Decomposing Creativity: Two Small Datasets Combining Originality Ratings and Metaphor Annotations

Emilie Sitter, Sina Zarrieß, Omar Momen, Berenike Herrmann
CRC 1646 – Linguistic Creativity in Communication
Faculty of Linguistics and Literary Studies
Bielefeld University, Germany
{emilie.sitter,sina.zarriess,omar.hassan,berenike.herrmann}@uni-bielefeld.de

Abstract

We introduce MetaphOrig, two small datasets comprising two genre-specific collections of spatial descriptions for the study of linguistic creativity and Non-Literal Expressions (NLEs). The sentence-level spatial descriptions were extracted from two distinct genre- and time-specific source corpora. Both source corpora comprise German texts: literary prose from the 18th to the 20th century (KOLIMO) and factual travel reports from the 21st century (Wikivoyage). Along with the spatial descriptions, the datasets contain sentence-level originality ratings obtained through crowdsourcing and from four different LLMs (GPT-5, Qwen2.5-32B-Instruct, Mistral-Small-3.2-24B-Instruct, and Llama-3.2-3B), and word-level metaphor annotations. We provide the MetaphOrig datasets, including all annotations, to the community. The datasets can be used for further research on linguistic creativity or metaphor, either in one specific textual domain or comparatively across the two domains. We conduct an illustrative study on the datasets, treating originality as a proxy of textual creativity. In both datasets, we investigate potential correlations between sentence-level originality ratings and the density of metaphorical expressions within each sentence. We find the correlation to be present only in the KOLIMO dataset. A comparison of human and LLM originality ratings shows that this pattern holds for both types of ratings.

Keywords: Linguistic Creativity, Annotation, Crowdsourcing, Metaphor, Originality, MIPVU, Large Language Models

1. Introduction

This paper introduces MetaphOrig, two small datasets that aim to bridge the gap between holistic, crowdsourced originality ratings at the text-level and a fine-grained, linguistic analysis of individual rhetorical devices. The exploratory study emphasizes metaphors, which are among the most prominent Non-Literal Expressions (NLEs) as a potential contributor to linguistic creativity.

According to the standard definition of creativity, the phenomenon comprises the two dimensions of originality and effectiveness (Runco and Jaeger, 2012). In this paper, we particularly focus on originality as its presumably most important dimension (Diedrich et al., 2015). We understand originality as the perceived novelty, unconventionality, or unexpectedness of a text for a specific reader.

Studies in psychology or psycholinguistics often aim to assess the creativity of ideas, artistic works, or linguistic productions. To this end, human assessments of creativity are oftentimes collected through crowdsourcing or by ratings of domain experts (Qian and Plucker, 2017).

In contrast, linguistics and NLP rarely conceptualize creativity as a holistic property of an idea or utterance. Instead, they tend to focus on specific textual features and formal characteristics that may influence how creative an utterance is perceived to be by humans (Weinstein et al., 2022; Zedelius et al., 2019). Metaphors provide a particularly salient example of such a feature, as they are often assumed to enhance the perceived creativity of a sentence. They are one of the most extensively researched forms of potentially creative and original language use (Kohl et al., 2020), and they are widespread or even ubiquitous in language and thought (Lakoff and Johnson, 2003). Yet, in NLP research, there is relatively little work and hardly any existing datasets that combine annotations of metaphors, on the one hand, and of creativity or originality, on the other hand.

According to Lakoff and Johnson's Conceptual Metaphor Theory (CMT), metaphor is a cognitive mechanism that structures human understanding by mapping complex experiences onto more concrete or familiar domains. Metaphorical expressions vary widely across genres in form, function, and cognitive effects, as well as in their degree of creativity or originality (Steen et al., 2010a; Herrmann, 2015; Momen et al., 2026). Even in presumably creative domains such as advertisement or literary texts, many metaphorical expressions are conventional, but still contribute to perceived creativity (Steen et al., 2010a; Dorst, 2015; Burgers et al., 2015). In the context of NLP, metaphors thus continue to pose challenges for current approaches to annotation and detection because of their context dependence and varying degrees of conventionality (Maudslay and Teufel, 2022; Ye et al., 2025).

The MetaphOrig datasets consist of two small, genre-specific datasets of sentence-level descriptions of spatial scenes, annotated with originality ratings and metaphor labels. Spatial descriptions

Proceedings of Learning Non-Literal Expressions with Small Data @ LREC 2026, pages 119–132
11 May 2026.

Rating Item	Example 1 (Literary)	Example 2 (Non-literary)
Original Text	*Die Glastüren zur Veranda standen offen, und der Duft des Flieders drang herein, der wie eine Mauer aus weißem und hellblauem Gewölk den Garten einhegte.*	*São Roque: Diese Stadt liegt an der Nordküste von Pico und ist bekannt für ihre schönen Naturschwimmbecken, in denen Besucher baden können.*
Translation (own)	*The glass doors to the terrace were open, and the scent of the lilac drifted in, enclosing the garden like a wall of white and light blue clouds.*	*São Roque: This town is located on the north coast of Pico and is famous for its beautiful natural pools where visitors can swim.*
Source	Eduard von Keyserling, *Abendliche Häuser* (1940)	Wikivoyage, *Pico* (2023)
Metaphors	Metaphor Density: 0.23913 Direct Metaphors: 4 ("Mauer", "weißen", "hellblauen", "Gewölk") Indirect Metaphors: 1 ("standen offen")	Metaphor Density: 0.047619 Direct Metaphors: 0 Indirect Metaphors: 1 ("liegt")
Originality Ratings	Humans: 4.8 GPT-5: 5.0 Qwen2.5-32B-Instruct: 4.6 Mistral-Small-3.2-24B-Instruct: 5.0 Llama-3.2-3B: 4.6	Humans: 2.1 GPT-5: 1.0 Qwen2.5-32B-Instruct: 2.0 Mistral-Small-3.2-24B-Instruct: 3.0 Llama-3.2-3B: 3.3

Table 1: Example sentences from the MetaphOrig datasets with respect to metaphors and averaged originality ratings by humans and LLMs.

commonly appear as textual elements in different types of genres, including travel reports or novels. Such descriptions are comparable content-wise and can be understood in isolation. We therefore assume that human raters can assess the linguistic originality of these controlled sentences more easily than that of randomly selected passages which would differ substantially in terms of content.

The spatial descriptions were extracted from two fundamentally different textual domains, yielding two genre-specific datasets: 18th to 20th century literary prose on the one hand and factual travel reports on the other (see the example sentence in Table 1 from a literary prose text alongside a non-literary sentence from a travel report). The older literary prose presumably seems highly literary to contemporary readers, while the travel reports are characterized by factual and, presumably, more non-figurative language.

For each sentence in both datasets, we additionally collected originality ratings via online crowdsourcing and annotated word-level metaphors following the MIPVU procedure (Steen et al., 2010b; Herrmann et al., 2019). These annotations result in two original datasets of sentences matched with both sentence-level originality scores and word-level metaphor information.

We showcase a potential use-case of the MetaphOrig datasets by presenting an analysis that correlates metaphor density with originality according to the human judgments as well as to LLM ratings of originality. In a first step, the two very different datasets are used to analyze the two formally distinct textual domains in terms of originality and metaphoricity. We examine whether the relationship between human originality ratings and metaphor use varies across them. In a second step, we explore whether LLMs can be leveraged for rating originality similarly to a crowdsourced collection of ratings. In addition to the human originality ratings, we thus collect ratings from four different LLMs. We aim to explore not only the extent to which LLMs approximate human originality judgments, but also whether their originality ratings appear to be grounded in similar textual cues.

2. Background

2.1. Creativity Assessment

In creativity research, the goal of creativity assessments is to determine the creativity of particular products, as opposed to, e.g., the creativity of individuals (such as writers) or processes (their writing processes). The creativity of products, such as texts, is typically assessed through human ratings. Most of these human-evaluation approaches include Likert scale ratings of novelty or originality collected from experts or via crowdsourcing (Hennessey et al., 2011; Kaufman et al., 2009), structured pairwise comparisons where raters select the more creative item from a pair (Crompvoets, 2025; Cao et al., 2026), and ranking tasks in which multiple items are ordered according to perceived creativity or originality (Do Dinh et al., 2018). One of the most influential approaches in this area is the Consensual Assessment Technique (CAT) introduced by Amabile (1982). While CAT raters are usually domain experts, a certain degree of subjectivity in the ratings is inevitable (Qian and Plucker,

2017). Mozaffari (2013) proposed an analytical rubric that assesses multiple sub-dimensions of creativity rather than relying on one single, holistic creativity score. All these approaches share the common assumption that creativity can be inferred from the collective judgments of human raters.

2.2. Metaphors and Creativity

The most influential approach to metaphor in linguistics is the Conceptual Metaphor Theory (CMT), introduced by Lakoff and Johnson in their seminal work *Metaphors We Live By* (1980) (Lakoff and Johnson, 2003). In linguistics, metaphor is still predominantly understood according to CMT as a cognitive mechanism by which a (typically abstract) *target domain* is conceptualized in terms of a more concrete *source domain*. While conceptual metaphors, the underlying cognitive mechanisms of a metaphor, are typically highly conventional, their linguistic realizations can vary widely and may be realized as novel and creative utterances (Kövecses, 2020). This can be observed especially in literary text which allows authors and poets to create novel linguistic metaphors by creatively re-interpreting conventional and everyday conceptual metaphors (Semino and Steen, 2008; Kövecses, 2010).

The Metaphor Identification Procedure (MIP) (Pragglejaz Group, 2007) and its extension by VU Amsterdam (MIPVU) (Steen et al., 2010b) are currently among the most common ways to annotate metaphor in linguistics. However, these annotation procedures do not capture the degree of creativity, novelty, or deliberateness of a metaphor. Existing approaches to annotate metaphor novelty as one dimension of creativity are based either on crowdsourcing (Do Dinh et al., 2018) or they are dictionary-based (Egg and Kordoni, 2022). According to Reimann and Scheffler (2024), dictionary-based annotations align more reliably with expert metaphor annotations.

Further, more detailed annotation protocols of metaphors are based on Deliberate Metaphor Theory (DMT) (Steen, 2017) and aim to identify whether a metaphor is used potentially deliberately (Reijnierse et al., 2018; Dipper et al., 2024). DMT suggests that non-deliberate metaphors are always conventional.

In the context of working with potentially creative or deliberate metaphors, the distinction between direct and indirect metaphors introduced by MIPVU is also worth considering (Steen et al., 2010a). Direct metaphors, often comprising longer phrases, are typically realized in the form of similes (comparisons) and explicitly marked in the discourse, e.g., by 'as' or 'like' (Steen et al., 2010a) (see example (1) in Table 1). They are particularly important when studying creativity because they might be strongly related to originality and deliberate usage. Indirect metaphors, by contrast, largely occur on the word and phrase level and are far more frequent, conventionalized, and more often used non-deliberately (Steen, 2017). They are less likely to contribute strongly to a sentence's overall originality.

When creating the MetaphOrig datasets, we opted for a parallel annotation setting: we collected standard creativity ratings on the sentence-level, and independent metaphor annotations on the word level according to MIPVU.

2.3. LLMs in creativity and metaphor research

To date, a variety of computational approaches have been developed to assess creativity. Acar (2025) provides an overview of the evolution of computational methods for assessing creativity.

Research on using LLMs to predict creativity ratings has so far come mainly from the social sciences and from psychology. Many of these experiments rely on human creativity ratings of either texts of different domains or of responses to creativity tests, such as the Alternative Uses Task (AUT). Luchini et al. (2025) demonstrated that LLMs outperform measures of semantic distance in terms of predicting human originality ratings of creativity test responses. Organisciak et al. (2023) found similar results for AUT responses and Laverghetta et al. (2025) for solutions to design problems. Rabeyah et al. (2024) point out the high correlation of multiple LLMs with each other in creativity scoring of AUT responses. More specifically, in the humanities, in a study with a particular focus on creative writing, Kim and Oh (2025) could demonstrate high consistency and performance of LLM ratings. Not specifically related to originality, but in terms of assessing the quality of narrative text, Chiang and Lee (2023) demonstrated that LLMs are a potential alternative to human ratings.

Specific work on annotating metaphors or their originality or novelty in the context of NLP and Digital Humanities has been carried out by several researchers. Some studies aimed to apply LLMs as annotators of word-level metaphors (Hicke and Kristensen-McLachlan, 2024; Reimann and Scheffler, 2025; Sánchez-Montero et al., 2025). They demonstrate that under certain conditions, LLM-based annotations can approximate human ones, but also highlight the brittleness of prompting approaches and the sensitivity to specific prompts. DiStefano et al. (2024) fine-tuned two language models on human creativity ratings of metaphors, suggesting that these models may be able to approximate certain aspects of how humans interpret figurative language. Momen et al. (2026) explored LLM surprisal as a predictor of metaphor novelty. They found a positive correlation between word-

level surprisal of metaphors and their novelty ratings in multiple datasets. In the present paper, we focus on comparing LLM originality ratings to human ratings, leaving the modeling of (creative) metaphor detection for future work.

3. Introducing the MetaphOrig datasets

This section describes the construction of the two datasets of MetaphOrig. We provide the datasets, including all annotations on GitHub.

3.1. Spatial descriptions datasets

Both MetaphOrig datasets consist of 100 spatial descriptions extracted from German texts. They were selected randomly from two larger datasets of sentences that have been annotated manually based on annotation guidelines for identifying descriptions of spatial scenes (Sitter et al., 2025). Such spatial descriptions aim to describe concrete, static spatial, scenic surroundings without any immediate actions. We focus on spatial descriptions for several methodological reasons: First, spatial descriptions are common and integral textual elements across different genres and can be found in both source corpora of the MetaphOrig datasets. Second, their comparability in terms of content facilitates the assessment of originality at the linguistic level, as it reduces the risk that raters conflate linguistic creativity with originality of the content. Third, the selected spatial descriptions are comprehensible in isolation and can be presented to raters without additional textual context. This restriction allows for stronger experimental control and ensures a substantial degree of comparability between individual items, although it may limit the generalizability of the findings to entire texts.

3.2. The KOLIMO dataset

The first small dataset consists of 100 spatial descriptions that originally were extracted from KOLIMO, the "Corpus of Literary Modernity", (Horstmann, 2019; Horstmann and Akazawa, 2024; Herrmann and Lauer, 2018). This corpus comprises literary fictional prose texts mainly from the 18th to 20th centuries, while most of the sentences have been extracted from texts published in the 19th century. We chose a literary source corpus for building the dataset because metaphors are a particularly important rhetorical device in literary writing and we expected a high amount of non-literal expressions in these texts. Moreover, to the best of our knowledge, there is currently no prior study that systematically collected originality ratings for individual literary sentences using crowdsourcing methods. Using 18th–20th century language complicates comparisons with contemporary texts. However, it is difficult to obtain and distribute annotated data for contemporary literary texts due to copyright reasons. This limitation particularly affects highbrow canonical works, which are often regarded as the most creative representatives of their genre. Consequently, we opted for texts that are both publicly accessible and literary prestigious in character instead of contemporary text material.

3.3. The Wikivoyage dataset

The second small dataset consists of 100 spatial descriptions extracted from the Wikivoyage corpus (Nolda, 2024; Wikimedia Foundation Inc., 2025). Wikivoyage is a collection of non-literary travel reports (21st century). Most sentences are factual and plain (see example (2) in Table 1). This dataset can thus be expected to contain less figurative and less original language use, compared to the KOLIMO dataset.

4. Ratings and annotations

For each of the textscMetaphOrig datasets, we collected originality ratings via crowdsourcing and from four different LLMs. The metaphors in the sentences were annotated by trained experts following the MIPVU procedure (Steen et al., 2010b). This section describes the obtained ratings and the annotations.

4.1. Human originality ratings

We collected the ratings for all sentences in both small datasets as part of a large-scale rating study on linguistic creativity on Prolific. Prolific is a crowdsourcing platform that is primarily used in the social sciences and aims to ensure the high quality of the data collected through quality checks. Remuneration for the raters was calculated based on the German minimum wage at the time of data collection (May 2025).

A total of 120 L1 German speakers rated all items on a six-point Likert scale for originality. 10 people in total rated each individual item in an incomplete between-groups design that was applied to reduce the cognitive load on the participants.

Participants were presented with 50 items each (one third each from the KOLIMO and the Wikivoyage dataset, one third filler items). We provided no explicit genre information and stated that all presented items were taken from everyday language texts. This was intended to ensure a uniform extra-textual context for both datasets and thus maximize the comparability of the ratings. At the same time,

our aim was to examine the extent to which textual genre cues alone affect originality judgments, without explicitly priming raters toward a literary reading mode or evoking genre expectations that might influence their ratings (Knoop and Blohm, 2025). However, given that humans are often able to infer genre solely by reading a text without relying on any explicit extra-textual genre cues (Knoop et al., 2024), the KOLIMO example (1) in Table 1 is still likely to be considered as literary but not the Wikivoyage example (2). Some participants explicitly commented in the feedback field at the end of the study that they identified the KOLIMO sentences as passages from literary text despite being presented as everyday language.

In addition to a German instruction text (Appendix A), participants received examples of one very original and one very conventional sentence. Rating each individual sentence was not mandatory. The mean number of raters per sentence is thus 9.97 (range 9–10, SD = 0.171). The inter-rater agreement is moderate (Krippendorff's Alpha = 0.525), likely reflecting the inherently subjective nature of perceived originality.

Importantly, we do not aim to model how texts of the KOLIMO dataset were perceived at the time of their publication. We are rather interested in modeling the "synchronous" originality judgments of today's readers which are made from a contemporary perspective. The temporal distance to the texts' publication presumably contributes to a higher degree of literariness as perceived by today's readers. From a contemporary perspective, literariness may not always be clearly separable from originality. We assume that the perceived literariness of these historically distant texts may contribute to higher originality ratings among today's readers. By contrast, readers at the time of publication may not necessarily have perceived texts with similar stylistic features as equally marked or original.

4.2. LLM originality ratings

We obtained LLM annotations by three instruction-tuned locally hosted LLMs, covering a spectrum of different sizes, and one commercial LLM, GPT-5. The largest locally hosted model, Qwen2.5-32B-Instruct (Yang et al., 2024), has previously proven particularly well suited for the automatic extraction of spatial descriptions (Sitter et al., 2025). This suggests that it might be well adapted to this kind of data. Mistral-Small-3.2-24B-Instruct-2506 (Mistral AI, 2025) represented another family of LLMs and a medium-sized model. The smallest model, Llama-3.2-3B (et al., 2024), was chosen to investigate how well a relatively small LLM works for this task.

All LLM prompts were standardized as much as possible to emulate the human annotation process. The models received basic information about the rating study in the *system prompt*. The *user prompt* contained the same instructions the human raters received. Following a few-shot prompting approach, the user prompt contained examples for one highly original and one highly conventional sentence. In addition to the instructions for human raters, we specified the desired output format in the LLM instructions (see appendix B).

To approximate the Prolific study setup with 10 human raters per sentence, we prompted each model 10 times per sentence. Temperature was set to 1.0 to introduce variability and approximate the diversity of human raters. A final score per sentence for each model was obtained by averaging all 10 ratings.

4.3. Metaphor annotations

We annotated metaphors at the word level applying the Metaphor Identification Procedure of VU Amsterdam (MIPVU) (Steen et al., 2010b) and particularly its specification for German (Herrmann et al., 2019). Applying MIPVU, annotators assessed for each individual word in a sentence whether its contextual sense shows a meaning that is distinct from its more basic, typically concrete sense, while understood by comparison to it (such as "*enclosing*" in example (1), referring to a scent instead of a physical enclosure). More basic senses were looked up in Duden.

All metaphor information was annotated collaboratively by three different annotators trained on the MIPVU procedure (one of whom was the first author of this paper). The annotators cross-checked each other's annotations; cases of disagreement were discussed in detail.

MIPVU allows to annotate both direct and indirect metaphors regardless of their degree of conventionality or novelty. Dictionary-based approaches to annotating metaphor novelty would not be able to capture all direct metaphors and possibly fail to do justice to the creative potential of literary texts. The main annotation categories are "Non-MRW" for words used in their literal sense, "indirect", and "direct", as well as "WIDLII" ("When in doubt, leave it in.") for borderline cases. Markers of direct metaphors are annotated as "Mflag". Following MIPVU, each word of a direct metaphor has to be annotated as metaphorical.

MIPVU annotations can be used to calculate a "metaphor density" for each sentence. Metaphor density is the ratio of metaphorically to non-metaphorically used words in a sentence. To prevent the metaphor density score from being overly skewed toward direct metaphors, we consider only content words and annotated non-content words within direct metaphors as "stopwords". Stopwords within direct metaphors can themselves be indirect metaphors and therefore annotated as "indirect":

	KOLIMO	Wikivoyage
Mean sentence length	23.38	16.15
Median sentence length	21.00	16.00
Metaphor words overall	16.60%	8.57%
Direct metaphor words	2.46%	0.00%
Indirect metaphor words	14.14%	8.57%
Mean metaphor density per sentence	0.1374	0.0870
Sentences with at least one metaphorical word	86%	69%

Table 2: Descriptive statistics for datasets of MetaphOrig

(1) In der Moderluft schlotternd, sahen sie Fröschlach als eine Stadt aus grünen Kachelöfen [...].
Shivering in the cold air, they saw Fröschlach as a city of green tiled stoves [...]. (own translation)

In Example (1), the word *als* 'as' is annotated as "Mflag", marking the onset of a direct metaphor. The phrase *eine Stadt aus grünen Kachelöfen* 'a city of green tiled stoves' constitutes the direct metaphor as a whole, whereas the article *eine* 'a' and the preposition *aus* 'of' can be treated as stopwords. Strictly applying MIPVU, the preposition *aus* itself is a (highly conventional) metaphor and thus annotated as "indirect".

When calculating the metaphor density, we do not differentiate between direct and indirect metaphors. Ambiguous metaphors ("WIDLII") are counted as half. Multi-word expressions are treated as single words. The metaphor density thus represents a comparable score of metaphoricity for all sentences across both entire datasets (Steen et al., 2010a).

5. Analysis

To demonstrate a potential use of MetaphOrig, we analyze the relationship between human and LLMs' originality ratings and examine the role of metaphors in shaping these ratings. All statistical analyses were conducted in R. Relationships between metaphor density, human, and LLM originality ratings were modeled using Cumulative Link Models (CLMs), which allow the analysis of ordinal data. We used the `ordinal` package for ordinal regression models (Christensen, 2023). To facilitate comparison across models, we also compute Spearman's rank correlations. To ensure comparability across corpora and raters, metaphor density and all ratings were rescaled to the same range using the `rescale()` function from the `scales` package in R (Wickham et al., 2011).

5.1. Results

Descriptive Statistics The source and genre differences between the datasets of MetaphOrig yield substantial differences in linguistic form. Even very basic descriptive statistics reflect that the spatial descriptions extracted from KOLIMO are highly literary, i.e., descriptions from KOLIMO exhibit a higher average sentence length and a higher degree of metaphoricity (Table 2) in comparison to Wikivoyage descriptions. Yet, it is worth noting that even in Wikivoyage, 69% of sentences feature at least one metaphor.

Previous comparative research has shown differences in metaphor type use between genres. Even though literary texts do not necessarily exhibit the highest overall metaphor density, they do have the highest frequency of direct metaphors (Steen et al., 2010a; Dorst, 2015; Herrmann, 2015). Such a difference in the use of direct metaphors is also evident in our datasets (Table 2). Since most direct metaphors are deliberate, literary spatial descriptions such as those from KOLIMO can also be expected to contain considerably more deliberate metaphors.

Figure 1 reports the mean, median, range, and standard deviation of the averaged human and LLM originality ratings for each small dataset. Mean and median ratings in each rating setting are higher for the KOLIMO dataset than for the Wikivoyage dataset.

The boxplots also show bigger ranges of originality ratings (by both humans and LLMs) for the KOLIMO dataset than for the Wikivoyage one. Moreover, they reveal that only humans and GPT-5 assign ratings of 1 (not original at all), while the other three LLMs assign ratings of at least 1.6. LLMs tend to give higher ratings than humans, except GPT-5 in the Wikivoyage dataset.

Correlating human originality ratings and metaphor density To analyse the effect of word-level metaphors on human originality ratings of full sentences, we modeled humans' and LLMs' originality ratings with CLMs as an ordinal outcome predicted by the metaphor density, the dataset, and these predictors' interaction. Table 3 reports the effect of metaphor density on the originality ratings in both datasets. Figure 2 displays the human originality ratings by metaphor density, along with the ordinal regression lines computed by the CLM. Similar plots for all rater LLMs can be found in Appendix C. Only in the KOLIMO dataset, metaphor density has a significant positive effect on the human originality ratings. The Wikivoyage dataset shows no effect of metaphor density on originality ratings. While the human raters consistently rate KOLIMO sentences as more original than Wikivoyage sentences even when not metaphorical at all, Figure 2 shows that as

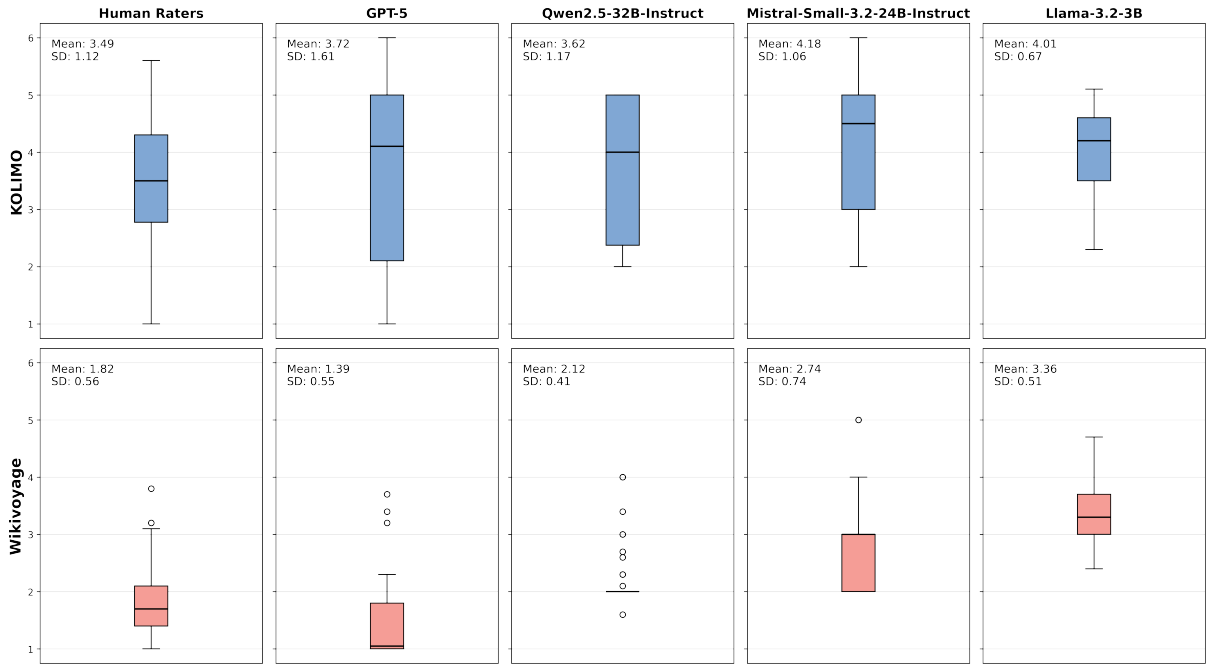


Figure 1: Summary statistics for originality ratings (scale 1–6) for humans and LLMs across the averaged ratings.

Rater	Data	β-coef	SE
Human Ratings	KOLIMO	5.57***	1.00
	Wikivoyage	0.79	0.94
GPT-5	KOLIMO	4.60***	0.89
	Wikivoyage	1.10	0.96
Qwen2.5-32B	KOLIMO	4.21***	0.99
	Wikivoyage	0.83	1.54
Mistral-Small-3.2-24B	KOLIMO	3.94***	0.99
	Wikivoyage	-0.51	0.96
Llama-3.2-3B	KOLIMO	2.06**	0.78
	Wikivoyage	-0.26	0.93

Table 3: Effect of word-level **metaphor density** on originality ratings (from humans and instruct-tuned LLMs). Higher β-coefficients indicate that higher metaphor density predicts higher originality ratings. Stars denote significance (*$p < .05$, **$p < .01$, ***$p < .001$). Standard Error (SE) quantifies the uncertainty of the estimated coefficient.

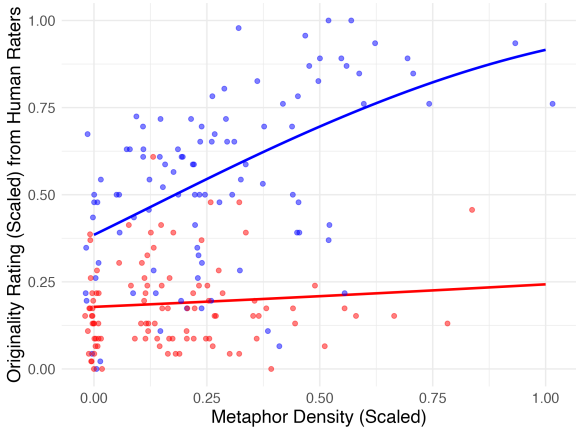


Figure 2: Averaged human ratings per sentence and ordinal regression lines for each corpus (blue = KOLIMO, red = Wikivoyage)

metaphor density increases, the gap between the KOLIMO and Wikivoyage dataset regression lines widens. This illustrates that (i) literary language is generally perceived as more original by humans in our data, and that (ii) within the literary genre, there is a strong impact of metaphors on originality for human readers.

Correlating human and LLM originality ratings To assess the alignment of human and LLM originality ratings, we fitted one CLM for each rater LLM with the human rating as ordinal dependent variable and the LLM rating, the source corpus (KOLIMO vs. Wikivoyage), and their interaction as predictors. These complex models provided significantly better fits than simpler alternatives with less predictors. We also computed Spearman rank-order correlations between each LLM's ratings and human ratings. Results of the CLMs and Spearman correlations are reported in Table 4. The beta coefficients of the CLMs indicate significant alignment between humans and LLMs for both datasets. Rank-order correlations between human and LLM ratings are stronger for the KOLIMO dataset across all four models. GPT-5 shows the highest correlation with human ratings in the KOLIMO dataset, followed by

Qwen2.5-32B-Instruct, whereas Mistral-Small-3.2-24B-Instruct shows the highest correlation in the Wikivoyage one.

Correlating LLM originality ratings and metaphor density The analysis shows that metaphor density has a generally positive effect on the ratings; i.e. the higher the metaphor density, the higher the originality rating. Just as for the human raters, this effect is significant only in the KOLIMO dataset (highly significant for GPT-5, Qwen2.5-32B-Instruct, and Mistral-Small-3.2-24B-Instruct; significant for Llama-3.2-3B) but not in the Wikivoyage dataset. Similar to humans, LLMs rate sentences in the KOLIMO dataset as more original, regardless of their metaphor density—however, the higher the metaphor density, the stronger its impact on originality ratings (see Figure C).

5.2. Discussion

Using the MetaphOrig datasets, the showcase study investigated correlations between originality ratings and metaphor density, as well as between human ratings and LLM ratings.

We find that in general, sentences in the literary KOLIMO dataset are perceived as considerably more original (Figure 1). The statistical analysis also gives reason to assume that originality correlates with metaphor density in the literary KOLIMO dataset. This effect does not hold in the Wikivoyage dataset (Table 3). Generally, these results indicate interesting and potentially complex relationships between creativity, NLEs, and genre, calling for more research on this understudied topic.

Even though the MIPVU annotations do not provide information on the deliberateness or novelty of metaphors, the absence of direct metaphors in the Wikivoyage dataset and the presence of similes in the KOLIMO dataset (Table 2) may partially explain the observed differences. Because of their textual literariness (comparably more frequent and more patterned rhetorical devices), we assume that texts in the KOLIMO source corpus generally contain more deliberate metaphors. A promising direction for future research would be to explicitly annotate and systematically analyze this type of metaphor in the MetaphOrig datasets.

Correlating human and LLM ratings of sentence originality demonstrates that LLMs' assessments significantly align with how humans rate originality. This alignment is much stronger in the KOLIMO dataset than in the Wikivoyage travel report dataset. Just as for human raters, higher LLM originality ratings in the KOLIMO dataset appear to be associated with greater metaphor density. This aligns with our aim of investigating whether LLMs rely on textual cues similar to those underlying human originality judgments, and suggests that metaphor density may function as one such shared cue. Among the LLMs tested, GPT-5 shows the highest Spearman correlation with human ratings for the KOLIMO dataset. For the non-literary Wikivoyage dataset, Mistral-Small-3.2-24B-Instruct-2506 performs best, however, with GPT-5 yielding almost similar results. Overall, this suggests that for this specific rating task of sentence originality, GPT-5 is the most suitable model (Table 4). However, since GPT-5 can be accessed only via OpenAI's API, its usage incurs a certain monetary cost ($4.48 for the batch API in total for both datasets).

One possible further usage of the introduced small datasets could be a more explicit focus on the application of LLMs for metaphor identification in different textual domains. Using an LLM to generate the metaphor annotations that we obtained manually would enable a direct comparison between human and model-based annotations.

Importantly, in this illustrative study, we do not aim to explain any variation in both human and LLM originality ratings by metaphors alone. Originality is likely influenced by a range of interacting stylistic and rhetorical features, such as sentence length, grammatical patterns, or repetition schemas that were not controlled in this study. While the results indicate that metaphors are a particularly salient feature for perceived originality in literary texts, other linguistic features may play a more prominent role in non-literary domains. Future work with the introduced small datasets should therefore consider metaphor as only one feature of a broader set of rhetorical and stylistic devices that can occur in the sentences.

In addition, many sentences in the KOLIMO dataset might automatically seem more poetic and thus more original because of their temporal distance from contemporary language. While identical analyses can be applied to both introduced datasets, this aspect of their design should therefore be treated with caution. In future work, the dataset collection could be expanded with additional genre-specific subsets. This would enable more fine-grained comparisons between different types of literary genres (e.g., highbrow and lowbrow literature) and allow for more systematic analyses across historical periods.

6. Conclusion

We introduce the MetaphOrig datasets, consisting of two small datasets of sentence-level spatial descriptions from different source corpora. We provide the sentences, sentence-level originality ratings from humans and LLMs, and word-level metaphor annotations following the MIPVU procedure.

Model	Corpus	β-coefficient	SE	Spearman Correlation
GPT-5	KOLIMO	8.76***	0.90	**0.8306**
	Wikivoyage	9.02***	1.82	0.5169
Qwen2.5-32B	KOLIMO	6.27***	0.74	0.7731
	Wikivoyage	3.80*	1.50	0.3506
Mistral-Small-3.2-24B	KOLIMO	9.00***	0.93	0.7631
	Wikivoyage	5.96***	1.03	**0.5294**
Llama-3.2-3B	KOLIMO	7.10***	0.92	0.6658
	Wikivoyage	4.17***	1.05	0.3799

Table 4: Slopes from CLMs showing the **strength of alignment** (β-coefficient) between human and instruction-tuned LLM originality ratings in the two corpora. Stars denote significance (*$p < .05$, **$p < .01$, ***$p < .001$). Standard Error (SE) quantifies the uncertainty of the estimated coefficient. Spearman correlations are for direct comparison between LLMs.

MetaphOrig is thus among the first datasets combining approaches from psychological creativity research assessing originality holistically with approaches from linguistics and rhetoric that focus on specific, smaller textual units. Through the inclusion of LLM ratings, the datasets allow for systematic comparisons between human and LLM perceptions of originality.

A showcase study demonstrates how analyses can explicitly connect these perspectives in practice. It suggests that human ratings of originality can be traced back to metaphor density in a literary textual context and a similar pattern applies to LLM ratings. However, deliberateness of metaphor needs to be explored in future work, including conventionality and novelty of metaphor, as well as other stylistic factors and historical distance. Such analyses can draw on our small datasets and build upon the showcase study presented in this paper.

Limitations

The model hyperparameters for the prompting experiments were not individually optimized. Adjusting parameters such as the model temperature could potentially improve alignment with human ratings. Likewise, prompt design introduces variability. Different structuring of the prompts might have yielded different outcomes. Another limitation of this study is that it only considers the originality criterion of creativity. For a more comprehensive view on creativity, follow-up studies must take a closer look at the dimension of success as well. Moreover, we acknowledge that the provided contextual information (presenting the sentences of the KOLIMO dataset as everyday language) in the rating study may have shaped the ratings, for instance by making stylistic features that are conventional in literary texts appear more original. A follow-up study will therefore systematically investigate how different priming conditions and explicit genre labels affect originality ratings.

Acknowledgements

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – CRC-1646, project number 512393437, project A05.

Bibliographical References

Selcuk Acar. 2025. Creativity Assessment, Research, and Practice in the Age of Artificial Intelligence. *Creativity Research Journal*, 37(2):181–187. Publisher: Routledge _eprint: https://doi.org/10.1080/10400419.2023.2271749.

Teresa M. Amabile. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5):997–1013. Place: US Publisher: American Psychological Association.

Christian Burgers, Elly A. Konijn, Gerard J. Steen, and Marlies A.R. Iepsma. 2015. Making ads less complex, yet more creative and persuasive: the effects of conventional metaphors and irony in print advertising. *International Journal of Advertising*, 34(3):515–532.

Qian Cao, Xiting Wang, Yuzhuo Yuan, Yahui Liu, Fang Luo, and Ruihua Song. 2026. Evaluating text creativity across diverse domains: a dataset and large language model evaluator. In *The Fourteenth International Conference on Learning Representations*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Rune H. B. Christensen. 2023. ordinal—Regression Models for Ordinal Data.

EAV Crompvoets. 2025. *Behind the scenes of pairwise comparison for educational measurement*. Ridderprint.

Jennifer Diedrich, Mathias Benedek, Emanuel Jauk, and Aljoscha C. Neubauer. 2015. Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts*, 9(1):35–40. Place: US Publisher: Educational Publishing Foundation.

Stefanie Dipper, Adam Roussel, Alexandra Wiemann, Won Kim, and Tra-My Nguyen. 2024. Guidelines for the Annotation of Intentional Linguistic Metaphor. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 53–58, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Paul V. DiStefano, John D. Patterson, and Roger E. Beaty. 2024. Automatic Scoring of Metaphor Creativity with Large Language Models. *Creativity Research Journal*. Publisher: Routledge _eprint: https://doi.org/10.1080/10400419.2024.2326343.

Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. 2018. Weeding out Conventionalized Metaphors: A Corpus of Novel Metaphor Annotations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1424, Brussels, Belgium. Association for Computational Linguistics.

Lettie Dorst. 2015. More or different metaphors in fiction? A quantitative cross-register comparison. *Language and Literature*, 24:3–22.

Markus Egg and Valia Kordoni. 2022. Metaphor annotation for German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2556–2562, Marseille, France. European Language Resources Association.

Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models. ArXiv:2407.21783 [cs].

B.A. Hennessey, Teresa M. Amabile, and J.S. Mueller. 2011. Consensual Assessment. In *Encyclopedia of Creativity*, pages 253–260. Elsevier.

J. Berenike Herrmann. 2015. High on metaphor, low on simile? An examination of metaphor type in sub-registers of academic prose. In J. Berenike Herrmann and Tony Berber Sardinha, editors, *Metaphor in Language, Cognition, and Communication*, volume 4, pages 163–190. John Benjamins Publishing Company, Amsterdam.

J. Berenike Herrmann and Gerhard Lauer. 2018. Korpusliteraturwissenschaft. Zur Konzeption und Praxis am Beispiel eines Korpus zur literarischen Moderne. *Osnabrücker Beiträge zur Sprachtheorie*, 92:127–156.

J. Berenike Herrmann, Karola Woll, and Aletta G. Dorst. 2019. Linguistic metaphor identification in German. *Metaphor Identification in Multiple Languages. MIPVU around the world*. ISBN: 9789027204721.

Rebecca M. M. Hicke and Ross Deans Kristensen-McLachlan. 2024. SCIENCE IS EXPLORATION: Computational Frontiers for Conceptual Metaphor Theory. In *CHR 2024: Computational Humanities Research Conference,*, pages 1105–1116, Aarhus, Denmark.

Jan Horstmann. 2019. KOLIMO: Korpus der literarischen Moderne. *forTEXT. Literatur digital erforschen*.

Jan Horstmann and E Akazawa. 2024. Ressourcenbeitrag: Korpus der literarischen Moderne (KOLIMO). *forTEXT, Korpusbildung*, 1(2). Medium: PDF,XML Publisher: Universitäts- und Landesbibliothek Darmstadt.

James C. Kaufman, John Baer, and Jason C. Cole. 2009. Expertise, Domains, and the Consensual Assessment Technique. *The Journal of Creative Behavior*, 43(4):223–233. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2162-6057.2009.tb01316.x.

Sungeun Kim and Dongsuk Oh. 2025. Evaluating Creativity: Can LLMs Be Good Evaluators in Creative Writing Tasks? *Applied Sciences*, 15(6):2971. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

Christine A. Knoop and Stefan Blohm. 2025. Literarinesses—A Bag of Three-Sided Coins. *Literature*, 5(3):21.

Christine A. Knoop, Thomas Nehrlich, Sabrina Aristei, Oliver Lubrich, Kirsten Stark, Alexander Enge, Werner Sommer, and Rasha Abdel Rahman. 2024. The usual miracles: How narrative style affects the processing of counterintuitive concepts. *Psychology of Aesthetics, Creativity, and the Arts*.

Katrin Kohl, Marianna Bolognesi, and Ana Werkmann Horvat. 2020. The Creative Power of Metaphor. In Katrin Kohl, Rajinder Dudrah, Andrew Gosler, Suzanne Graham, Martin Maiden, and Wen-chin Ouyang, editors, *Creative Multilingualism*, 1 edition, pages 25–46. Open Book Publishers, Cambridge, UK.

Zoltán Kövecses. 2010. *Metaphor: a practical introduction*, 2. edition edition. Oxford University Press, Oxford.

Zoltán Kövecses. 2020. Conceptual metaphor theory. In Elena Semino and Zsófia Demjén, editors, *The Routledge handbook of metaphor and language*, first issued in paperback edition, Routledge handbooks in linguistics, pages 13–27. Routledge, Taylor & Francis Group, London New York.

George Lakoff and Mark Johnson. 2003. *Metaphors We Live By*. University of Chicago Press, Chicago, IL.

Antonio Jr. Laverghetta, Tuhin Chakrabarty, Tom Hope, Jimmy Pronchick, Krupa Bhawsar, and Roger E. Beaty. 2025. How do Humans and Language Models Reason About Creativity? A Comparative Analysis. ArXiv:2502.03253 [cs].

Simone A. Luchini, Ibraheem Muhammad Moosa, John D. Patterson, Dan Johnson, Matthijs Baas, Baptiste Barbot, Iana Bashmakova, Mathias Benedek, Qunlin Chen, Giovanni E. Corazza, Boris Forthmann, Benjamin Goecke, Sameh Said-Metwaly, Maciej Karwowski, Yoed N. Kenett, Izabela Lebuda, Todd Lubart, Kirill G. Miroshnik, Felix-Kingsley Obialo, Marcela Ovando-Tellez, Ricardo Primi, Rogelio Puente-Díaz, Claire Stevenson, Emmanuelle Volle, Aleksandra Zielińska, Janet G. van Hell, Wenpeng Yin, and Roger E. Beaty. 2025. Automated assessment of creativity in multilingual narratives. *Psychology of Aesthetics, Creativity, and the Arts*. Place: US Publisher: Educational Publishing Foundation.

Rowan Hall Maudslay and Simone Teufel. 2022. Metaphorical Polysemy Detection: Conventional Metaphor meets Word Sense Disambiguation. ArXiv:2212.08395 [cs].

Mistral AI. 2025. Mistral-Small-3.2-24B-Instruct-2506.

Omar Momen, Emilie Sitter, Berenike Herrmann, and Sina Zarrieß. 2026. Surprisal and metaphor novelty: Moderate correlations and divergent scaling effects. *arXiv preprint arXiv:2601.02015*.

Seyedeh Hamideh Mozaffari. 2013. An Analytical Rubric for Assessing Creativity in Creative Writing. *Theory and Practice in Language Studies*, 3.

Andreas Nolda. 2024. Wikivoyage-Korpus: Korpusquellen der deutschen Sprachversion von Wikivoyage im TEI-Format.

Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 49:101356.

Pragglejaz Group. 2007. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, 22(1):1–39.

Meihua Qian and Jonathan A Plucker. 2017. Creativity assessment. In *Creativity and innovation*, pages 223–234. Routledge.

Abdullah Al Rabeyah, Fabrício Góes, Marco Volpe, and Talles Medeiros. 2024. Do LLMs Agree on the Creativity Evaluation of Alternative Uses? ArXiv:2411.15560 [cs].

W. Gudrun Reijnierse, Christian Burgers, Tina Krennmayr, and Gerard J. Steen. 2018. DMIP: A Method for Identifying Potentially Deliberate Metaphor in Language Use. *Corpus Pragmatics*, 2(2):129–147.

Sebastian Reimann and Tatjana Scheffler. 2024. When is a Metaphor Actually Novel? Annotating Metaphor Novelty in the Context of Automatic Metaphor Detection. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 87–97, St. Julians, Malta. Association for Computational Linguistics.

Sebastian Reimann and Tatjana Scheffler. 2025. Using Large Language Models to Perform MIPVU-Inspired Automatic Metaphor Detection. In *Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II)*, pages 10–21, Vienna, Austria. Association for Computational Linguistics.

Mark A. Runco and Garrett J. Jaeger. 2012. The Standard Definition of Creativity. *Creativity Research Journal*, 24(1):92–96. Publisher: Routledge _eprint: https://doi.org/10.1080/10400419.2012.650092.

Alec Sánchez-Montero, Gemma Bel-Enguix, Sergio-Luis Ojeda-Trueba, and Gerardo Sierra. 2025. Prompting metaphoricity: Soft labeling with large language models in popular communication of science tweets in Spanish. In *Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II)*, pages 45–56, Vienna, Austria. Association for Computational Linguistics.

Elena Semino and Gerard J. Steen. 2008. Metaphor in Literature. In Jr. Gibbs, Raymond W., editor, *The Cambridge Handbook of Metaphor*

and Thought, Cambridge Handbooks in Psychology, pages 232–246. Cambridge University Press, Cambridge.

Emilie Sitter, Omar Momen, Florian Steig, J. Berenike Herrmann, and Sina Zarrieß. 2025. Annotating Spatial Descriptions in Literary and Non-Literary Text. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 308–325, Vienna, Austria. Association for Computational Linguistics.

Gerard J. Steen. 2017. Deliberate Metaphor Theory: Basic assumptions, main tenets, urgent issues. *Intercultural Pragmatics*, 14(1):1–24. Publisher: De Gruyter Mouton.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, and Tina Krennmayr. 2010a. Metaphor in usage. 21(4):765–796. Publisher: De Gruyter Mouton Section: Cognitive Linguistics.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. 2010b. *A Method for Linguistic Metaphor Identification. From MIP to MIPVU*, volume 14 of *Converging Evidence in Language and Communication Research (CELCR)*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Theresa J. Weinstein, Simon Majed Ceh, Christoph Meinel, and Mathias Benedek. 2022. What's Creative About Sentences? A Computational Approach to Assessing Creativity in a Sentence Generation Task. *Creativity Research Journal*, 34(4):419–430. Publisher: Routledge _eprint: https://doi.org/10.1080/10400419.2022.2124777.

Hadley Wickham, Thomas Lin Pedersen, and Dana Seidel. 2011. scales: Scale Functions for Visualization. Series Title: CRAN: Contributed Packages.

Wikimedia Foundation Inc. 2025. Wikivoyage – Freie Reiseinformationen rund um die Welt.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 Technical Report. ArXiv:2407.10671 [cs].

Fengying Ye, Shanshan Wang, Lidia S. Chao, and Derek F. Wong. 2025. Unveiling LLMs' Metaphorical Understanding: Exploring Conceptual Irrelevance, Context Leveraging and Syntactic Influence. ArXiv:2510.04120 [cs].

Claire M. Zedelius, Caitlin Mills, and Jonathan W. Schooler. 2019. Beyond subjective judgments: Predicting evaluations of creative writing from computational linguistic features. *Behavior Research Methods*, 51(2):879–894.

A. Rater Instructions

A.1. Original German Instruction

Wie originell und überraschend ist diese Beschreibung für Sie? Wird in der Textstelle ein Ort so beschrieben, wie Sie es in einem alltagssprachlichen Text nicht erwartet hätten? Gibt es eine frische Perspektive oder einen unerwarteten oder neuen Ausdruck?

Originelle Beschreibungen müssen nicht unbedingt extrem kunstvoll und ausgeschmückt sein. Entscheidend ist, ob Sie überrascht sind darüber, dass jemand einen Ort auf eine solche Art und Weise beschreibt.

Beispiele:

wenig originell: *Am frühen Morgen lag die Straße, die durch einige abgelegene Orte führte, im Nebel.*

sehr originell: *Die Straße verlief im Nebel des frühen Morgens und machte sich die Mühe, winzige Orte zu besuchen, die streng genommen nicht auf ihrem Weg lagen.*

A.2. English Translation

How original or surprising is this description for you? Does it describe a place in a way you would not have expected in everyday language? Is there a fresh perspective or an unexpected or new phrasing?

Original descriptions are not necessarily very elaborate and ornate. What is relevant is whether you are surprised about someone describing a place in such a way.

Examples:

not very original: *Early in the morning, the road that led through some remote places lay in the fog.*

very original: *The road ran away in the mist of the early morning, going to some trouble to visit tiny towns which were not, strictly speaking, on its way.*

B. LLM Prompts

B.1. Original German Prompt

System Prompt

> Du bist Annotator auf Prolific mit Deutsch als Erstsprache.
>
> Du nimmst an der Studie ‘Verstehen von Beschreibungen’ teil und bewertest Beschreibungen auf auf einer Skala von 1 bis 6.
>
> Die Beschreibungen stammen aus Kontexten des alltäglichen Lebens, wie Postkarten, Textnachrichten oder Wikipedia. Alle Beschreibungen haben die Gemeinsamkeit, dass sie räumliche Gegebenheiten beschreiben.

User Prompt

> Wie originell und überraschend ist diese Beschreibung für Sie? Wird in der Textstelle ein Ort so beschrieben, wie Sie es in in einem alltagssprachlichen Text nicht erwartet hätten? Gibt es eine frische Perspektive oder einen unerwarteten oder neuen Ausdruck?
>
> Originelle Beschreibungen müssen nicht unbedingt extrem kunstvoll und ausgeschmückt sein. Entscheidend ist, ob Sie überrascht sind darüber, dass jemand einen Ort auf eine solche Art und Weise beschreibt.
>
> **Beispiele:**
>
> wenig originell: *Am frühen Morgen lag die Straße, die durch einige abgelegene Orte führte, im Nebel.*
>
> sehr originell: *Die Straße verlief im Nebel des frühen Morgens und machte sich die Mühe, winzige Orte zu besuchen, die streng genommen nicht auf ihrem Weg lagen.*
>
> **Beschreibung zur Bewertung:** {sentence}
>
> Bewerte den Satz mit einer Zahl zwischen 1 und 6 und begründe deine Entscheidung. Antworte ausschließlich im JSON-Format mit folgenden Feldern: ‘Bewertung’ und ‘Begründung’.

B.2. English Translation

System Prompt

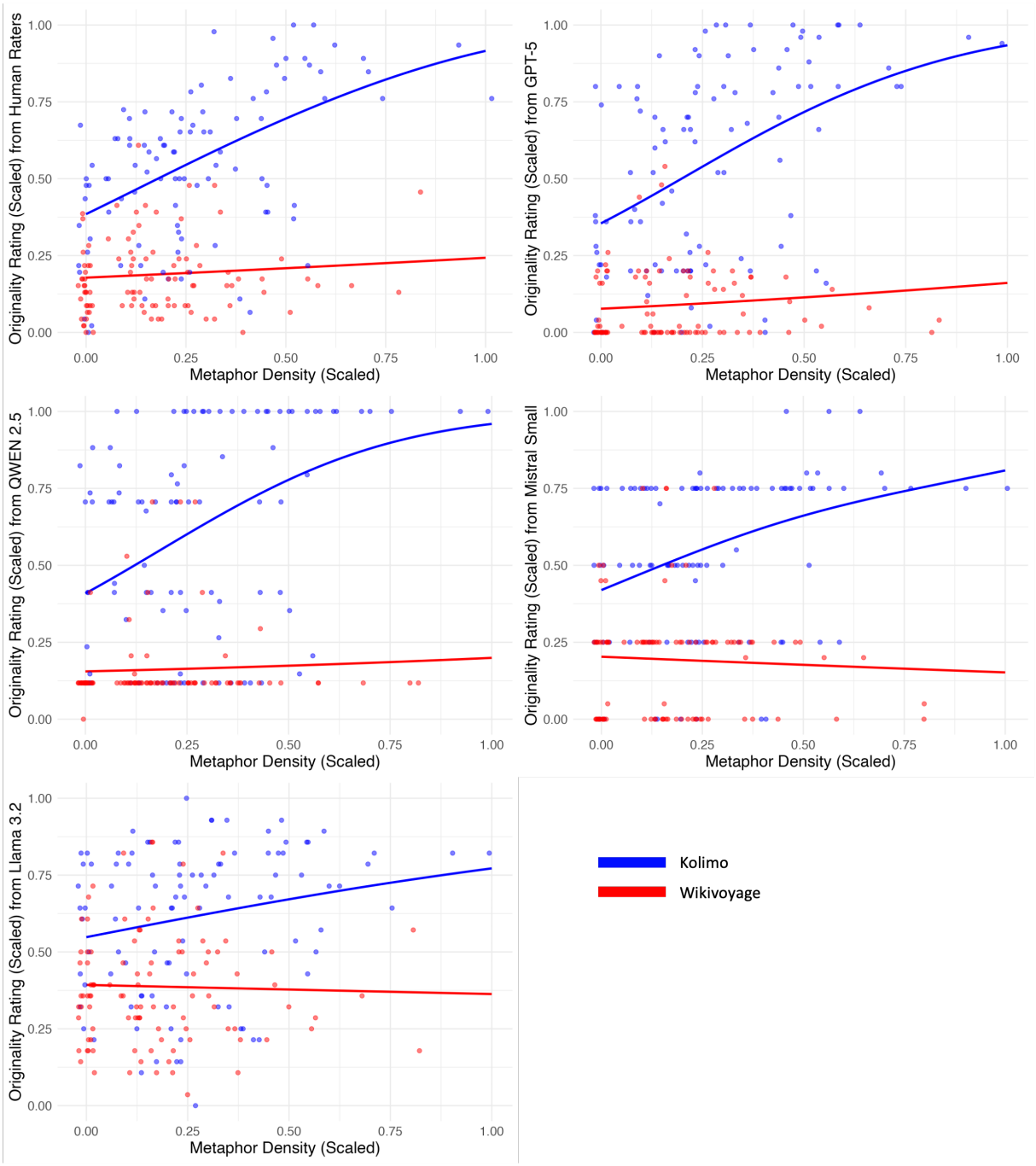
> You are an German native speaker annotator on Prolific.
>
> You are participating in the study ‘Understanding Descriptions’ and evaluating descriptions on a scale from 1 to 6.
>
> The descriptions come from everyday contexts, such as postcards, text messages, or Wikipedia. All descriptions have in common that they describe spatial surroundings.

User Prompt

> How original and surprising is this description to you? Does the passage describe a place in a way that you would not expect in everyday language? Is there a fresh perspective or an unexpected or new expression?
>
> Original descriptions do not necessarily have to be extremely artistic and embellished. What is relevant is whether you are surprised that someone would describe a place in such a way.
>
> **Examples:**
>
> not very original: *Early in the morning, the road that led through some remote places lay in the fog.*
>
> very original: *The road ran away in the mist of the early morning, going to some trouble to visit tiny towns which were not, strictly speaking, on its way.*
>
> **Description for evaluation:** {sentence}
>
> Rate the sentence with a number between 1 and 6 and explain your decision. Respond exclusively in JSON format with the following fields: ‘rating’ and ‘reason’.

C. Averaged ratings per sentence and ordinal regression lines

Increase/decrease of lines for the Wikivoyage dataset is not significant.


Originality Rating (Scaled) from Human Raters
Originality Rating (Scaled) from GPT-5
Originality Rating (Scaled) from QWEN 2.5
Originality Rating (Scaled) from Mistral Small
Originality Rating (Scaled) from Llama 3.2
Metaphor Density (Scaled)
0.00
0.25
0.50
0.75
1.00
Kolimo
Wikivoyage