Novelty Isn't Just Surprisal: Diverging Scaling Effects in Metaphor Novelty Prediction Revealed by Corpus-based and Synthetic Datasets

Anonymous ACL submission

Abstract

Metaphor comprehension involves complex semantic processes and linguistic creativity, which makes it an interesting task for probing LLMs. This study investigates whether surprisal, a probabilistic measure of predictability in LMs, correlates with human annotations of metaphor novelty. We analyse surprisal from 26 LLMs on one corpus-based and two synthetic metaphor novelty datasets. Results show that causal LLMs produce significant yet moderate correlations with novelty scores, while masked language models (MLMs) fail to capture metaphorical surprise. Our proposed cloze-surprisal method improves these correlations by conditioning on full-sentence context. We further identify divergent scaling patterns: on corpus-based data, correlation strength decreases with model size (inverse scaling effect), but on synthetic data, it increases (quality-power hypothesis). We conclude that while surprisal can partially account for human judgments of metaphor novelty, it remains limited as a metric of linguistic creativity.[1]

1 Introduction

Recent advances in language modeling 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 (1) „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, though, as in example (2) "The *arrested* water" (Figure 1) establish a more unusual or novel mapping that can be considered creative.

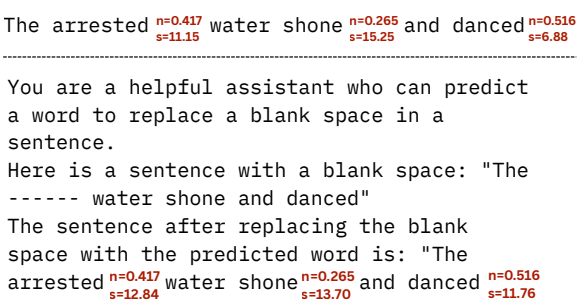

Figure 1: A sentence from VUA-METANOV with metaphor-novelty annotations (n) and surprisal measures (s) from GPT-2-base. The top part shows the direct-surprisal measure, while the bottom part shows the cloze-surprisal measure.

The distinction between conventional and novel metaphors is well established in theory (Bowdle and Gentner, 2005; Gibbs, 2005) and studied in experimental work on human metaphor 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, corpus-based and computational work on metaphor detection and understanding does not often take this distinction 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; Reimann and Scheffler, 2024). In this work, we propose to fill this gap and study questions on the processing of novel and conventional metaphors with the help of language models. In particular, we ask whether language models show increased processing difficulty for novel compared to conventional metaphors, as has been found for

[1] Code and data will be released here: https://tinyurl.com/2n7nd3j3.

humans (Arzouan et al., 2007; Lai et al., 2009).

To address these questions, 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. Oh and Schuler (2023b) observe a so-called inverse scaling effect when testing surprisal estimates from GPT-2 models of different sizes, showing that the surprisal computed with smaller model sizes achieved a better fit with human reading times than larger model sizes. 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 psychometric predictive power.

In this paper, we investigate surprisal on metaphor words computed with LLMs as a metric of metaphor novelty. We find significant, yet weak-to-moderate correlations between LLMs' surprisal and metaphor novelty annotations. We perform our experiments on three different datasets coming from corpus-based and experimental/synthetic annotation setups, using LMs of different sizes and architectures. Most interestingly, we observe effects supporting the "inverse scale" pattern (Oh et al., 2025) on a corpus-based annotations dataset, and contrary patterns supporting the counter-argument, "Quality-Power" hypothesis (Wilcox et al., 2023) on the experimental and synthetic datasets. We perform a deeper analysis on the level of genre splits, revealing that genre (and the LMs' perplexity on text of this genre) affects the quality of surprisal as a predictor of metaphor novelty significantly. Finally, we introduce a new variant of computing surprisal, *cloze-surprisal*, for recording word-level surprisal from causal LLMs in a way that includes the right context of the word in its conditional probability. We find that this method can boost the correlation of surprisal to metaphor novelty annotations by a few points. In general, our experiments point to some interesting directions for studying linguistic creativity with LMs and call 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 lexical units sampled from fiction, news, academic writing, and conversations 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 LLMs 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 such as GPT-2 and GPT-3 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 LLM-based surprisals align with human sensitivity to gradient grammaticality (Lau et al., 2017; Meister et al., 2021; Hu et al., 2020).

2.3 LM-Based Metrics for Metaphor Novelty

Recent work has explored diverse LM-based approaches—surprisal, embedding similarity, attention patterns, or activation probing—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. Only two studies explicitly correlate LM-derived measures with human novelty judgments. Pedinotti et al. (2021) test 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 deviancy remain unclear. We extend this study towards more recent LLMs and broader comparisons between models, datasets, and genres. Bunescu and Uduehi (2022) introduce distribution-based surprise metrics—such as expectation–realisation gaps and KL divergences—from masked LM output distributions and report positive correlations with human novelty ratings on metaphor and humour datasets, particularly when combined with traditional information-content measures. While they compare novel metaphors against literal senses, our study aims to distinguish conventional from novel metaphors. We also focus on raw surprisal values from both causal and MLM models as a predictor of continuous and categorical metaphor novelty annotations.

3 Metaphor Novelty Datasets

This Section describes the three metaphor datasets we use in our experiments and explains how they annotate metaphor novelty.

3.1 Corpus-based Dataset

VUA-METANOV Following the MIPVU procedure, all lexical units in the VUAMC are annotated for being a metaphor-related word (MRW) or not. Out of 186,673 lexical units in VUAMC, 24,762 units are labelled as MRWs. While the original VUAMC does not include novelty annotations, we make use of the crowd-sourced novelty annotations by (Do Dinh et al., 2018). In their set-up, annotators were presented four sentences containing an MRW 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 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. Do Dinh et al. (2018) use a threshold of 0.5 to transform these continuous scores to binary labels (conventional/novel). They also filtered out non-content MRWs like prepositions and auxiliary verbs, resulting in 15,155 content MRWs (after merging compound terms) and corresponding novelty scores. The mean length of sentences is 29.4 words with a standard deviation of 16.5. Henceforth, we refer to this novelty-annotated dataset as VUA-METANOV. In Table 6, we present 16 examples from the dataset, four examples from each genre, where two of them have a score lower than 0.5 (conventional), and two have a score greater than or equal 0.5 (novel).

3.2 Synthetic Datasets

LAI2009-METANOV Our second dataset is based on (Lai et al., 2009)'s experimental study on metaphor processing from the area of neuropsychology. To investigate how our brains handle conventional and novel metaphors differently, Lai et al. (2009) construct a set of controlled experimental items featuring metaphors with two degrees of novelty. Two linguists selected 104 words (21 nouns, 42 verbs, and 21 adjectives), 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 (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). The mean length of sentences is 6.7 words with a standard deviation of 1.5. We refer to this dataset as LAI2009-METANOV from here on. Table 6 presents 6 examples from this dataset, a conventional and a novel example from 3 different words.

GPT4o-METANOV To test our experiments in a setting that is controlled (in contrast to VUA-METANOV) but features more varied sentence lengths, structures and degrees of novelty (in contrast to LAI2009-METANOV), we construct a synthetic dataset by prompting GPT4o to generate sentences that include conventional and novel metaphoric senses. We do not assume that LLMs are able to generate ideal novel metaphors, but we are interested in exploring a new setting of 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 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. The mean length of sentences is 11.4 words with a standard deviation of 2.5. We refer to this dataset as GPT4o-METANOV. Prompts used to construct this dataset are listed in the Appendix. Table 6 presents 4 examples from this dataset, a conventional and a novel example from 2 different words.

4 Methods & Experiments

We describe our experiments designed to investigate whether and to what extent surprisal scores computed from LLMs can predict human annotations of metaphor novelty.

4.1 Surprisal as a metric of metaphor novelty

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 (in the case of causal LMs), this quantity is called surprisal and computed as in Equation 1[2].

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

$$\text{(Causal)}\ \ \text{Surprisal}(w_i) = -\log\, p(w_i \mid w_{<i})$$
$$\text{(MLM)}\ \ \text{Surprisal}(w_i) = -\log\, p(w_i \mid \mathbf{w}_{\setminus i}) \quad (1)$$

In our experiments, we measure word-level surprisal of metaphoric words in their sentence-level context. We feed every sentence to an LLM in an independent teacher-forced forward pass, and record the surprisal of the target (metaphoric) word(s). For MLM models, we mask the metaphoric word within the input sentence and compute its surprisal after the forward pass. In case of multiple metaphoric words in one sentence (only in VUA-METANOV), we record the surprisal of each metaphor word in an independent forward pass. When the target word is split into multiple tokens, we retain only the maximum surprisal value of the split tokens as a representative for the entire word.

Cloze-surprisal A potential inherent weakness of word-level surprisal computed from causal models is that it does not take the right context of the metaphoric word into account. This is an issue when sentences do not end in the metaphorical word, as in the LAI2009-METANOV dataset (see Section 3.2), but feature various positions of metaphorical word as in naturally occurring corpus data (see Section 3.1). To address this, we experiment with a new strategy for computing word-level surprisal autoregressively, while introducing the full sentence context: for each metaphor word, we input a prompt (as illustrated in Figure 1) that indicates a cloze test and show the whole sentence to the model except the metaphor word (replaced by a blank space), and then spell out the whole sentence again as if the model is solving the cloze test. We measure the surprisal of the word at the second occurrence of the sentence, ensuring that the left and right context are included in the conditional probability of the metaphor word. We apply this strategy to the VUA-METANOV dataset only, as in the other two datasets, the metaphoric words often come close to or at the end of the sentence.

4.2 Evaluating the correlation

To determine whether LMs distinguish between conventional and novel metaphors, we measure the correlation between surprisal and metaphor novelty scores. For continuous novelty scores, we compute Pearson's and Spearman's correlations. For binary labels (conventional or novel), we use the Rank-

biserial correlation. We also estimate the potential of surprisal as a discriminator for binary novelty labels using the Area Under the ROC Curve (AUC).

4.3 Settings

Surprisal measures are derived from the learnt $p(x)$ of pre-trained LLMs 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 examine how genre-specific scales influence the correlation between surprisal and human novelty annotations.

Models We examine the two types of LLM architectures, causal LMs and MLMs, as causal LLMs are currently prevalent while MLMs might provide more natural surprisal scores for metaphor novelty, given their bidirectional attention heads. We examine 3 families of causal LLMs (GPT-2, Llama 3 and Qwen2.5) (Radford et al., 2019; Grattafiori et al., 2024; Yang et al., 2024) and 4 families of MLM LLMs (BERT, RoBERTa, DeBERTa and ModernBERT) (Devlin et al., 2019; Zhuang et al., 2021; He et al., 2021; Warner et al., 2025). For each model family, we select 2-4 different sizes, based on availability, to represent the effect of model size on the correlation. Recent causal LLMs (e.g Llama 3 and Qwen 2.5) are often published as a pre-trained and an additionally instruction-tuned variant. Hence, we also consider the instruction-tuned variants whenever possible.

Datasets As explained in Section 3, we perform experiments on three different datasets (Table 5). As VUA-METANOV has continuous novelty scores, we evaluate the correlation of its continuous scores to surprisal. In addition, we convert these scores to binary labels and evaluate their correlation to surprisal, allowing us to compare the results of VUA-METANOV to LAI2009-METANOV and GPT4o-METANOV which only have binary novelty annotations.

Genre variables VUA-METANOV 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 differ from one model to another, and from one genre to another. For example, a model that is pre-trained on a lot of fiction texts would be (on average) less surprised than a model pre-trained on texts not including fiction. Perplexity of a model on a certain dataset is the exponential of the average word-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-METANOV by feeding sentences one by one to the model, measuring word-level surprisals accordingly and averaging and exponentiating them to obtain perplexity. This yields higher values of perplexity than common values reported in the literature as we use a shorter context (single sentences).

Model Variant	**Pearson** r	**Spearman** ρ
	direct | cloze	direct | cloze
GPT-2-base (137M)	.372 | **.446**	.382 | **.452**
GPT-2-medium (380M)	.303 | .424	.349 | .430
GPT-2-large (812M)	.324 | .411	.339 | .418
GPT-2-xlarge (1.6B)	.313 | .392	.330 | .401
Qwen2.5 (0.5B)	.348 | .360	.351 | .359
Qwen2.5 (0.5B) It.	.329 | .332	.342 | .332
Qwen2.5 (7B)	.303 | .307	.296 | .293
Qwen2.5 (7B) It.	.277 | .213	.272 | .207
Qwen2.5 (14B)	.285 | .289	.277 | .272
Qwen2.5 (14B) It.	.232 | .204	.228 | .192
Llama3.2 (1B)	.309 | .401	.305 | .402
Llama3.2 (1B) It.	.310 | .311	.309 | .308
Llama3.2 (3B)	.293 | .340	.286 | .339
Llama3.2 (3B) It.	.297 | .249	.294 | .245
Llama3.1 (8B)	.281 | .274	.272 | .263
Llama3.1 (8B) It.	.280 | .204	.274 | .190

Table 1: Correlation estimates between surprisal and continuous novelty scores in VUA-METANOV. All results in are statistically significant at the 0.01 level. It. refers to Instruction-tuned models.

4.4 Summary

Our experiments rely on a large collection of surprisal measures for each metaphor word in the three datasets: 42 different surprisal (including direct and cloze) values for 15,155 metaphoric words of the VUA-METANOV dataset and 26 different surprisal (direct only) values for 208 and 200 metaphoric words of LAI2009-METANOV and GPT4o-METANOV, respectively.

In Figure 4, we plot the distributions of surprisal values by model, in addition to metaphor novelty scores from VUA-METANOV. From these plots and further normal distribution tests, we find that novelty scores and surprisal values are not strictly normally distributed.

Model Variant	R-Biserial r_b	AUC
	direct | cloze	direct | cloze
GPT-2-base (137M)	**.523** | .468	**.761** | .734
GPT-2-medium (380M)	.501 | .495	.751 | .748
GPT-2-large (812M)	.482 | .496	.741 | .748
GPT-2-xlarge (1.6B)	.462 | .468	.731 | .734
Qwen2.5 (0.5B)	.507 | .470	.754 | .735
Qwen2.5 (0.5B) It.	.491 | .429	.746 | .715
Qwen2.5 (7B)	.430 | .472	.715 | .736
Qwen2.5 (7B) It.	.411 | .365	.705 | .682
Qwen2.5 (14B)	.395 | .477	.697 | .738
Qwen2.5 (14B) It.	.349 | .371	.675 | .686
Llama3.2 (1B)	.446 | .492	.723 | .746
Llama3.2 (1B) It.	.476 | .421	.738 | .710
Llama3.2 (3B)	.427 | .397	.713 | .699
Llama3.2 (3B) It.	.460 | .354	.730 | .677
Llama3.1 (8B)	.417 | .417	.708 | .708
Llama3.1 (8B) It.	.431 | .353	.715 | .676

Table 2: Correlation estimate (Rank-Biserial) and AUC between surprisal and binarised novelty scores in VUA-METANOV (score $\geq$ 0.5 = novel). All of the results in this table are statistically significant at the 0.01 level. It. refers to Instruction-tuned models.

5 Results

We found causal LMs to consistently outperform MLMs in all experimental settings. Therefore, the following result discussion focuses on causal LMs. Results on MLMs are presented in the Appendix.

5.1 Results on the corpus-based dataset

First, we analyse the correlation between surprisal and metaphor-novelty annotations in VUA-METANOV. Table 1 and 2 present correlation estimates for direct and cloze surprisal for continuous and binarised novelty scores, respectively.

Correlation effects All correlations of surprisal with VUA-METANOV novelty annotations shown in Tables 1 and 2 are statistically significant (at a significance level of 0.01). However, overall, the correlation coefficients indicate a weak correlation strength between 0.2 and 0.4. Certain GPT-2-based settings achieve a coefficient above 0.4, pointing to a moderate correlation effect. Thus, in general, we find that surprisals computed from SOTA LLMs show consistent correlation patterns with human metaphor novelty scores, but cannot explain the full variance in the data and cannot be regarded as a fully reliable predictor of metaphor novelty.

Surprisal vs. Cloze-Surprisal As shown in Tables 1 and 2, our proposed cloze-surprisal method improves the continuous correlation estimates of the overall best model variant (GPT-2-base), providing the strongest correlation across all model variants on VUA-METANOV (from $r = .372, \rho = .382$ to $r = \mathbf{.446}, \rho = \mathbf{.452}$). We observe similar improvements for cloze-surprisal on all the other GPT-2 variants. However, this pattern is not observed consistently in the binary correlation estimates and for other models than GPT-2.

Model comparison Generally, the smallest model variant across all causal models we tested, GPT-2-base, shows the strongest correlation estimates ($r = .372, \rho = .382, r_b = .523, AUC = .761$). Across all model families, we observe that the continuous correlations get consistently weaker as the model's size increases, which holds for direct and cloze-surprisal. Instruction-tuning shows different effects between model families: instruction-tuned Qwen obtains weaker correlation estimates (continuous and binary) in contrast to pre-trained variants. For Llama, instruction-tuning shows a deteriorating effect with cloze-surprisal and a positive effect with direct surprisal.

Genre effects Table 3 shows spearman's correlations for direct surprisal against the continuous VUA-METANOV novelty scores split by genre. These results yield a surprisingly consistent pattern: the correlation estimate for each model is strongest on fiction, reaching a coefficient of $\rho = \mathbf{.501}$ for GPT-2-base. The correlation is the weakest on conversations, where the coefficient drops to 0.25 for GPT-2-base and even more for other models. In general, the correlation strengths decrease as follows for almost all model variants: ($\rho_{fiction} > \rho_{news} > \rho_{academic} > \rho_{conversation}$).

Relating perplexity to correlation strength To investigate the connection between the inverse scaling effect (i.e., correlations decrease with increasing model size) and genre divergences, we measure the perplexity of each model variant on each genre split. Figure 2 plots Spearman's correlations against perplexity on each genre split, highlighting perplexity as a potential underlying factor: Figure 2 shows that the correlation between novelty scores and surprisal of LLMs increases as perplexity increases, suggesting that more powerful models underestimate processing difficulty in novel metaphors. However, this relation only holds to some extent: when perplexity increases beyond a certain level, as is the case for the Conversation genre, the correlation drops again.

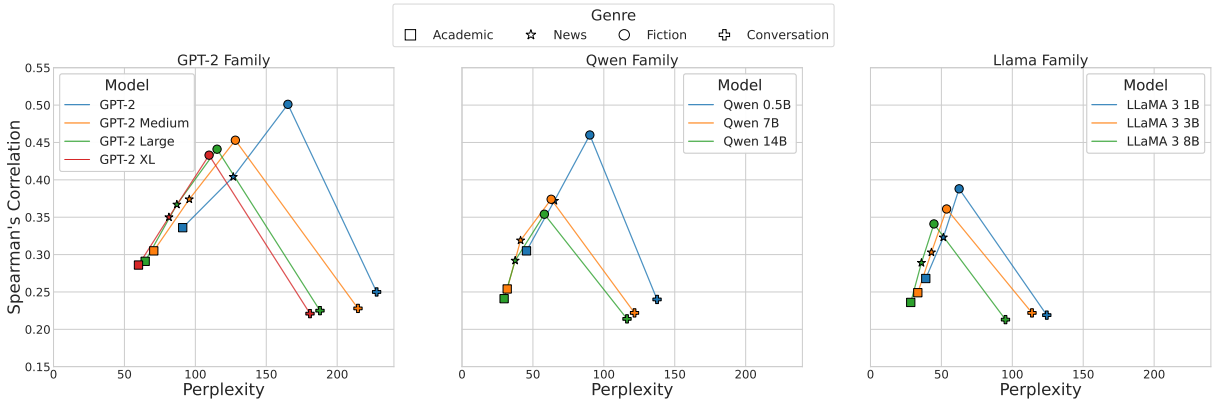

Figure 2: Spearman's correlations (surprisal - novelty in VUA-METANOV) as a function of perplexity, split by genre.

Model Variant	F	N	A	C
GPT-2-base (137M)	**.501**	.404	.336	.250
GPT-2-medium (380M)	.453	.374	.305	.228
GPT-2-large (812M)	.441	.367	.291	.225
GPT-2-xlarge (1.6B)	.433	.350	.286	.221
Qwen2.5 (0.5B)	.460	.372	.305	.240
Qwen2.5 (0.5B) It.	.445	.366	.299	.225
Qwen2.5 (7B)	.374	.319	.254	.222
Qwen2.5 (7B) It.	.348	.299	.225	.208
Qwen2.5 (14B)	.354	.292	.241	.214
Qwen2.5 (14B) It.	.277	.247	.211	.161
Llama3.2 (1B)	.388	.323	.268	.219
Llama3.2 (1B) It.	.393	.332	.267	.225
Llama3.2 (3B)	.361	.303	.249	.222
Llama3.2 (3B) It.	.369	.316	.251	.227
Llama3.1 (8B)	.341	.289	.236	.213
Llama3.1 (8B) It.	.349	.297	.230	.213

Table 3: Spearman's Correlation by genre on VUA-METANOV: Fiction (F), News (N), Academic (A) and Conversation (C). All results are statistically significant at the 0.01 level. It. refers to Instruction-tuned models.

Model Variant	Rank-Biserial r_b	AUC
GPT-2-base (137M)	.302	.651
GPT-2-medium (380M)	.392	.696
GPT-2-large (812M)	.421	.711
GPT-2-xlarge (1.6B)	.442	.721
Qwen2.5 (0.5B)	.385	.693
Qwen2.5 (0.5B) It.	.386	.693
Qwen2.5 (7B)	.507	.753
Qwen2.5 (7B) It.	.414	.707
Qwen2.5 (14B)	**.508**	**.754**
Qwen2.5 (14B) It.	.481	.741
Llama3.2 (1B)	.456	.728
Llama3.2 (1B) It.	.466	.733
Llama3.2 (3B)	.463	.731
Llama3.2 (3B) It.	.402	.701
Llama3.1 (8B)	.492	.746
Llama3.1 (8B) It.	.472	.736

Table 4: Correlation estimates between surprisal and novelty annotations in LAI2009-METANOV. All of the results in this table are statistically significant at the 0.01 level. It. refers to Instruction-tuned models.

5.2 Results on the synthetic datasets

Tables 4 and 9 report correlations between LLMs' surprisal and binary novelty annotations in LAI2009-METANOV and GPT4o-METANOV.

Similar patterns We observe similar ranges of correlation strengths ($r_b(.3-.6), AUC(.6-.8)$) in both datasets, compared to the results on binary annotations in VUA-METANOV. Generally, GPT-2-medium produces the strongest correlations on GPT4o-METANOV, yielding the strongest coefficients across all experiments in this study ($r_b = .578, AUC = .789$). On LAI2009-METANOV, Larget Qwen variant (14B) achieves the best correlations ($r_b = .508, AUC = .754$). Similar to VUA-METANOV, all correlation estimates for causal LMs are statistically significant.

Perplexity and correlation strength A striking finding in the LAI2009-METANOV and GPT4o-METANOV results is that model size correlates positively with correlation strength, which is the opposite of the pattern seen in VUA-METANOV. In both synthetic datasets, correlation strength increases as model size goes up. This trend is perfectly consistent across all model variants in LAI2009-METANOV (and is entirely reversed relative to VUA-METANOV). GPT4o-METANOV shows the same pattern, with a minor deviation within the GPT-2 family. We also measured perplexity for all model variants on both datasets: as model size increases, perplexity decreases, as expected. This contrast further reinforces the inverted pattern from VUA-METANOV, where higher perplexity was associated with stronger correlations (up to a point, after which it dropped).

6 Discussion

Is surprisal a good metric for metaphor novelty? Our experiments reveal a significant but moderate correlation ($\rho = .452$) between GPT-2's surprisal values and human annotations of metaphor novelty. This result is reinforced by comparable significant correlations observed across three additional datasets, each featuring distinct setups for novelty annotation. These findings extend prior work linking LLM-based surprisal to human processing difficulty, including reading times (Oh and Schuler, 2023a), eye-tracking data (Salicchi et al., 2023; Klein et al., 2025), and minimal-pair acceptability judgments (Hu et al., 2024; Tjuatja et al., 2025). However, compared to these more generic processing difficulty and acceptability effects, metaphor novelty seems to pose a substantially greater challenge for surprisal-based measurements. We believe that this may be due to deep semantic processing involved in metaphor interpretation. We hypothesise that the linguistic creativity of novel metaphors in particular requires measures beyond predictability and tailored to capturing novel interpretations and mappings between domains. Future work could enhance surprisal by incorporating semantic features, as suggested by (Salicchi et al., 2023; Meister et al., 2024).

Do large models weaken the correlation—or not? The effect of model size observed in the VUA-METANOV results aligns with prior research suggesting that as model size increases (or perplexity decreases), the correlation between surprisal and human-like behavioural measures tends to decline (Oh and Schuler, 2023a,b; Oh et al., 2024; Kuribayashi et al., 2024). However, we observe the opposite pattern in LAI2009-METANOV and GPT4o-METANOV. This inverse trend has also been reported in other empirical studies (Goodkind and Bicknell, 2018; Hao et al., 2020; Wilcox et al., 2020, 2023). We attribute these contrasting patterns to fundamental differences in datasets and annotations. For instance, Oh et al. (2024) suggest that larger models diverge from human-like expectations because they develop superhumanly complex associations between word frequencies, facilitated by their vast training data and high parameter capacity. Other work has highlighted their emergent capabilities (Hagendorff et al., 2023) and their super-human performance of LLMs in solving constructed, cognitive, e.g., vignette-based tasks (Binz and Schulz, 2023). Reflecting this reasoning in our context, we suspect that the synthetic datasets may resemble such artificial cognitive tasks that particularly large LMs have been found to excel at. The corpus-based data and their crowd-sourced novelty annotations, on the other hand, may reflect a more natural reading scenario where small foundation models seem to outperform LLMs in accounting for effects of predictability. This is underscored by our results on genre divergences, suggesting that the level of perplexity an LM reaches on a certain text type determines its ability to account for human surprisal. These results raise some important questions for future work on metaphor creativity, calling for a diversity of approaches that combine corpus-based and experimental settings.

Model Architecture and Alignment Effects Across all experiments, surprisals from causal LLMs consistently outperform those from MLMs in predicting metaphor novelty, with the latter showing weak or non-significant correlations. This supports the preference for causal architectures in surprisal-based analyses. Moreover, 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.

Cloze-surprisal strengthens correlation Our cloze-surprisal approach improves correlation across many model variants. In examples from Table 6, cloze-surprisal often raises the surprisal of the metaphorical word, consistent with the intuition that the full context is needed to process the metaphor. 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.

7 Conclusion

We have studied the distinction between conventional and novel metaphors and systematically investigated surprisal computed with LLMs 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 several 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. This opacity introduces the possibility of data leakage, especially for widely used corpora such as the BNC, which is the source of VUA-METANOV. Finally, although we carefully analyse effects of model architecture, size, and domain (genre), future work could adopt mixed-effects models to test the interaction of these variables.

Acknowledgments

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.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

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

Razvan C. Bunescu and Oseremen O. Uduehi. 2022. Distribution-based measures of surprise for creative language: Experiments with humor and metaphor. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 68–78, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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

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*.

Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie CY Chan, Andrew Lampinen, Jane X Wang, Zeynep Akata, and Eric Schulz. 2023. Machine psychology. *arXiv preprint arXiv:2303.13988*.

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*.

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*,

pages 75–86, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

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.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

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.

Keren Gruteke Klein, Shachar Frenkel, Omer Shubi, and Yevgeni Berzak. 2025. Eye tracking based cognitive evaluation of automatic readability assessment measures. *Preprint*, arXiv:2502.11150.

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.

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, Mario Giulianelli, and Tiago Pimentel. 2024. Towards a similarity-adjusted surprisal theory. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16485–16498, Miami, Florida, USA. Association for 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, 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.

Byung-Doh Oh, Hongao Zhu, and William Schuler. 2025. The inverse scaling effect of pre-trained language model surprisal is not due to data leakage.

In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1820–1827, Vienna, Austria. 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).

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.

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.

Lavinia Salicchi, Emmanuele Chersoni, and Alessandro Lenci. 2023. A study on surprisal and semantic relatedness for eye-tracking data prediction. *Frontiers in Psychology*, 14:1112365.

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.

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.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. 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.

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*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Details of the datasets

A.1 Statistics and Examples

In Table 5, we describe in numbers the three datasets under study. In Table 6, we list 16 examples from the VUA-METANOV dataset, 6 examples from the LAI2009-METANOV dataset, and 4 examples from the GPT4o-METANOV dataset.

A.2 GPT4o-METANOV Construction

To generate the sentences in GPT4o-METANOV, we prompt GPT4o 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.`"

B MLM Results and GPT4o-METANOV

B.1 MLM Results

In Table 7, we report the results of MLM models on VUA-METANOV. And in Table 8, we report the results of MLM models on LAI2009-METANOV.

B.2 GPT4o-METANOV Results

In Table 9, we report the results of both causal and MLM models on VUA-METANOV.

C Visualisations

C.1 Genres

In Figure 3, we plot the correlation of surprisal with metaphor novelty scores of VUA-METANOV on genre splits.

C.2 Distributions

In Figure 4, we plot the distributions of novelty scores and surprisal of the VUA-METANOV.

VUA-METANOV					LAI2009-METANOV			GPT4o-METANOV	
Genre	# Met.	Score$_{nov}$ mean | std.	# Nov. $>= 0.5$	L$_{sent}$ mean | std.	Class	#	L$_{sent}$ mean | std.	#	L$_{sent}$ mean | std.
Fiction	3170	-.005 | .271	94	26.0 | 16.5	Conventional	104	6.7 | 1.5	100	10.1 | 1.9
News	4712	.000 | .257	132	29.9 | 14.2	Novel	104	6.7 | 1.4	100	12.7 | 2.3
Academ.	5499	.003 | .239	102	34.9 | 16.0	All	208	6.7 | 1.5	200	11.4 | 2.5
Conver.	1774	-.000 | .236	25	17.5 | 15.9					
All	15155	.000 | .251	353	29.4 | 16.5	per class: *21 N, 42 V, 41 Adj.*			per class: *50 N, 50 V*	

Table 5: Distributions and statistics of the datasets under study. # Met. is the number of metaphor words, Score$_{nov}$ is the BWS novelty scores, # Nov. is the number of novel metaphors (Score$_{nov}$ >= 0.5). L$_{sent}$ is the length of sentences in words. We report the number of metaphors used in the sense of Nouns (*N*), Verbs (*V*) or Adjectives (*Adj.*) in the synthetic datasets.

VUA-METANOV					
Class	Conventional	Novel	Score	Surprisal	Cloze
Fiction	**1.** ' Tell him I am very sorry, but I must **fill** the quota. '	**2.** Adam might have escaped the file memories for years, **suppressed** them and jerked violently <14> by those events.	c: -.441 **n: .531**	9.211 **13.26**	6.887 **14.74**
	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.	**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.	c: .278 **n: .588**	10.99 **7.426**	15.17 **9.109**
News	**5.** One Mr Clarke can not duck away from if he wants to **avoid** a second Winter of Discontent ..	**6.** This was conveniently **encapsulated** in the first try.	c: -.094 **n: .500**	3.969 **8.586**	7.344 **14.35**
	7. Thrusts of resistance (mass demonstrations, resignations, tax rebellions, etc) would come in **crests**.	**8.** Travel: A **pilgrimage** sans progress Elisabeth de Stroumillo potters round Poitou	c: .382 **n: .514**	10.97 **8.563**	13.66 **15.31**
Academic	**9.** The Tehuana dress is by no means the most decorative variant or the **closest** to pre-Hispanic forms of clothing.	**10.** Interwoven with these images are subtler references to the **metaphorical** borderlines which separate Latin American <5> and North America.	c: -.194 **n: .529**	6.457 **9.117**	11.45 **12.78**
	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.	**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.	c: .294 **n: .514**	9.961 **5.648**	8.109 **9.469**
Conversat.	**13.** Me dad said he's had enough Well, we were debating whether to **give** it to you or not.	**14. Struggled** with it a little	c: -.633 **n: .552**	3.934 **17.02**	7.867 **12.31**
	15. That's an old **trick**.	**16.** Can you **sort** erm, madame out?	c: .310 **n: .567**	4.063 **8.734**	12.34 **10.13**
LAI2009-METANOV					
sank (V)	**17.** Upon hearing the news my spirits **sank**	**18.** Upon having the data my prediction **sank**	–	3.900 **10.10**	–
battle (N)	**19.** Those chess players are prepared for **battle**	**20.** Those plastic surgeons are prepared for **battle**	–	4.238 **5.758**	–
fragile (Adj)	**21.** His mental condition remains **fragile**	**22.** His website popularity remains **fragile**	–	6.508 **9.023**	–
GPT4o-METANOV					
anchor (N)	**23.** Her family was her emotional **anchor** during the crisis.	**24.** The smell of coffee became an **anchor** to mornings that no longer came.	–	3.615 **7.488**	–
navigate (V)	**25.** The software helps users **navigate** complex legal documents.	**26.** He tried to **navigate** the silence like a sailor without stars.	–	3.941 **8.391**	–

Table 6: Examples from each dataset. The metaphor word is in **boldface** within texts. 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>. Score is the novelty score, Surprisal is the GPT2-Base recorded surprisal, and Cloze is the GPT2-Base cloze-surprisal. In those three columns, values on top correspond to the conventional example, and on bottom to the novel example.

Model Variant	Continuous Novelty		Binary Novelty	
	Pearson r direct | cloze	Spearman ρ direct | cloze	Rank-Biserial r_b direct | cloze	AUC direct | cloze
Masked Language Models				
BERT-base-cased-110M	.243	.239	.056*	.479
BERT-large-cased-335M	.227	.218	.080*	.472
RoBERTa-base-125M	.259	.267	.233	.533
RoBERTa-large-355M	.238	.244	.168	.514
DeBERTa-v3-Xsmall-70M	.113	.105	.273	.584
DeBERTa-v3-Small-142M	.012*	.006*	.192*	.559
DeBERTa-v3-Base-184M	-.013*	-.009*	.075*	.530
DeBERTa-v3-Large-435M	-.075*	-.078*	.002*	.496
ModernBERT-Base-149M	.247	.241	.148*	.504
ModernBERT-Large-395M	.222	.208	.085*	.481

Table 7: Correlation estimates between surprisal and continuous/binary novelty scores in VUA-METANOV for MLM models. Estimates that are not statistically significant are marked by an asterisk (*)

Model Variant	Rank-Biserial r_b	AUC
Masked Language Models		
BERT-base-cased-110M	.066*	.533
BERT-large-cased-335M	.063*	.532
RoBERTa-base-125M	.245	.623
RoBERTa-large-355M	.260	.630
DeBERTa-v3-Xsmall-70M	-.002*	.499
DeBERTa-v3-Small-142M	-.031*	.484
DeBERTa-v3-Base-184M	.054*	.527
DeBERTa-v3-Large-435M	-.011*	.495
ModernBERT-Base-149M	.203	.602
ModernBERT-Large-395M	.263	.632

Table 8: Correlation estimates between surprisal and binary novelty scores in LAI2009-METANOV for MLM models. Estimates that are not statistically significant are marked by an asterisk (*)

Model Variant	Rank-Biserial r_b	AUC
Causal Language Models		
GPT-2-base (137M)	.514	.757
GPT-2-medium (380M)	**.578**	**.789**
GPT-2-large (812M)	.572	.786
GPT-2-xlarge (1.6B)	.554	.777
Qwen2.5 (0.5B)	.386	.693
Qwen2.5 (0.5B) It.	.418	.709
Qwen2.5 (7B)	.469	.735
Qwen2.5 (7B) It.	.374	.687
Qwen2.5 (14B)	.505	.752
Qwen2.5 (14B) It.	.399	.700
Llama3.2 (1B)	.485	.742
Llama3.2 (1B) It.	.440	.720
Llama3.2 (3B)	.488	.744
Llama3.2 (3B) It.	.440	.720
Llama3.1 (8B)	.529	.765
Llama3.1 (8B) It.	.487	.743
Masked Language Models		
BERT-base-cased-110M	.291	.646
BERT-large-cased-335M	.344	.672
RoBERTa-base-125M	.347	.674
RoBERTa-large-355M	.331	.666
DeBERTa-v3-Xsmall-70M	-.049*	.476
DeBERTa-v3-Small-142M	-.078*	.461
DeBERTa-v3-Base-184M	-.045*	.478
DeBERTa-v3-Large-435M	.024*	.512
ModernBERT-Base-149M	.386	.693
ModernBERT-Large-395M	.355	.678

Table 9: Correlation estimates between surprisal and novelty annotations of GPT4o-METANOV. All of the results in this table are statistically significant at the 0.01 level. It. refers to Instruction-tuned models.

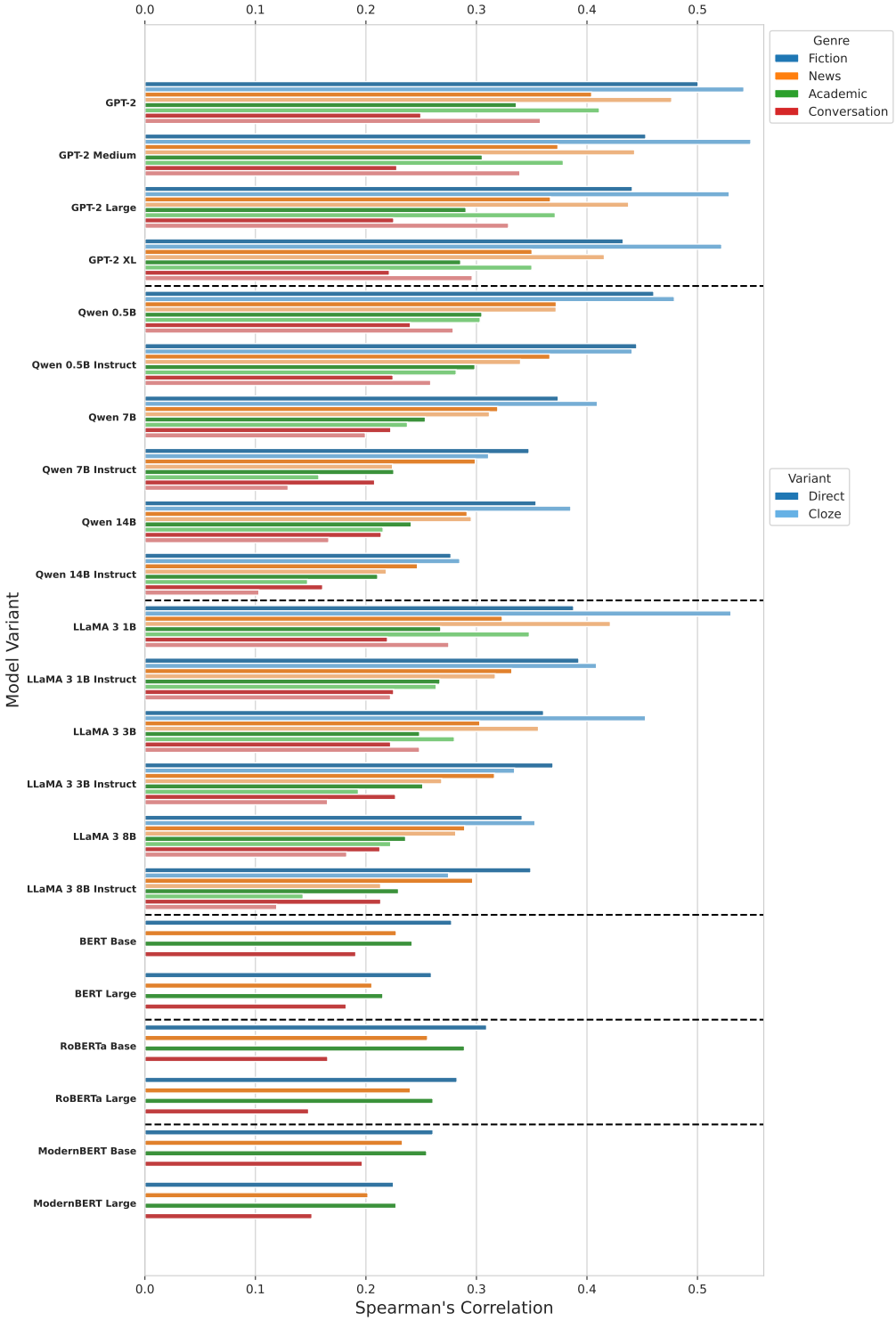

Figure 3: Spearman Correlation by model variants and genre splits

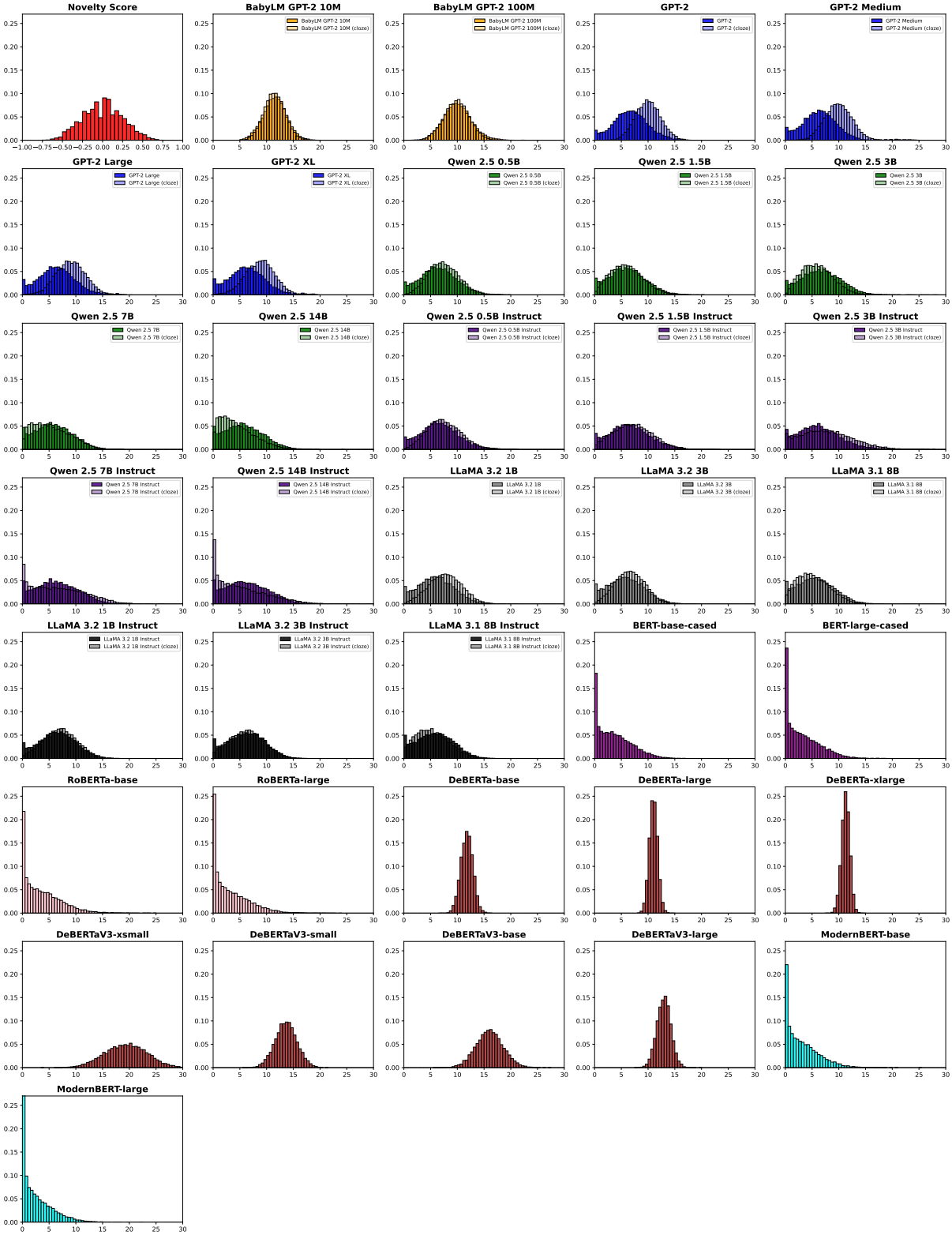

Figure 4: Distribution of metaphor novelty, and surprisals across 35 language models