Surprisal and Metaphor Novelty Judgments: Moderate Correlations and Divergent Scaling Effects Revealed by Corpus-Based and Synthetic Datasets

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

Abstract

Novel metaphor comprehension involves complex semantic processes and linguistic creativity, making it an interesting task for studying language models (LMs). This study investigates whether surprisal, a probabilistic measure of predictability in LMs, correlates with annotations of metaphor novelty in different datasets. We analyse the surprisal of metaphoric words in corpus-based and synthetic metaphor datasets using 16 causal LM variants. We propose a cloze-style surprisal method that conditions on full-sentence context. Results show that LM surprisal yields significant moderate correlations with scores/labels of metaphor novelty. We further identify divergent scaling patterns: on corpus-based data, correlation strength decreases with model size (inverse scaling effect), whereas on synthetic data it increases (quality–power hypothesis). We conclude that while surprisal can partially account for annotations of metaphor novelty, it remains limited as a metric of linguistic creativity.[1]

1 Introduction

Recent advances in language modelling have led to a renewed interest in studying linguistic creativity, an aspect of language that was notoriously challenging for traditional NLP systems. A well-known instance of creative language is metaphor, which arises through mapping of a source domain–a lexical unit's literal meaning–onto a target domain–its figurative meaning-(Lakoff and Johnson, 1980). Yet, not all metaphors are equally novel or creative as many of such mappings are highly conventionalized, such as in the famous example "to *attack* an argument," where the mapping from the source domain "WAR" onto the target domain "ARGUMENT" corresponds to a conventional sense of the verb "attack" which can be found in dictionaries. Other metaphors, as in "The *arrested* water" (Figure 1) establish a more unconventional or novel mapping that is considered creative.

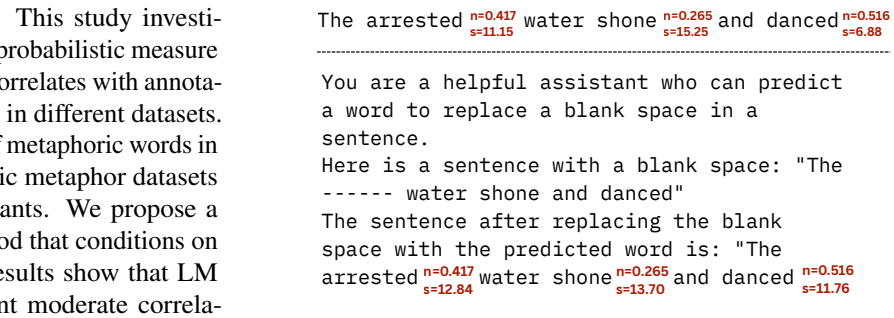

Figure 1: A sentence from VUA-ratings with metaphor-novelty ratings (n) and surprisal measures (s) from GPT2-base. The top part shows the direct-surprisal measure, while the bottom part shows the cloze-surprisal measure.

The distinction between conventional and novel metaphors is well established in theory (Bowdle and Gentner, 2005; Gibbs, 2005) and studied in experimental work on cognitive processing (Arzouan et al., 2007; Lai et al., 2009). Novel metaphors require greater interpretative effort compared to conventional metaphors, as the unfamiliar mapping requires speakers to construct new connections between domains (Philip, 2016).

However, separating novel metaphors from conventional ones remains a challenging task for human annotators (Do Dinh et al., 2018; Reimann and Scheffler, 2024). While computational methods on metaphor analysis do not often take the novelty dimension into account (Maudslay and Teufel, 2022), and some existing studies highlight that novel metaphors are more difficult to detect than conventional ones (Neidlein et al., 2020; Tong et al., 2024). In this work, we propose to fill this gap by studying questions related to LMs' processing of novel and conventional metaphors and investigating correlations between LM-based metrics and different setups of metaphor novelty annotations.

[1]Code and data are publicly available: https://github.com/OmarMomen14/surprisal-metaphor-novelty.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8113–8127

We draw on a line of research that has studied the ability of LMs to account for effects of difficulty in human sentence processing and that goes back to surprisal theory (Hale, 2001; Levy, 2008). Surprisal is computed with LMs as the negative log probability of a word in context and has been found to provide a robust predictor for human processing difficulty (e.g., of reading times) (Goodkind and Bicknell, 2018; Shain et al., 2024).

However, recent work draws a mixed picture in terms of which LMs can provide the most robust and cognitively plausible predictors for processing difficulty. Oh and Schuler (2023b) observe an inverse scaling effect when testing surprisal estimates from GPT-2 models of different sizes, showing that surprisal computed with smaller model sizes achieved a better fit with human reading times than larger models. Wilcox et al. (2023), on the other hand, train LMs of small and medium size on a range of languages and find that LM quality generally correlates with its predictive power of reading times.

In this paper, we investigate surprisal of metaphoric words computed using a selection of LMs as a metric of metaphor novelty. We find significant, positive and moderate correlations between LMs' surprisal and different metaphor novelty annotations. We perform our experiments on four different datasets coming from corpus-based and synthetic setups, using LMs of different sizes. Most interestingly, we observe effects supporting the "inverse scale" pattern (Oh and Schuler, 2023b) on the two corpus-based novelty datasets, and contrary patterns supporting the counter-argument, "Quality-Power" hypothesis (Wilcox et al., 2023) on the two synthetic datasets. We also investigate the effect of instruction-tuning on surprisal correlation to metaphor novelty. Moreover, we conduct a deeper analysis of the genre splits, revealing that genre, metaphor density, and LM perplexity are potential factors underlying the quality of surprisal as a predictor of metaphor novelty. Finally, we introduce a new method of computing surprisal, *cloze-surprisal*, to include the right context of the word in its conditional probability. We find that this method can boost the correlation of surprisal with metaphor novelty annotations by a few points. In general, our study establishes a promising direction for studying linguistic creativity with LMs and calls for novel measures and datasets that provide systematic annotations of metaphor novelty across genres.

2 Related Work

2.1 Metaphor Annotations

The most common annotation scheme in corpus-based metaphor studies is the Metaphor Identification Procedure Vrije Universiteit (MIPVU) (The Pragglejaz Group, 2007; Steen et al., 2010), designed as a reliable step-by-step framework for identifying metaphorically used words. This method was used to construct the VU Amsterdam Metaphor Corpus (VUAMC) (Steen et al., 2010), a large-scale, genre-balanced corpus containing 186,673 words sampled from fiction, news, academic writing, and conversations splits in the BNC Baby edition. The VUAMC has become a benchmark for metaphor research, inspiring developments in further English corpora (Beigman Klebanov et al., 2018; Mohammad et al., 2016) and more recent multilingual efforts (Sanchez-Bayona and Agerri, 2022; Egg and Kordoni, 2022).

Several studies have extended annotation beyond binary literal/metaphoric labels. Mohler et al. (2016) introduced a four-point metaphoricity scale, considering factors such as vividness and familiarity. Reijnierse et al. (2019) proposed the concept of *deliberateness*, distinguishing metaphors intended to be recognized as metaphors; novel metaphors are typically deliberate under this framework. Direct novelty annotations were first introduced by (Parde and Nielsen, 2018), who asked annotators to rate metaphorical word pairs from the VUAMC on a 0–3 novelty scale. Do Dinh et al. (2018) proposed a more comprehensive approach considering all metaphor words in VUAMC, and by aggregating ranked novelty judgments into continuous scores. More recently, Reimann and Scheffler (2024) proposed a dictionary-based method that labels a metaphor as novel when its contextual meaning is absent from dictionary entries.

Psycholinguistic studies also produce datasets of metaphors, but these are smaller in size and feature controlled synthetic sentences (fixed words, structures, sentence length, etc.), cf. (Cardillo et al., 2010, 2012; Roncero and de Almeida, 2014). Some of these also created datasets of metaphors that are classified as conventional or novel metaphors (Lai et al., 2009; Ahrens et al., 2024).

2.2 Predictive Powers of Surprisal

Surprisal estimates from LMs have been shown to capture human processing difficulty across multiple behavioural and neural measures. In self-paced

reading and eye-tracking studies, token-level surprisals from causal models significantly predict reading times beyond lexical and syntactic factors (Goodkind and Bicknell, 2018; Wilcox et al., 2020; Oh and Schuler, 2023a). Similar effects have been reported for acceptability judgments, where LM surprisals align with human sensitivity to grammaticality (Lau et al., 2017; Meister et al., 2021; Hu et al., 2020; Tjuatja et al., 2025).

2.3 LM-Based Metrics for Metaphor Novelty

Recent work has explored diverse LM-based approaches—prompting, surprisal, embedding similarity, and attention patterns—for the analysis of metaphorical language. Aghazadeh et al. (2022) demonstrate that metaphor-relevant information is encoded in mid-layer embeddings of multilingual pre-trained LMs. Ichien et al. (2024) show that GPT-4 generates interpretations of novel literary metaphors that are favoured by human judges over interpretations by college students, suggesting sensitivity to metaphorical meaning beyond lexical overlap. Pedinotti et al. (2021) tested BERT's masked token probabilities across conventional, novel, and nonsensical metaphors, showing that novel metaphors tend to receive lower probabilities, but distinctions between novelty and nonsense remained unclear. We extend this study towards more recent LMs and broader comparisons between models, datasets, and genres. Djokic et al. (2021) trained a BERT-based classifier to predict novelty scores jointly with the task of metaphor detection.

3 Metaphor Novelty Datasets

This Section describes the four datasets we use in our experiments and explains how metaphor novelty is annotated in different approaches. Datasets statistics are reported in Tables 6 and 7 (Appendix A.1).

3.1 Corpus-based Datasets

Following the MIPVU procedure, all words in the VUAMC are annotated as being a metaphor-related word (MRW) or not. Out of 186,673 words in VUAMC, 24,762 words (15,155 content words) are labelled as MRWs. While the original VUAMC does not include novelty annotations, we utilise two annotation studies that offer different annotation protocols for metaphor novelty on the same corpus.

VUA-ratings Do Dinh et al. (2018) collected crowd-sourced ratings of metaphor novelty for VUAMC. In their set-up, annotators were presented four sentences containing an MRW (content words only) and asked to select the best (most novel) and worst (most conventional). Each MRW appeared in six different best-worst scaling comparisons. These annotations were then converted into continuous Best-Worst Scaling scores (Kiritchenko and Mohammad, 2017) in the range of (-1, +1), with -1 being the most conventional and +1 being the most novel. Additionally, they also convert these scores to binary labels using a threshold of 0.5. This results in labelling 353 metaphors as novel out of 15,155 content metaphoric words in VUA.

VUA-dictionary Reimann and Scheffler (2024) proposed a dictionary-based method that labels a metaphor as novel when its contextual meaning is absent from dictionary entries. They applied this method to VUAMC. In particular, they re-annotated the potentially novel metaphors in VUA according to VUA-ratings and Reijnierse et al. (2019)'s metaphor deliberateness annotations (1,160 potentially novel metaphors in total). Their procedure resulted in labelling only 409 content[2] metaphoric words as novel out of the 1,160 potentially novel metaphors. We assume the remaining metaphors in VUA to be conventional in our study.

3.2 Synthetic Datasets

Another class of metaphor novelty datasets is synthetic datasets, which are mainly characterised by being generated from a fixed source. They usually have comparable sentences in terms of the target metaphoric words. We consider two cases of such datasets, a dataset from a psycholinguistic study concerned with novel metaphors. And a toy dataset that we generated from GPT-4o.

Lai2009 Lai et al. (2009) investigated how our brains handle conventional and novel metaphors differently. To this end, they constructed a set of controlled experimental items featuring metaphors with two degrees of novelty. Two linguists selected 104 words and constructed 4 sentences for each word, according to the Conceptual Metaphor Theory (CMT) (Lakoff and Johnson, 1980). For each word, the items include one (i) literal use, (i) conventional metaphor, (iii) novel metaphor, and

[2]We exclude non-content words from this dataset to allow comparison with VUA-ratings, which originally excluded non-content words.

(iv) anomalous use, with the target word as the last word in each sentence. Familiarity and interpretability tests showed a significant difference between the conventional and novel metaphoric senses. For our experiments, we select only the conventional and novel metaphor senses for each word, resulting in 208 sentences (104 conventional and 104 novel metaphors).

GPT-4o-metaphors To test our experiments in a setting that is more controlled (in contrast to VUA) but features more varied sentence lengths, structures and degrees of novelty (in contrast to Lai2009), we construct a synthetic dataset by prompting GPT-4o to generate sentences that include conventional and novel metaphoric senses. We do not assume that LLMs are able to generate ideal novel metaphors, or that this dataset is a benchmark of any kind; we are only interested in exploring a new quick setting of potential metaphor novelty annotations. We prompt GPT-4o to generate 5 verbs and 5 nouns that can be used in a metaphoric sense. For each word, we prompt the model again to generate 10 different sentences using the target word in a conventional metaphor sense, and 10 different sentences in a novel metaphor sense. This results in a dataset of 200 sentences, 100 contain conventional metaphors, and 100 contain novel metaphors.

4 Methods & Experiments

We describe our experiments designed to investigate whether and to what extent surprisal scores computed from LMs correlate with different annotations of metaphor novelty.

4.1 Surprisal of a target word

In information theory (Shannon, 1948), the information content of an event x with probability $p(x)$ is defined as: $I(x) = -\log p(x)$. In the context of LMs trained to predict the next token in a sequence, this quantity is called surprisal and computed for a w_i in a sequence such that: $\text{Surprisal}(w_i) = -\log p(w_i \mid w_{<i})$[3].

In our experiments, we measure word-level surprisal of metaphoric words in their sentence-level context. We feed every sentence to an LM in an independent teacher-forced forward pass, and record surprisal of the target (metaphoric) word(s). We denote this quantity as *direct-surprisal* to distinguish it from *cloze-surprisal*.

Most LMs operate on subwords rather than words, so deriving word-level probabilities requires implementation choices that can affect surprisal estimates and, in turn, downstream measures (Pimentel and Meister, 2024; Lesci et al., 2025). For transparency and reproducibility, we report two choices in our implementation that affect the computed surprisal values. **First**, we choose to align tokens (subwords) to words by precomputing the offsets of the exact[4] target words in their corresponding sentences, and then searching for the minimal span of tokens in the tokenisation of the input sentence covering the target word's offsets. This usually results in a proper alignment with the target word in addition to a leading whitespace character (e.g. "`Ġarrested`" for "arrested"), with very rare cases when the last token is attached to a punctuation (e.g. `["Ġindivid", "ual", "ism,"]` for "individualism"). We also apply the corrections made by (Pimentel and Meister, 2024; Oh and Schuler, 2024) that address the problem of leading whitespace in tokenisations of most causal language models. **Second**, surprisal is computed using the conditional probability of the current word being the target word given the preceding words in a sentence. This makes computing surprisal of the first word in a sentence an issue[5]. To compute surprisal of the first word in a sentence, we prepend the input sentence with a "beginning of sequence" special token.

Cloze-surprisal A key limitation of standard (left-to-right) surprisal as a proxy for semantic contextual properties (e.g., metaphor novelty) is that it does not condition on future (right) context in the sentence. This matters in naturally occurring corpus data (e.g VUA datasets), where metaphorical words can appear in many sentence positions, and in some cases in the GPT-4o-metaphors dataset (see Table 8). To incorporate the right context while retaining an autoregressive setup, we compute *cloze-surprisal*: for each target word, we prompt the model to predict a missing word in a sentence, and we replace the target in the sentence by a blank space, and then we append the same sentence again as the intended completion (Figure 1). We measure

[3]We use log of base e for all surprisal/perplexity computations in our study.

[4]We make sure no punctuation or any other characters outside of the word boundary are included inside the offsets.

[5]In some cases in the VUA datasets, the target word is the first word in the sentence.

surprisal for the target word at its position in the second occurrence of the sentence. Because the first occurrence exposes the full right context, the resulting conditional probability effectively incorporates both left and right contexts of the target word.

4.2 Evaluating the correlation

To determine whether LMs' surprisal correlates with different metaphor novelty annotations, we use multiple correlation metrics between surprisal and metaphor novelty scores/labels. For continuous novelty scores (as in VUA-ratings), we compute **Pearson's** r (Pearson, 1895) and **Spearman's** ρ (Spearman, 1904) correlations. For binary labels (conventional vs. novel), we compute the **Rank-biserial** r_b (Glass, 1966) correlation. Rank-biserial estimates the probability that a random observation from the set of novel metaphors has a larger surprisal than one from the set of conventional metaphors, minus the reverse probability. We also estimate the potential of surprisal as a discriminator for binary novelty labels using the Area Under the ROC Curve **(AUC)** (Fawcett, 2006). We also compute the significance of all these estimates. All our metrics and tests do not assume normality, except for Pearson's, for which we also provide a non-parametric alternative (Spearman).

4.3 Settings

Surprisal measures are derived from the learnt $p(x)$ of pre-trained LMs and, hence, depend substantially on the model architecture and the training data. Also, human annotations of metaphor novelty depend on the annotation process. Additionally, novelty norms may vary by genre: a metaphor considered novel in academic texts might appear conventional in fiction. We thus investigate how all these factors affect the correlation by experimenting with multiple settings.

Models We examine three families of decoder-only causal LMs (GPT-2, Llama 3 and Qwen2.5) (Radford et al., 2019; Grattafiori et al., 2024; Yang et al., 2024). For each model family, we select 3-4 different sizes, to represent the effect of model size on the correlation. To investigate the effect of instruction-tuning on the correlation, we include the instruction-tuned variants of Llama 3 and Qwen2.5.

Datasets As explained in Section 3, we perform experiments on four datasets (Tables 6 and 7). As VUA-ratings has continuous novelty scores, we evaluate the correlation of its continuous scores to surprisal. In addition, we convert these scores to binary labels using a threshold of 0.5 and evaluate their correlation to surprisal, allowing us to compare the results of VUA-ratings to VUA-dictionary, Lai2009 and GPT-4o-metaphors, which only have binary novelty annotations.

Genre variables VUA provides genre splits (see Section 3.1), allowing us to analyse the correlation between surprisal and novelty separately for each genre.

Perplexity Scales of surprisal can differ from one genre to another, and from one model to another. To this end, we investigate perplexity as a potential factor in our study. Perplexity of a model on a certain dataset is the exponential of the average token-level surprisal of all tokens in the dataset. In that sense, perplexity indicates how much a certain model is "surprised" on average when predicting a certain dataset. We measure models' perplexity on the genre splits of VUA by feeding sentences one by one to the model, measuring token-level surprisals accordingly and averaging and exponentiating them to obtain perplexity. This yields higher values of perplexity than common values reported in literature, as we use a shorter context (single sentences).

4.4 Summary

Our experiments rely on a large collection of surprisal measures for each metaphor word in the four datasets: 32 different surprisal values (including direct and cloze) for 15,155 metaphoric words of the VUA dataset and for 208 and 200 metaphoric words of Lai2009 and GPT-4o-metaphors, respectively. In Figure 3 (Appendix A.6), we plot the distributions of surprisal values by model, in addition to metaphor novelty scores/labels from VUA. From these plots and further normal distribution tests, we find that novelty scores and surprisal values are not strictly normally distributed.

5 Results

We report the correlation measures for the corpus-based datasets in Table 1, and for synthetic datasets in Table 2. We also report the gains (in terms of rank-biserial) of the instruction-tuned variants over their base models in Table 3. And the gains of the cloze-surprisal method over the direct-surprisal method in Table 4.

Model	VUA-ratings				VUA-dict.	
	r	ρ	r_b	auc	r_b	auc
GPT2-base	.419	.417	.638	.819	.581	.791
GPT2-med	.389	.383	.600	.800	.557	.778
GPT2-large	.381	.373	.585	.793	.539	.769
GPT2-xl	.373	.362	.566	.783	.539	.769
Llama-1B	.345	.329	.532	.766	.480	.740
Llama-3B	.328	.308	.502	.751	.446	.723
Llama-8B	.314	.293	.488	.744	.431	.716
Qwen-0.5B	.384	.377	.598	.799	.543	.771
Qwen-7B	.334	.314	.502	.751	.456	.728
Qwen-14B	.316	.295	.470	.735	.430	.715

Table 1: Correlation estimates between surprisal and novelty scores/labels in the corpus-based datasets. All results are statistically significant at the 0.001 level.

Model	Lai2009		GPT-4o-met.	
	r_b	auc	r_b	auc
GPT2-base	.276	.638	.511	.756
GPT2-med	.362	.681	.586	.793
GPT2-large	.397	.699	.629	.814
GPT2-xl	.414	.707	.587	.794
Llama-1B	.450	.725	.508	.754
Llama-3B	.451	.725	.511	.755
Llama-8B	.483	.742	.557	.778
Qwen-0.5B	.374	.687	.382	.691
Qwen-7B	.494	.747	.469	.734
Qwen-14B	.504	.752	.536	.768

Table 2: Correlation estimates between surprisal and novelty labels in the synthetic datasets. All results are statistically significant at the 0.001 level.

5.1 Overall Correlation Results

Generally, we find LMs' direct surprisal values correlate positively with metaphor novelty annotations across the four datasets (Tables 1 and 2). All correlation estimates are statistically significant for both the corpus-based and synthetic datasets [6]. Overall, for direct surprisal, the largest Pearson correlation $r = .419$, largest Spearman correlation $\rho = .417$, largest rank-biserial correlation $r_b = .638$ and largest $AUC = .819$ come from GPT2-base on VUA-ratings. These values indicate a significant positive correlation, yet its strength is moderate.

We also find that the correlation strengths' ranges differ across the different annotation datasets. By comparing the rank-biserial estimate r_b across the four datasets, we find its ranges as follows: (.47-.64) for VUA-ratings, (.43-.58) for VUA-dictionary, (.28-.50) for Lai2009, and (.38-.63) for GPT-4o-metaphors. These results show that surprisal correlates more strongly with the corpus-based data than with the synthetic data. Also, surprisal correlates more strongly with human ratings on VUA than the dictionary-based annotation approach. While the most controlled dataset (Lai2009) gets the weakest correlations with surprisal.

5.2 Model Size Effects

The effect of model size is consistent (but diverging in direction) across the two dataset types. We plot the rank-biserial correlations of the ten model variants for the four datasets in Figure 2. For the two corpus-based datasets (blue), surprisal–novelty correlation is monotonically decreasing as model size increases (per model family). On the other hand, for the two synthetic datasets (red), surprisal–novelty correlation increases as model size increases, with a single minor exception at the GPT2 model family on the GPT-4o-metaphors dataset.

[6]The relatively large corpus-based datasets (15,155 datapoints) can bias the significance values; however, we get similar significance values for the relatively small synthetic datasets (100 datapoints).

Model	VUA-r	VUA-d	Lai	GPT-4o
Llama-1B-It.	+4.0	+1.5	+0.4	−5.1
Llama-3B-It.	+4.0	+2.4	−5.8	−5.6
Llama-8B-It.	+1.6	+0.6	−2.3	−5.1
Qwen-0.5B-It.	−1.6	−0.8	+0.1	+2.2
Qwen-7B-It.	−1.7	−1.9	−8.9	−13.7
Qwen-14B-It.	−3.9	−3.1	−3.7	−14.0

Table 3: Instruction-tuning % gains (over corresponding base variants) in Rank-biserial's correlation estimates between surprisal and novelty scores/labels in the four datasets: **VUA-r**atings, **VUA-d**ictionary, **Lai**2009 and **GPT-4o**-metaphors.

5.3 Instruction-tuning Effects

We report percentage gains in rank-biserial correlation for the instruction-tuned variants over their base variants across the four datasets in Table 3. This shows the effect of extracting surprisals from an instruction-tuned variant over a base variant with the same experimental setting. We do not add a prompt/instruction to the instruction-tuned models' inputs. The results suggest that instruction-tuning does not always improve the correlation between surprisal and novelty annotations. Only Llama instruction-tuned variants could improve the correlations over their base variants on the VUA datasets; otherwise, instruction-tuning deteriorates the correlations over the base variants. Also, instruction-tuning fails to improve the correlations

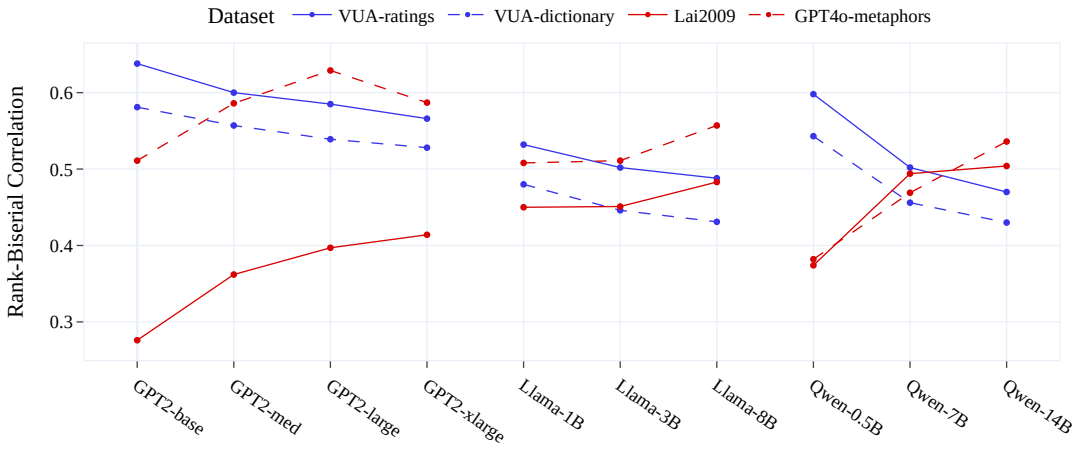

Figure 2: Effect of model size on correlation between surprisal and metaphor novelty annotations from the four datasets. Corpus-based datasets (blue) show a negative scale effect, while synthetic datasets (red) show a positive scale effect.

over more basic models such as GPT2-base.

5.4 Cloze-surprisal

In Table 4, we report the percentage gains in rank-biserial correlations for cloze-surprisal over direct-surprisal from same model variants across the four datasets. We find cloze-surprisal to be boosting the correlations with extra points in many cases; however, it also deteriorates the correlations heavily in other cases. For the GPT2 model family, cloze-surprisal consistently improves the corpus-based datasets' correlations, yielding the strongest correlations in this study. Out of the 32 recorded surprisals on each dataset, cloze surprisal from GPT2-base achieves the highest correlations on VUA-ratings, with $r = .499$, $\rho = .499$, $r_b = .687$ and $auc = .843$. However, cloze-surprisal of GPT2 models consistently deteriorates the correlations on the synthetic datasets. Llama and Qwen model families also show positive gains from cloze surprisal in most of the cases, however, not as consistently as for GPT2 models on VUA. Most importantly, unlike the GPT2 model family, Llama and Qwen boost correlations on the synthetic datasets except for a very few cases. Cloze-surprisal achieves the overall strongest correlation on the Lai2009 dataset using the Qwen-14B model, with $r_b = .539$ and $auc = .779$. Also, it achieves the overall strongest correlation on the GPT-4o-metaphors dataset using the Llama-8B model, with $r_b = .684$ and $auc = .842$.

To investigate the effect of incorporating the whole sentence context in surprisal computation, we plot the distribution of LMs' direct and cloze surprisals of metaphoric words in VUA in Figure 3 (Appendix A.6). We observe that the distribution of direct-surprisal is often right-skewed with many values near zero. While cloze-surprisal shifts the distribution away from zero and approaches a normal distribution. Furthermore, we plot the distribution of GPT2-base direct and cloze surprisal of metaphoric words in VUA at different sentence positions in Figure 4 (Appendix A.6). Again, we observe that cloze-surprisal shifts the distributions away from near zero values. However, against our intuition, the effect of cloze-surprisal is not more impactful in cases where the metaphor is located earlier in the sentence than later. The effect of cloze-surprisal, although positive, is not related to the position of the metaphoric word in the sentence.

5.5 Genre Splits

In Table 5, we report the rank-biserial correlations between GPT2-base surprisals and binary novelty labels of VUA-ratings and VUA-dictionary on each genre split separately. We also report the share of novel metaphors within each genre to illustrate the variance in metaphor density across genres. Additionally, we report the perplexity of GPT2-base on each genre split separately.

For VUA-ratings, we find a positive relation between genres' metaphor density and the correlation of surprisal with novelty. We also observe a positive relation between perplexity and surprisal-

Model	VUA-r	VUA-d	Lai	GPT-4o
GPT2-base	+4.9	+4.0	−7.7	−16.7
GPT2-med	+6.8	+4.5	−11.8	−26.9
GPT2-large	+6.2	+8.2	−14.2	−25.1
GPT2-xl	+6.3	+2.5	−16.0	−28.1
Llama-1B	+7.3	+8.0	−13.3	−15.7
Llama-1B-It.	−3.8	−1.2	−2.4	−10.3
Llama-3B	−0.5	+5.6	+0.3	+0.7
Llama-3B-It.	−11.1	−2.3	+4.7	+6.5
Llama-8B	+1.9	+2.7	+2.5	+12.7
Llama-8B-It.	−7.8	−0.2	+6.0	+10.9
Qwen-0.5B	−2.1	+2.1	−2.5	+6.2
Qwen-0.5B-It.	−4.5	−1.6	−3.9	+7.0
Qwen-7B	+3.7	+5.6	+2.4	+12.8
Qwen-7B-It.	−6.9	−1.6	+9.1	+8.0
Qwen-14B	+5.8	+7.8	+3.5	+6.9
Qwen-14B-It.	−4.4	+0.7	+4.6	+12.8

Table 4: Cloze-surprisal % gains (over direct-surprisal from the same model) in Rank-biserial's correlation estimates between surprisal and novelty scores/labels in the four datasets: **VUA-r**atings, **VUA-d**ictionary, **Lai**2009 and **GPT-4o**-metaphors.

novelty correlations that is only violated on the conversation split. On the other hand, for VUA-dictionary, the relation between metaphor density and surprisal-novelty correlation is not preserved. While the perplexity positive relation with surprisal-novelty correlation is more concrete. Interestingly, in contrast to VUA-ratings, the dictionary-based annotations of the conversation split correlate strongly with surprisal values despite its tiny number of annotated novel metaphors (11 out of 1774 metaphors) and high perplexity (134).

Genre		VUA-ratings		VUA-dictionary	
	ppl.	r_b	Nov. %	r_b	Nov. %
Fiction	108	.693	2.97	.478	3.56
News	89	.653	2.80	.413	3.88
Academic	73	.588	1.85	.413	1.85
Conversation	134	.482	1.41	.800	0.62
All	96	.638	2.33	.581	2.70

Table 5: GPT2-base perplexity (**ppl.**), correlation between GPT2-base surprisal and novelty labels (r_b) and novel metaphor percentage (**Nov. %**) in VUA-ratings and VUA-dictionary per genre.

6 Discussion

Is surprisal a good metric for metaphor novelty? Across multiple novelty annotation setups—human ratings, dictionary-based binary labels, experimental contrastive items designed by experts, and LLM-generated conventional vs. novel senses—we find consistently moderate associations between LM surprisal and metaphor novelty (best: $r = .49, \rho = .50, r_b = .69, \text{AUC} = .84$). While these results are not directly comparable to prior work due to differences in datasets and task formulations, their magnitudes are broadly in line with reported surprisal–behavior associations in acceptability judgments (Tjuatja et al., 2025) and reading times (Oh and Schuler, 2023a). The closest point of reference in the metaphor novelty literature is (Parde and Nielsen, 2018), who report $r = .44$ when predicting novelty scores from a wide range of linguistic features. They additionally report moderate inter-annotator agreement, underscoring that novelty itself is challenging to measure reliably by humans. At the same time, surprisal has clear theoretical limits as a standalone predictor: interpreting a metaphor involves semantic integration and cross-domain mapping, not only predictability. We therefore expect surprisal to be most informative when combined with measures that more directly target novel interpretations and domain mappings, especially for highly creative (novel) metaphors.

Model Sizes & Dataset Types: Our results introduce a great opportunity to further understand the underlying factors of opposing negative and positive effects of model sizes on correlations with psycholinguistic and cognitive features. The negative and positive effects are present together in our study, and clearly contrasted by the type of metaphor novelty dataset under experiment.

The negative effect of model size observed in corpus-based datasets mirrors the inverse scaling effect reported in reading time studies (Oh and Schuler, 2023b; Wilcox et al., 2025). Recently, (Oh et al., 2024) argued that this effect is largely driven by word frequency, with larger models assigning increasingly non-human-like expectations to rare words. Since corpus-based novelty scores are often correlated with word frequency (Do Dinh et al., 2018; Reimann and Scheffler, 2024), frequency can be an underlying factor for this negative scaling effect in our study. By contrast, the synthetic datasets control for lexical identity and frequency: conventional and novel senses are elicited for the same set of words, and stimuli are constructed to cleanly separate the two senses. Under these controls, scaling improves alignment with novelty labels, agreeing with reports of positive scaling effects in other behavioural settings such as acceptability judgments and other reading time studies (Tjuatja et al., 2025; Wilcox et al., 2023).

Overall, we see our results as evidence that these diverging scaling effects are due to the nature of the dataset types. While corpus-based datasets reflect metaphor novelty mainly through lexical properties such as word frequency, synthetic datasets more directly isolate the conventional–novel distinction by controlling for lexical properties confounds.

Cloze-surprisal and Instruction-tuning: Our cloze-surprisal approach improves correlation across many model variants. In Figures 3 and 4, cloze-surprisal is found to be shifting the surprisal values away from near-zero values and pushing towards a normal distribution.
In examples from Table 8, cloze-surprisal often raises the surprisal of the metaphorical word, consistent with the intuition that the full context is needed to process the metaphor and realise its true predictability. Moreover, in cases where the metaphorical word begins a sentence (e.g. example 14), direct surprisal is naturally high—regardless of novelty—whereas cloze-surprisal successfully moderates such inflated values.
Instruction-tuning—despite its goal of aligning model outputs with human intent—does not enhance the human-likeness of surprisal. In fact, similar to prior findings (Kuribayashi et al., 2024), instruction-tuned models tend to reduce alignment between predicted probabilities and human annotation scores.

Genre Effects: Our results show a significant effect of sentences' genre on surprisal correlation with metaphor novelty. We suspect the difference in genre's novel metaphor density can be an underlying factor. Also, our observations on perplexity relations to the correlations trigger the possibility that the amount of pretraining data contributing to each genre can significantly affect LM-based methods of detecting novel metaphors.

7 Conclusion

We have studied the distinction between conventional and novel metaphors and systematically investigated surprisal computed with LMs as a metric for metaphor novelty. In general, our experiments show some potential for surprisal in predicting aspects of linguistic creativity, but also call for novel measures and datasets that provide systematic annotations of metaphor novelty across genres and across corpus-based and experimental settings.

Limitations

We acknowledge a couple of limitations in this work. First, the scarcity of high-quality metaphor novelty annotations in existing literature constrains both coverage and generalizability. Second, we rely on pretrained language models whose training data and processes are not fully disclosed. Additionally, although we carefully analyse the effects of model architecture, size, and domain (genre), future work could adopt mixed-effects models to test the interaction of these variables.

Acknowledgments

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – CRC-1646, project number 512393437, project A05. We acknowledge the anonymous reviewers and area chairs for their valuable comments and feedback. Furthermore, we thank Özge Alaçam, Annett Jorschick, Vicky Tzuyin Lai and Tiago Pimentel for their cooperation and responsiveness to our inquiries.

References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.

Kathleen Ahrens, Christian Burgers, and Yin Zhong. 2024. Making the unseen seen: The role of signaling and novelty in rating metaphors. *Journal of Psycholinguistic Research*, 53(3):36.

Yossi Arzouan, Abraham Goldstein, and Miriam Faust. 2007. Brainwaves are stethoscopes: Erp correlates of novel metaphor comprehension. *Brain Research*, 1160:69–81.

Beata Beigman Klebanov, Chee Wee (Ben) Leong, and Michael Flor. 2018. A corpus of non-native written English annotated for metaphor. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 86–91, New Orleans, Louisiana. Association for Computational Linguistics.

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

E. R. Cardillo, G. L. Schmidt, A. Kranjec, and A. Chatterjee. 2010. The neural bases of metaphoric expression comprehension. *Behavior Research Methods*, 42(3):651–660.

E. R. Cardillo, C. E. Watson, G. L. Schmidt, A. Kranjec, and A. Chatterjee. 2012. Selective metaphor impairments after left, not right, hemisphere injury. *Frontiers in Psychology*, 3:87.

Goran Djokic, Ekaterina Shutova, and Verna Dankers. 2021. Episodic memory demands modulate novel metaphor use during event narration. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*.

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.

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.

Tom Fawcett. 2006. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874.

Raymond W. Gibbs, Jr. 2005. *Embodiment and Cognitive Science*. Cambridge University Press.

Gene V. Glass. 1966. Note on rank-biserial correlation. *Educational and Psychological Measurement*, 26(3):623–631.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint*.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Nicholas Ichien, Dušan Stamenković, and Keith J. Holyoak. 2024. Large language model displays emergent ability to interpret novel literary metaphors. *Preprint*, arXiv:2308.01497.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. 2024. Psychometric predictive power of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1983–2005, Mexico City, Mexico. Association for Computational Linguistics.

Vicky Tzuyin Lai, Tim Curran, and Lise Menn. 2009. Comprehending conventional and novel metaphors: An erp study. *Brain Research*, 1284:145–155.

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

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Pietro Lesci, Clara Meister, Thomas Hofmann, Andreas Vlachos, and Tiago Pimentel. 2025. Causal estimation of tokenisation bias. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28325–28340, Vienna, Austria. Association for Computational Linguistics.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Rowan Hall Maudslay and Simone Teufel. 2022. Metaphorical polysemy detection: Conventional metaphor meets word sense disambiguation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 65–77, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).

Arthur Neidlein, Philip Wiesenbach, and Katja Markert. 2020. An analysis of language models for metaphor recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3722–3736, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Byung-Doh Oh and William Schuler. 2023a. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921, Singapore. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2023b. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Byung-Doh Oh and William Schuler. 2024. Leading whitespaces of language models' subword vocabulary pose a confound for calculating word probabilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3464–3472, Miami, Florida, USA. Association for Computational Linguistics.

Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2644–2663, St. Julian's, Malta. Association for Computational Linguistics.

Natalie Parde and Rodney Nielsen. 2018. A Corpus of Metaphor Novelty Scores for Syntactically-Related Word Pairs. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Karl Pearson. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242.

Paolo Pedinotti, Eliana Di Palma, Ludovica Cerini, and Alessandro Lenci. 2021. A howling success or a working sea? testing what BERT knows about metaphors. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 192–204, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gill Philip. 2016. Conventional and novel metaphors in language. In Elena Semino and Zsófia Demjén, editors, *The Routledge Handbook of Metaphor and Language*, page 14. Routledge, London / New York. "Conventional and Novel Metaphors in Language".

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18358–18375, Miami, Florida, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*.

W. Gudrun Reijnierse, Christian Burgers, Tina Krennmayr, and Gerard J. Steen. 2019. Metaphor in communication: the distribution of potentially deliberate metaphor across register and word class. *Corpora*, 14(3):301–326.

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.

C. Roncero and R. G. de Almeida. 2014. The role of expectedness and predication in metaphor comprehension. *Language and Cognition*, 6(2):141–168.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2022. Leveraging a new Spanish corpus for multilingual and cross-lingual metaphor detection. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 228–240, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3 and 4):379–423, 623–656. Reprinted with corrections.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15:72–101.

G.J. Steen, A.G. Dorst, J.B. Herrmann, A.A. Kaal, T. Krennmayr, and T. Pasma. 2010. *A method for linguistic metaphor identification. From MIP to MIPVU.* Number 14 in Converging Evidence in Language and Communication Research. John Benjamins.

The Pragglejaz Group. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

Lindia Tjuatja, Graham Neubig, Tal Linzen, and Sophie Hao. 2025. What goes into a LM acceptability judgment? rethinking the impact of frequency and length. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2173–2186, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3517–3536, Bangkok, Thailand. Association for Computational Linguistics.

Ethan Wilcox, Clara Meister, Ryan Cotterell, and Tiago Pimentel. 2023. Language model quality correlates with psychometric predictive power in multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7503–7511, Singapore. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *CoRR*, abs/2006.01912.

Ethan Gotlieb Wilcox, Michael Y. Hu, Aaron Mueller, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Adina Williams, Ryan Cotterell, and Tal Linzen. 2025. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, 144:104650.

An Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, and .... 2024. Qwen2.5 technical report. *arXiv preprint*.

A Appendix

A.1 Statistics and Examples

In Tables 6 and 7, we describe in numbers the four datasets under study. In Table 8, we list 16 examples from the VUA datasets, 6 examples from the LAI2009 dataset, and 4 examples from the GPT-4o-metaphors dataset.

A.2 GPT-4o-metaphors Construction

To generate the sentences in GPT-4o-metaphors, we prompt GPT-4o once for each word in the dataset. As we planned for 10 different words, we prompted the model 10 consecutive times to construct this dataset. `"I am curating a dataset to be used in a study about metaphoric knowledge in pretrained language models PLMs e.g. GPT-2. The dataset should consist of sentences that correspond to target nouns/verbs. For each target noun/verb, there should be 10 sentences. The target noun/verb should be used 10 times using conventional metaphoric meanings of the target noun/verb, and 10 times using novel metaphoric meanings of the target noun/verb. Please suggest a target noun, and generate 20 sentences following the requirements above: 10 sentences with Conventional Metaphor usages, and 10 sentences with Novel Metaphor usages."`

A.3 Models Details

All the LMs used in this study are based on the HuggingFace HF models hub. Hence, we list here the exact HF models' IDs that we used in our experiments. We use default HF parameters when forwarding the inputs through the model layers. For models meta-data and parameters, please refer to the models' cards on HF using the links below:

1. openai-community/gpt2
2. openai-community/gpt2-medium
3. openai-community/gpt2-large
4. openai-community/gpt2-xl
5. meta-llama/Llama-3.2-1B
6. meta-llama/Llama-3.2-1B-Instruct
7. meta-llama/Llama-3.2-3B
8. meta-llama/Llama-3.2-3B-Instruct
9. meta-llama/Llama-3.1-8B
10. meta-llama/Llama-3.1-8B-Instruct
11. Qwen/Qwen2.5-0.5B
12. Qwen/Qwen2.5-0.5B-Instruct
13. Qwen/Qwen2.5-7B
14. Qwen/Qwen2.5-7B-Instruct
15. Qwen/Qwen2.5-14B
16. Qwen/Qwen2.5-14B-Instruct

			VUA-ratings		VUA-dictionary
Genre	# Metaphors	L_{sent} mean | std.	Novelty Score mean | std.	# Novel >= 0.5	# Novel
Fiction	3170	26.0 | 16.5	-.005 | .271	94	113
News	4712	29.9 | 14.2	.000 | .257	132	183
Academic	5499	34.9 | 16.0	.003 | .239	102	102
Conversation	1774	17.5 | 15.9	-.000 | .236	25	11
All	15155	29.4 | 16.5	.000 | .251	353	409

Table 6: Distributions and statistics of the corpus-based datasets. # Metaphors is the number of metaphoric words in corpus, L_{sent} is the length of sentences in words. Novelty Score is (Do Dinh et al., 2018) novelty scores, # Novel is the number of novel metaphors (Novelty Score >= 0.5 for VUA-ratings).

	Lai2009		GPT-4o-metaphors	
Label	#	L_{sent} mean | std.	#	L_{sent} mean | std.
Conventional	104	6.7 | 1.5	100	10.1 | 1.9
Novel	104	6.7 | 1.4	100	12.7 | 2.3
All	208	6.7 | 1.5	200	11.4 | 2.5

Table 7: Distributions and statistics of the synthetic datasets. # is the number of instances, and L_{sent} is the length of sentences in words.

A.4 Use of AI Assistants

AI assistants were used during manuscript preparation only for limited linguistic editing to improve clarity and style, and for writing auxiliary code (e.g., for visualisations). They were not used for scientific reasoning, evaluation decisions, or interpretation of results; all analyses and conclusions were drawn by the authors.

A.5 Scientific Artifacts

VUA datasets are publicly available datasets and intended for scientific research, and we follow this purpose in this study. Lai2009 dataset has copyright regulations; the author of this dataset approved our usage of the dataset, and she promised to share the data with anyone interested in reproducing our experiments. We double-checked that the datasets and prompts do not contain any personal data.

A.6 Distributions Visualisations

In Figure 3, we plot the distributions of novelty scores/labels and surprisal of the VUA datasets. In Figure 4, we plot scatter plots between VUA-ratings scores and GPT2-base surprisal from both direct and cloze methods, grouping instances based on the location of the metaphoric word within the sentence.

VUA datasets				
Sentence	ratings	dictionary	direct	cloze
1. ‘ Tell him I am very sorry, but I must **fill** the quota. ’	-0.441	conventional	9.043	7.317
2. Adam might have escaped the file memories for years, **suppressed** them and jerked violently <14> by those events.	0.531	novel	14.23	14.69
3. It was an excitement that <11> and I had long dreamed of that scatter of tiny, **magically** named islands strewn across one third of a globe.	0.278	conventional	10.17	16.08
4. The seemingly random and <11> designed to disguise a boat’s shape from the **prying** eyes of U-Boat captains, so it <10> in the Bahamas.	0.588	conventional	7.903	12.46
5. One Mr Clarke can not duck away from if he wants to **avoid** a second Winter of Discontent	-0.094	conventional	3.165	7.454
6. This was conveniently **encapsulated** in the first try.	0.500	conventional	7.918	14.54
7. Thrusts of resistance (mass demonstrations, resignations, tax rebellions, etc) would come in **crests**.	0.382	novel	12.46	16.84
8. Travel: A **pilgrimage** sans progress Elisabeth de Stroumillo potters round Poitou	0.514	conventional	8.188	14.81
9. The Tehuana dress is by no means the most decorative variant or the **closest** to pre-Hispanic forms of clothing.	-0.194	conventional	6.466	11.92
10. Interwoven with these images are subtler references to the **metaphorical** borderlines which separate Latin American <5> and North America.	0.529	conventional	10.16	12.57
11. This is often linked with a supposed denunciatory effect — the idea that the mandatory life sentence **denounces** murder as emphatically as possible <18> this crime.	0.294	conventional	11.02	12.61
12. He certainly held deep convictions as to the <9>, but at least a part of his apparent **hostility** was assumed for the occasion, a hard <7> in the end.	0.514	conventional	5.662	9.781
13. Me dad said he’s had enough Well, we were debating whether to **give** it to you or not.	-0.633	conventional	3.752	8.038
14. Struggled with it a little	0.552	conventional	17.13	14.92
15. That’s an old **trick**.	0.310	conventional	4.013	11.92
16. Can you **sort** erm, madame out?	0.567	conventional	8.820	9.591
LAI2009				
Sentence		label	direct	cloze
17. Upon hearing the news my spirits **sank**		conventional	4.358	11.82
18. Upon having the data my prediction **sank**		novel	10.28	13.27
19. Those chess players are prepared for **battle**		conventional	5.373	12.32
20. Those plastic surgeons are prepared for **battle**		novel	6.691	13.28
21. His mental condition remains **fragile**		conventional	7.251	13.34
22. His website popularity remains **fragile**		novel	9.538	13.95
GPT-4o-metaphors				
23. Her family was her emotional **anchor** during the crisis.		conventional	4.322	14.12
24. The smell of coffee became an **anchor** to mornings that no longer came.		novel	7.518	12.42
25. The software helps users **navigate** complex legal documents.		conventional	3.830	5.575
26. He tried to **navigate** the silence like a sailor without stars.		novel	8.041	8.402

Table 8: Examples from each dataset. The metaphor word is in **boldface** within sentences. For simpler presentations, we remove some words from long sentences and replace them with a tag of the number of words removed, e.g. <11>. **rating** is the (Do Dinh et al., 2018) score, and **dictionary** is the (Reimann and Scheffler, 2024) annotation. **direct** is the GPT2-Base recorded surprisal, and cloze is the GPT2-Base cloze-surprisal. Paired examples are picked randomly for VUA datasets from the same genres to illustrate the differences between conventional and novel instances according to “ratings”. Sentences 1-4 are from Fiction, 5-8 from News, 9-12 from Academic and 13-16 from Conversation. Also obviously, (Do Dinh et al., 2018) ratings and (Reimann and Scheffler, 2024) annotations do not agree in many cases.

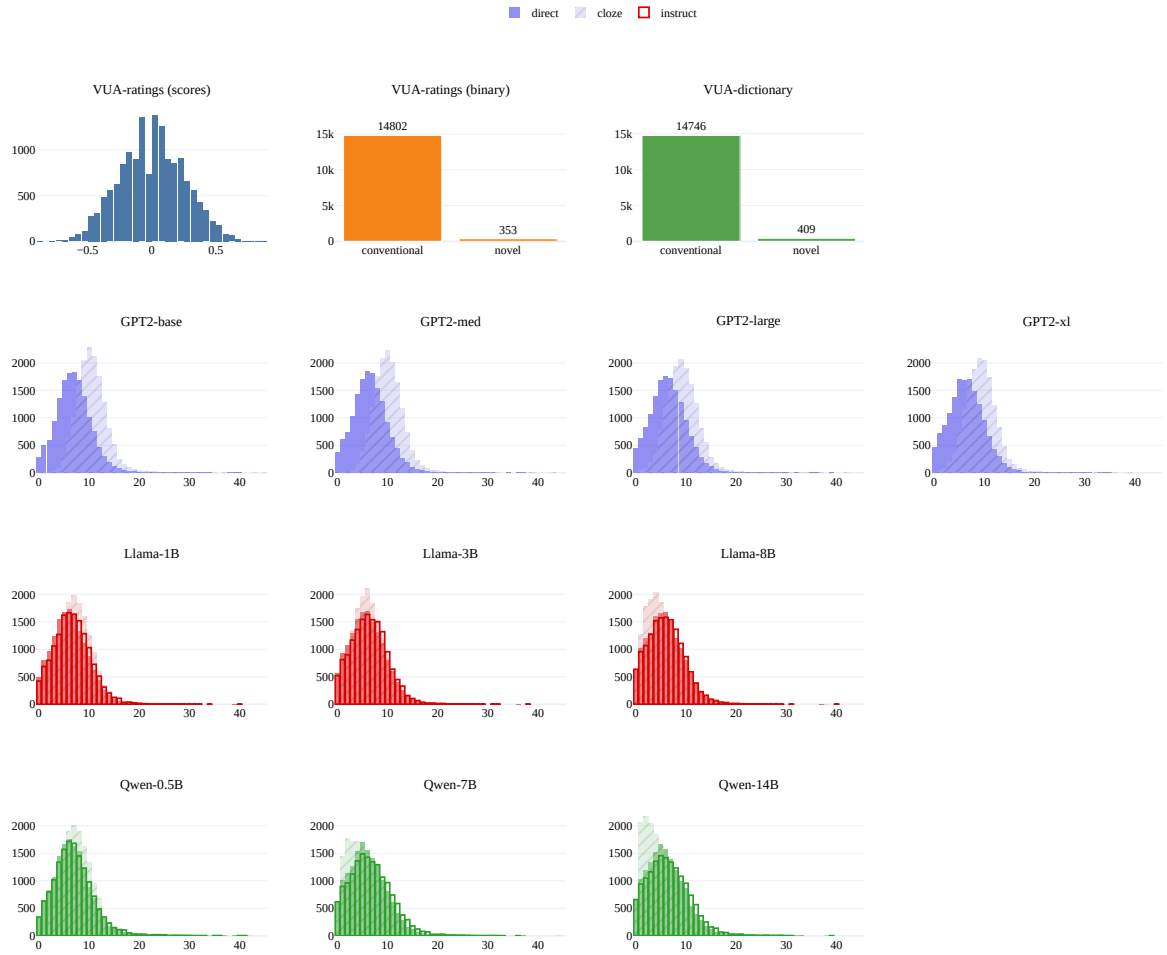

Figure 3: Histograms of VUA metaphor novelty annotations, and GPT2-base direct surprisal (dark colour) of content metaphoric words (15,155) across 10 LMs. Additionally, cloze-surprisal of the same models is plotted in (dashed) light colour, and surprisal from instruct-tuned variants (whenever applicable) is plotted with borderlines.

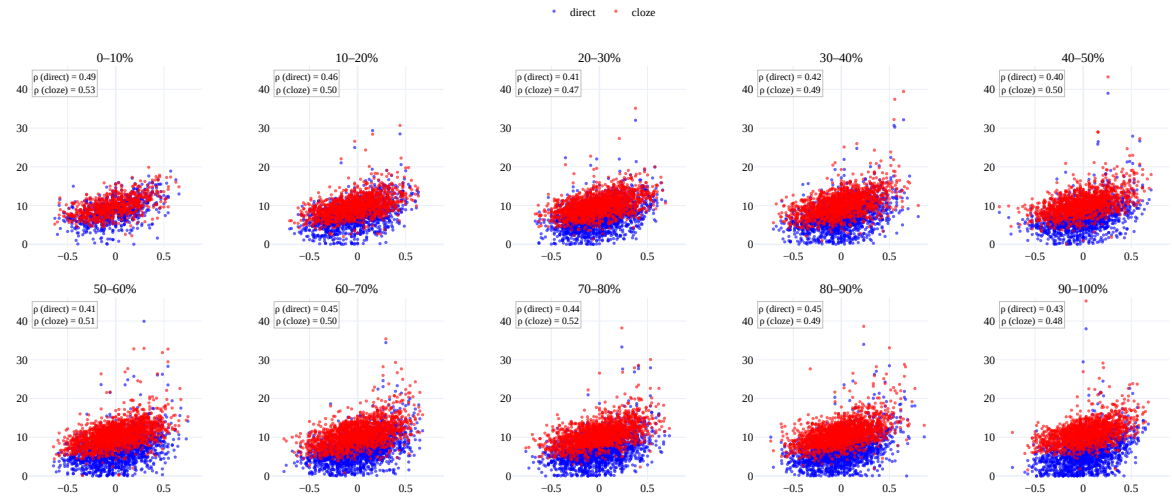

Figure 4: Scatter plots between metaphor novelty scores of VUA-ratings on the x-axis; and direct surprisal (blue) and cloze surprisal (red) (from GPT2-base) on the y-axis at different relative positions of metaphoric words in sentences. 0-10% means metaphoric words located at the first $1/10 * Length$ of the sentence.