

Group	n	Mean SD	Mean error	Median error
High disagreement	10	1.151	0.790	0.700
Low disagreement	10	0.636	0.584	0.525

Table 11: Automatic-evaluation error for held-out metaphors with high and low human-rating disagreement.

Although error was descriptively greater among the high-disagreement examples, the difference was not statistically significant under either a Mann–Whitney test ($U = 58.00$, $p = .571$) or Welch's t-test ($t = 1.07$, $p = .303$). The comparison is exploratory given the small number of held-out metaphors.

C Participant Information

Participants (Table 12) represented a range of linguistic, cultural, and educational backgrounds. The study recorded native or first language, self-reported English proficiency, cultural or geographic background, familiarity with linguistics, literature, or metaphor, age or educational band, and recruitment method.

Characteristic	Sample description
Included annotators	16
English proficiency	B2 or fluent
Number of native languages	9
Age range	18–73
Median age	21
Linguistics or literature training	3 participants
Recruitment method	Personal contact

Table 12: Aggregate characteristics of the annotator sample.

D Reproducibility Details

D.1 Annotation Protocol

Participants completed the annotation task through a Google Forms survey. The survey header presented the task objective, the 1–5 scoring rubric, and the six evaluation criteria (Figure 4). The survey was completed independently and asynchronously. Participants did not complete the task in a single supervised session and did not communicate with one another during annotation. Responses were exported from Google Forms as a CSV file for analysis.

D.2 Explanation Generation and Benchmark Construction

The benchmark contains 100 metaphor explanations drawn from a range of physical, social, cultural, and abstract conceptual domains. Source metaphors were curated from publicly available metaphor collections and supplemented with idiomatic expressions compiled by the authors from general usage. Candidate explanations were generated through the web interfaces of three commercial large language models: GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro. A standardized prompt template was used across systems. Explanations were normalized to approximately 25 words and to a broadly uniform, professional tone.
The generation prompt was based on the following template:

> **Role:** You are a computational linguistics expert.
>
> **Task:** Provide a definition for the metaphor: "Frozen with fear."
>
> **Constraint 1:** The definition must be approximately 25 words.
>
> **Constraint 2, targeted or mixed failures:** The definition must intentionally fail the specified subset of criteria while passing all others.
>
> **Constraint 2, reference explanations:** The definition must pass all six criteria.

The following criteria definitions were provided:

1. **Non-Circularity:** The definition must not restate the metaphor using its own words, or explain it via another metaphor.

2. **Origin and Cultural Accuracy:** The definition must identify the specific cultural, historical, or domain-specific origin of the metaphor (e.g., Chess, Myth, Navy), where applicable. This criterion applies only when the metaphor's source or target domain is traceable to a specific cultural tradition, historical event, game, field, or institution; a fail means the definition describes the meaning but ignores a traceable source.

Evaluation of Metaphor Definitions: Participant Guide

Objective: You are helping evaluate how well AI-generated definitions explain common metaphors. You will see a **Metaphor** and its **Definition**. Please rate the definition on 6 criteria using a scale of **1 (Poor/Fails)** to **5 (Excellent/Perfect)**.

The Scoring Rubric (1-5)

- **5:** Perfect. Meets all requirements of the criteria.
- **4:** Good. Very minor issues, but clear and accurate.
- **3:** Acceptable. Satisfactory, but lacks depth or has small errors.
- **2:** Poor. Major issues with accuracy, clarity, or logic.
- **1:** Fails. The definition is completely wrong or violates the rule entirely.

The 6 Criteria Explained

1. **Non-Circularity:** The definition must explain the concept without using the words from the metaphor itself. It also should not explain a metaphor using a *different* metaphor (e.g., don't define "Love is a battlefield" as "a romantic war").
2. **Origin/Cultural Key:** If a metaphor is tied to history, myth, or a specific game (like Chess), the definition **must** mention that origin.
    - *Note: If the metaphor is universal/common and needs no origin, give this a **5**.*
3. **Mapping (Logic Bridge):** Does it explain the *specific trait* shared by both things? (e.g., for "Heart of stone," it must mention "unfeeling" or "hard", not just that the heart is a mineral).
4. **Emergence:** Does it capture the deeper, figurative "spark"? It should not be a purely literal or robotic dictionary entry.
5. **Accessibility:** Is it written in clear, simple English? Avoid overly complex jargon or confusing sentence structures.
6. **Purpose:** Does it explain *why* or in *what situation* someone would use this phrase?

Figure 4: Participant instructions reproduced as displayed in the Google Forms survey.

3. **Emergence:** The definition must capture the figurative meaning that emerges from the metaphor, not merely describe the literal, physical objects involved.
4. **Accessibility:** The definition must use clear, natural language, avoiding dense academic jargon.
5. **Purpose:** The definition must explain the emotional or situational intent behind using the phrase (why or when someone would use it).
6. **Mapping:** The definition must identify the specific trait shared by the source and target domains (e.g., "vastness," "heat," "sharpness"), rather than defining the target without the linking trait.

Generated explanations were manually checked against the intended quality profile. Outputs that were hallucinated, off target, or inconsistent with the requested failure pattern were filtered and regenerated until the intended benchmark distribution was obtained.

The final benchmark contains:

- **15 reference explanations** designed to satisfy all six criteria;
- **60 targeted failures**, with ten explanations designed to violate each individual criterion; and
- **25 mixed failures** designed to violate multiple criteria simultaneously.

D.3 Automatic Evaluation Pipeline

The automatic evaluator consists of two stages. The first stage predicts the six criterion scores and the overall human judgment from the metaphor–explanation pair. The second stage maps the six predicted criterion scores to a final overall quality estimate.

Stage 1: Criterion Prediction

We fine-tuned `bert-base-uncased` as a multi-output regression model. The metaphor and candidate explanation were concatenated into a single input sequence, allowing the encoder to model the relation between the expression and its proposed

Hyperparameter	Value
Estimator	`GradientBoostingRegressor`
Number of estimators	200
Learning rate	scikit-learn default
Maximum tree depth	scikit-learn default
Subsample	scikit-learn default
Loss	scikit-learn default
Random state	42

Table 13: Stage-2 Gradient Boosting configuration.

Hyperparameter	Value
Base checkpoint	`bert-base-uncased`
Task head	`BertForSequenceClassification`
Number of labels	7
Problem type	Regression
Loss function	Mean squared error
Optimizer	AdamW
Learning rate	2×10^{-5}
Training batch size	2
Evaluation batch size	Hugging Face default
Weight decay	0.01
Epochs	50, fixed
Early stopping	None
Dropout	0.1, BERT default
Max. sequence length	128 tokens
Data-split seed	42
Training/init. seed	42
Hardware	Apple M2
Backend	PyTorch MPS, with CPU fallback
Approx. training time	3 hours

Table 14: Hyperparameters used to fine-tune the Stage-1 BERT regressor.

explanation. A linear regression head mapped the encoder representation to seven continuous targets: the six criterion scores and the overall quality score. BERT was selected as a computationally lightweight alternative to an LLM-based judge and to reduce the risk of self-preference effects that may arise when a generative model evaluates outputs produced by similar systems.

Stage 2: Overall-Score Prediction

The six criterion predictions produced by Stage 1 were used as input features to a Gradient Boosting regressor implemented in scikit-learn. The model predicts the final overall quality score. Gradient Boosting was used through five-fold `GroupKFold` cross-validation on the training portion, with grouping performed by metaphor to avoid leakage between explanations associated with the same metaphor. The model used 200 boosting estimators. Hyperparameters are reported in Table 13.

D.4 Training Configuration

Table 14 reports the Stage-1 training settings. Hyperparameters were set manually using standard defaults for BERT fine-tuning on a small regression dataset. The model was trained and tested using 1,600 annotation instances derived from 16 annotators and 100 metaphor explanations. Train–test splitting was performed at the metaphor level to prevent leakage. The held-out evaluation set contained 20 metaphors.

D.5 Reproducibility Checklist

- **Dataset size:** 100 metaphor explanations.
- **Annotators:** 16.
- **Annotation targets:** six criterion scores plus overall quality.
- **Split strategy:** metaphor-level train–test split.
- **Held-out set:** 20 metaphors.
- **Stage-1 model:** `bert-base-uncased` with a seven-output regression head.
- **Stage-2 model:** scikit-learn Gradient Boosting regressor.
- **Primary software:** PyTorch, Hugging Face Transformers, and scikit-learn.
- **Hardware:** Apple M2 using the PyTorch MPS backend, with CPU fallback where required.
- **Reported runs:** one train–test split and one training run, not an average across random seeds.