DEATH OF THE NOVEL(TY): BEYOND n-GRAM NOVELTY AS A METRIC FOR TEXTUAL CREATIVITY

Arkadiy Saakyan Columbia University Najoung Kim Boston University

Smaranda Muresan Columbia University **Tuhin Chakrabarty** Stony Brook University

ABSTRACT

N-gram novelty is widely used to evaluate language models' ability to generate text outside of their training data. More recently, it has also been adopted as a metric for measuring textual creativity. However, theoretical work on creativity suggests that this approach may be inadequate, as it does not account for creativity's dual nature: novelty (how original the text is) and appropriateness (how sensical and pragmatic it is). We investigate the relationship between this notion of creativity and n-gram novelty through 7542 expert writer annotations (n = 26) of novelty, pragmaticality, and sensicality via *close reading* of human and AI-generated text. We find that while n-gram novelty is positively associated with expert writer-judged creativity, $\approx 91\%$ of top-quartile expressions by n-gram novelty are not judged as creative, cautioning against relying on n-gram novelty alone. Furthermore, unlike human-written text, higher n-gram novelty in open-source LLMs correlates with lower pragmaticality. In an exploratory study with frontier close-source models, we additionally confirm that they are less likely to produce creative expressions than humans. Using our dataset, we test whether zero-shot, few-shot, and finetuned models are able to identify creative expressions (a positive aspect of writing) and non-pragmatic ones (a negative aspect). Overall, frontier LLMs exhibit performance much higher than random but leave room for improvement, especially struggling to identify non-pragmatic expressions. We further find that LLM-as-a-Judge novelty scores from the best-performing model were predictive of expert writer preferences. 1

1 Introduction

Advances in large language models (LLMs) have led to their widespread applications in writing. In fact, recent studies (Handa et al., 2025) (Anthropic) and (Chatterji et al., 2025) (OpenAI) show that writing assistance remains one of the main use cases of LLMs. At the same time, researchers have been raising concerns about how writing assistance tools can reduce collective human creativity via homogenization effects (Doshi & Hauser, 2024; Kobak et al., 2025; Zhang et al., 2025), proliferation of AI slop (Chakrabarty et al., 2025a; Shaib et al., 2025) or copying from training data (McCoy et al., 2023).

Such challenges lead to a growing need for robust textual creativity evaluation. Recently, tools like WiMBD (Elazar et al., 2024), Rusty-DAWG (Merrill et al., 2024), infini-gram(-mini) (Liu et al., 2024; Xu et al., 2025) have been developed to efficiently search LLMs' pretraining corpora and evaluate the novelty of their generations. Building on these tools, Lu et al. (2025) introduced a metric called CREATIVITY INDEX, which places significant weight on n-gram novelty – lack of occurrence of textual fragments in some large (several trillion tokens) corpora – for measuring creativity of text. However, literature on the psychology of creativity would consider such approach as not fully adequate: based on the widely adopted definition of

¹We will release the collected dataset and models on github.com/asaakyan/ngram-creativity.

creativity, novelty is a necessary but not sufficient criterion (Sawyer & Henriksen, 2024; Runco & Jaeger, 2012; Csikszentmihalyi et al., 1997; Amabile, 1983; Jackson & Messick, 1965).

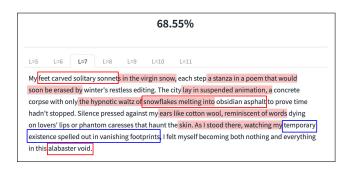


Figure 1: High Creativity Index (68.5%) of an excerpt with many n-gram novel (text not highlighted in red) yet non-sensical (red bounding box) and non-pragmatic (blue bounding box) expressions.

While prior work (Chakrabarty et al., 2024b) has relied on expert human evaluation for measuring creativity of long form text via fine-grained rubrics, LLMs struggle to imitate expert judgments, thereby making it difficult to scale and automate such an approach. To further our understanding of textual creativity, we operationalize it according to the standard definition of creativity (Runco & Jaeger, 2012) through novelty and appropriateness components. We further decompose appropriateness into sensicality (making sense on its own) and pragmaticality (making sense in context). An expression generated by an LLM may be novel with respect to the pretraining corpora but make little sense in the broader context

of the passage, precluding it from being judged as creative. For instance, expressions in bounding boxes in Figure 1 are novel with respect to pretraining but add no meaning due to being non-sensical ("feet carved solitary sonnets") and non-pragmatic ("temporary existence spelled out in vanishing footprints"). At the same time, an expression may not be n-gram novel with respect to pretraining but still be creative. For instance, That's the bottom of the heart, where blood gathers. has very low n-gram novelty, but was rated as creative in the context of the given passage by an expert writer because of how it comes across as emotionally foreshadowing to the reader (see Table 1).

To tackle these issues, we conduct a study on the relation between n-gram novelty – low occurrence in trillion token-level corpora – and creativity by asking expert writers to $close\ read\ AI$ - and human-written passages. Mixed-effects regression model analysis accounting for individual rater variation and other confounding factors revealed a negative relationship between n-gram novelty and pragmaticality in open-source models, with 91% of highly n-gram novel expressions not judged as creative. This cautions against relying on n-gram novelty alone to evaluate creativity. To understand whether LLM-as-a-Judge frameworks could be used instead, we utilize the collected dataset to evaluate how well frontier LLMs and finetuned models can emulate expert human judgments on creativity and pragmaticality of text. We further validate our best performing model on out-of-distribution data and find a strong alignment between the novelty scores generated by the LLM-as-a-Judge model and expert preferences on writing quality.

2 Data

OPERATIONALIZING CREATIVITY We operationalize creativity through human-judged novelty and appropriateness components (Runco & Jaeger, 2012). We decompose appropriateness into two subcomponents: *sensicality* and *pragmaticality*. Sensicality is whether the expression makes sense standalone (for example, an expression like "*he tended at cloud finger*" does not make sense by itself, or in other words, semantically infelicitous). Pragmaticality refers to the expression's fit within the context of the passage. Pragmaticality violation can encompass a range of errors, including logical incoherence (e.g., "*Alice felt*").

great" immediately followed by "Despite being sad, Alice..."), or sounding awkward or odd in that context.² Creative expressions are those that are simultaneously judged by a human as sensical, pragmatic, and novel. Text that is creative in this sense may not necessarily be n-gram novel (see Table 1).

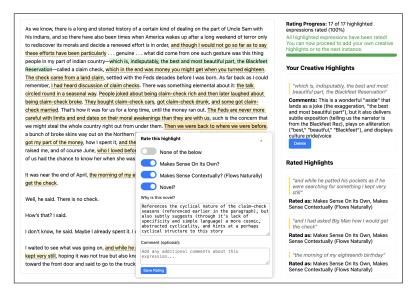


Figure 2: Example of the annotation interface and an expert writer's annotation.

SURVEY INSTRUMENT TO obtain the annotations, we took inspiration from the practice of *close reading* (Smith, 2016), a literary analysis technique of carefully analyzing small parts of the passage to make sense of it as a whole, "putting the author's choices under a microscope" (Brummett, 2018). We first split the passage by punctuation into "atomic" expression units, and then sample roughly 50% of them for annotation based on the percent of novel n-grams within the expression (see Appendix A.1). These expressions are pre-highlighted. Then, we ask the annotators to rate each of the pre-highlighted

expressions in the passage for their sensicality, pragmaticality, and novelty (see the user interface in Figure 2). In addition, the interface allows experts to highlight any creative expressions (which satisfy the sensicality, pragmaticality and novelty requirements) that were not already pre-highlighted. The annotators were required to provide rationales for why the expression was creative, and they could also leave an optional comment on other aspects. We provide detailed instruction and examples to the annotators with simplified terms for sensicality ("Makes sense on its own?") and pragmaticality ("Makes sense contextually? (Flows naturally)")—see Appendix A.3. An example of an annotation is shown in Figure 2, where the pop-up window displays the annotation for a pre-highlighted expression, and in green the annotator has highlighted a new creative expression.

RECRUITMENT We recruit professional writers (n=26) via listservs of top Master of Fine Arts writing programs in the US, as well as published writers with a Masters in English Writing or Literature through the Upwork platform. Each annotator first underwent training on at least one passage during which any instruction misunderstanding were corrected. Annotators were paid \approx USD \$100 per annotated batch of 10 passages, which took approximately 2–3 hours.

ANNOTATION For human text samples, we collect 50 passages of roughly 400 words each from stories published in the New Yorker.³ For LLM-generated samples, we rely on fully open-source (open weights and open training data) LLMs: OLMo (7B) (Groeneveld et al., 2024) and OLMo-2 (32B) (Team OLMo et al., 2025), considering both the older and the improved, larger versions to understand whether progress

²Pragmaticality is a more stringent requirement than sensicality (an expression cannot be pragmatic if nonsensical). We make this distinction to analyze more nuanced effects of context rather than atomic expression-level nonsensicality. Furthermore, our annotations suggest that pragmaticality is a precondition for novelty; we discuss this later.

³One the most prestigious literary magazines for publishing fiction.

in language modeling and benchmark performance led to improvements in writing quality. The passages were chosen such that they are not present in the corresponding open-source LLM training corpora (see Appendix A.2). We assign each model 25 passages from the human sample (*seed passages*), and prompt them to generate a similar length passage based on the content obtained from the summary of the human text (see details in Appendix C.1). To ensure reasonable completion time for each annotator task, we divide the total 100 passages into 10 batches of 10 passages each, containing 5 human and 5 AI passages. 3 distinct annotators are assigned for each batch. To gauge the ceiling LLM performance, we conduct a smaller-scale follow-up study⁴ on 2 frontier models: GPT-5 and Claude 4.1, each generating 5 passages to make up one additional batch of 10 texts assigned to 4 randomly selected annotators from our pool.

STATISTICS Overall, 7542 pre-highlighted expression annotations were collected for 2514 unique expressions, with 427 unique expressions rated as novel by at least one annotator, 535 as non-pragmatic and 216 as non-sensical. Pragmaticality and sensicality were largely preconditions for novelty judgments, as only 3\% and 5\% of novel annotations were also marked as non-sensical and non-pragmatic, respectively. In addition to the pre-highlighted expression ratings, the annotators highlighted 226 new expressions as creative (satisfying the sensicality, pragmaticality, and novelty criteria at the same time). Due to low prevalence of expressions annotated as novel (7%) and non-pragmatic (9%), traditional agreement metrics would suffer from prevalence bias and hence would not be suitable (Eugenio & Glass, 2004; Brennan & Silman, 1992; Feinstein & Cicchetti, 1990). Hence, we report the Free-Marginal Multirater Kappa κ_{free} (Randolph, 2005) using the statsmodels package (Skipper & Josef, 2010). For novelty, mean κ_{free} across the 10 batches is 0.78 with standard deviation at 0.11, while for pragmaticality it is 0.72 with a 0.12 standard deviation.

3 **METHODS**

MEASURING *n*-**GRAM NOVELTY** We use the infinigram package (Liu et al., 2024) to estimate *n*-gram novelty with respect to LLM training corpora. Specifically, we use ∞ -probability, which allows to assign a probability to any expression using backoff to the longest expression present in the corpus \mathcal{D} :

$$P_{\infty}(w_i \mid w_{1:i-1}) = \frac{\text{cnt}(w_{i-(n-1):i-1}w_i \mid \mathcal{D})}{\text{cnt}(w_{i-(n-1):i-1} \mid \mathcal{D})}, \quad \text{where } n = \max n' \in [1, i] \mid \text{cnt}(w_{i-(n'-1):i-1} \mid \mathcal{D}) > 0$$

We can then use the consecutive probabilities to compute the perplexity of an expression as a proxy for n-gram novelty.⁵ We use the respective OLMo and OLMo-2 training corpora⁶ as the reference \mathcal{D} .

MODELING Annotation of textual creativity is inherently subjective. To account for annotator variation as well as other confounding factors (topic of the passage, generation model), we turn to hierarchical/multilevel modeling (Gelman & Hill, 2007) commonly used to handle nested data in behavioral and social sciences (Baayen et al., 2008; Yarkoni, 2022; Kaufmann et al., 2025) and account for confounders in language model evaluation (Lampinen et al., 2022). Since our target variables are binary judgments of sensicality, pragmaticality, and novelty, we fit multilevel logistic regression models. For perplexity, we log-standardize the variable to reduce the skew and for easier interpretation. For instance, to model the relationship between n-gram novelty (measured by perplexity) and creativity, we fit the following model for the probability of the expression being labeled as creative (i.e. sensical, pragmatic, and novel simultaneously):

⁴Our main study focuses on LLMs for which pretraining corpora is known and hence the only models that allow accurate n-gram novelty estimation.

⁵Defined in a standard way as $\left(\prod_{i=1}^N P_\infty(w_i|w_{i-1})\right)^{-\frac{1}{N}}$ ⁶Dolma-v1.7 (2.6T tokens) for OLMo and v4_olmo-2-0325-32b-instruct (4.2T) tokens for OLMo-2

$$\operatorname{logit} \left(\mathbb{P} \left[\operatorname{creative}_i = 1 \right] \right) = \beta_0 + \beta_1 \cdot \operatorname{ppl-log_std}_i + u_{g[i]}^{(\operatorname{gen.src})} + u_{a[i]}^{(\operatorname{annot})} + u_{p[i]}^{(\operatorname{para})} + u_{p[i]}^{(\operatorname{para})} + u_{a[i]}^{(\operatorname{para})} + u_{a[i]}$$

where g[i], a[i], p[i] are the indices of the generation source (human, OLMo, OLMo-2), annotator, and seed passage id of expression i. At the generation-source level, we represent variation in the baseline creativity prevalence across models by the varying intercept $u_{g[i]}^{(\mathrm{gen.src})}$. Variation on each annotator's baseline threshold for calling something creative is represented by the varying intercept $u_{a[i]}^{(\mathrm{annot})}$, while variation on how often expressions from different passages are judged creative is represented by the varying intercept $u_{p[i]}^{(\mathrm{para})}$. Group-level coefficients are distributed normally with respective group variances and estimated with the 1 me 4 R package (Bates et al., 2015). See detailed specifications in Appendix B.

4 RESULTS

n-Gram novelty Predicts creativity But is not a reliable metric Although standardized log perplexity is significantly associated with creativity (OR (Odds Ratio) ≈ 1.95 per SD (Standard Deviation), p < 0.001), approximately 91% of top-quartile n-gram novel expressions are not judged as creative (sensical, pragmatic, and novel at the same time) by any of the annotators. Further, a substantial portion of creative expressions have very low perplexity: approximately 24% of creative expressions fall below the mean perplexity, and 7% are in the lowest quartile. This demonstrates that while high perplexity is predictive of creativity, a non-negligible fraction of creative expressions are not n-gram novel, and the vast majority of highly n-gram novel expressions are not creative, cautioning against relying solely on n-gram novelty to capture creativity. In Table 1 we provide examples of expressions that were judged as creative by the annotators along with their justifications, but have a very low log-standardized perplexity (among the lowest in the dataset). Observing the annotator justifications, it is possible that contextual reasons rather than solely n-gram novelty make an expression creative (more examples in Appendix A.4).

n-GRAM NOVELTY NEGATIVELY IMPACTS PRAGMATICALITY IN OPEN-SOURCE LLMs We fit a logistic regression model on whether an expression was rated as pragmatic using standardized log perplexity, generation source, and their interaction as predictors, with varying intercepts for annotators and seed passage ids. Instances that were labeled as not sensical were removed to focus on cases where expressions do not make sense *in the context*, rather than simply not felicitous standalone.

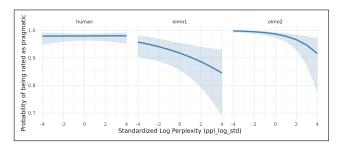


Figure 3: Predicted probability of being rated as pragmatic for different values of log-std perplexity, by generation source. Bands indicate 95% CIs. Annotator and paragraph intercepts correspond to population-level fixed effects.

Figure 3 demonstrates how *n*-gram novelty affects the probability of an expression being rated as pragmatic according to our model. We find negative slopes for both OLMo and OLMo-2 generated text, while we see no such effect for human-written text. Linear hypothesis tests show no evidence that n-gram novelty affects pragmaticality for human-written text $(\beta = 0.01, SE = 0.09, p = 0.94).$ In contrast, higher n-gram novelty negatively affects pragmaticality of AI-generated text: OLMo (β = $-0.18, \chi^2(1) = 4.96, p = 0.026$ and OLMo-2 ($\beta = -0.48, \chi^2(1) =$ 13.89, p < 0.001).

This indicates that as open-source LLMs try to generate more n-gram novel text, they are less likely to generate expressions that make sense contextually. In Table 1, we show examples of highly n-gram novel expressions that were rated as non-pragmatic.

Expression	Justification	PPL	Source
	— Low n -gram Novelty but Creative —		
antique thirty-foot oak beams now sold for three thousand dollars.	The novelty in this sentence comes from the rhythmic usage of numbers; "single" and "thirty-foot" along with "three thousand."	-2.6	human
knowing that he wouldn't.	I found this to be a creative expression because it offers insight into Tongsu's mind and surprises the reader with its blunt honesty.	-2.0	human
That's the bottom of the heart, where blood gathers.	The metaphor is vivid and physiological, merging the act of writing with the body in a tactile, somber way. It's also emotionally foreshadowing.	-1.5	human
	— High n-gram Novelty but Non-Pragmatic —		
and said the morning blessings in a whisper that embarrassed the chairs		2.37	GPT-5
the child who was at once a present fundamental fact but	In isolation and in context, this is a misadventure of poor word choice, punctuation, and structure.	2.36	human
each word spoken a thread in the slow weaving of trust	[] saying the act of weaving is made up of physical threads is mixing metaphors;[]	2.09	OLMo-2
tape suspended between her fingers like a frozen gesture		1.40	Claude-4.1

Table 1: Top: expressions judged creative despite low n-gram novelty (negative log-standardized perplexity). Bottom: non-pragmatic expressions with high n-gram novelty (positive log-standardized perplexity).

AN EXPLORATORY STUDY WITH FRONTIER CLOSED MODELS We conduct an exploratory follow-up study with two frontier models: GPT-5 and Claude-4.1 (see details in Appendices C.2, B.3).

Contrast	Odds Ratio(OR)	95% CI	p-value
Claude-4.1 / Human	0.515	[0.283, 0.937]	0.024
GPT-5 / Human	0.445	[0.235, 0.842]	0.007
OLMo / Human	0.525	[0.391, 0.705]	<0.001
OLMo-2 / Human	0.624	[0.466, 0.835]	< 0.001

Table 2: Estimated marginal means (EMM) contrasts comparing creativity of AI and human expressions. OR < 1 indicate model expressions were less likely to be judged creative compared to humans.

Expanding on the original dataset, we compare the probability of expression creativity across generation sources (human, OLMo, OLMo-2, GPT-5, Claude-4.1) by fitting a logistic regression with generation source as the predictor, including random intercepts for annotators and seed passages. Table 2 shows that the probability of an expression to be judged creative is significantly higher for humans compared to LLMs. Although not statistically significant (likely due to

small sample of the exploratory study and imprecise estimation of n-gram novelty as we have no access to the models' pretraining data), we observe a similar trend of a negative impact of n-gram perplexity when comparing frontier model and human expressions: the interaction terms between perplexity and generation source were negative for both Claude ($\beta = -0.33, p = 0.33$) and GPT-5 ($\beta = -0.28, p = 0.26$), but

slightly positive for the human baseline ($\beta = 0.11, p = 0.54$). Future work could utilize our close reading annotation scheme to conduct a higher-powered study.

LACK OF EVIDENCE THAT AI LIKELIHOOD PREDICTS NON-CREATIVITY One plausible hypothesis is that if a text seems more AI-generated, it is judged by humans as less creative. We fit a logistic regression predicting the probability of the expression being judged creative (simultaneously novel, pragmatic, and sensical) or pragmatic given the AI likelihood scores from one of the leading AI-generated text detectors, Pangram (Emi & Spero, 2024). The detector perfectly classified AI texts assigning all of them a score of 100%; human texts were assigned varying small likelihoods of sounding AI (see Figure 8). Log-standardized AI likelihood scores were used as a passage-level predictor in the logistic regression model with varying intercepts for annotators and passages. We did not find evidence that passages with higher (or lower) AI likelihood scores systematically differed in the average novelty or pragmaticality of their expressions. A one-standard deviation increase in a passage's AI likelihood was associated with a small, non-significant increase in the odds that an expression was judged creative ($\beta = 0.06$, SE = 0.11, p = 0.602) and a non-statistically significant decrease in the probability that it was judged pragmatic ($\beta = -0.24$, SE = 0.14, p = 0.094).

Writing Quality Reward Models Predict Creativity and Pragmaticality Chakrabarty et al. (2025a) has recently explored training scalar reward models on a corpus, LAMP, that consists of <AI-generated, Expert-edited> pairs of text with implicit preference judgments (edited > original) (Chakrabarty et al., 2025b).Unlike for AI-likelihood scores, a logistic regression on the passage-level writing quality reward scores showed significant positive effect of higher reward model scores for both creativity ($\beta = 0.24$, SE = 0.04, OR = 1.27, p = < 0.001) and (separately) pragmaticality ($\beta = 0.28$, SE = 0.05, OR ≈ 1.32 p < 0.001). This indicates that existing reward models can be predictive of positive and negative aspects of writing. However, these models output a single passage-level score without any explanation. In the following section, we explore whether LLMs can also explain why a passage may be deemed creative or non-pragmatic.

5 Can LLMs replicate human close reading judgments of creativity?

To understand whether LLMs can be used as reward models for textual creativity, we evaluate their ability to find both the positive aspects of writing (human-judged novel expressions), as well as the negative ones (non-pragmatic expressions). We formulate a close reading task similar to the annotation setting using our dataset containing passages \mathcal{P} , as well as a set of ground-truth novel expressions (or ground-truth non-pragmatic expressions, respectively) $\mathcal N$ provided by 3 annotators. For prompts and hyperparameters, see Appendix C.

CLOSE READING TASK DEFINITION Given a passage, the model needs to extract novel (or non-pragmatic) expressions $\hat{\mathcal{N}}$. We consider expression \hat{n} to be an approximate match with expression n if one is a subset of the other or their Levenstein ratio (Levenshtein, 1966; Bachmann, 2025) is $\geq 90\%$. We compute the F1 score as follows: an expression \hat{n} is a true positive if it has an approximate match with some expression in \mathcal{N} , and a false positive otherwise. In addition, all n that do not have a match in $\hat{\mathcal{N}}$ are considered as false negatives.

LLM SETUP We test 3 frontier reasoning models: GPT-5, Claude 4.1, and Gemini 2.5 Pro. For the zero-shot setting, we provide a prompt analogous to the instructions we gave to the human annotators. For the few-shot setting, we additionally provide examples from 3 passages set aside for few-shot prompting (≈ 15 novel expression and non-pragmatic expression examples) chosen randomly among passages with a median number of novel or non-pragmatic annotations. For *finetuned model evaluation*, we finetune a set of smaller open-source models (OLMo2-Instruct 7B, Qwen3 8B (Yang et al., 2025), Llama-3.1-Instruct 8B (Meta, 2024)) using LoRA (Hu et al., 2022), as well as GPT-4.1 using full-parameter finetuning on 60% of our dataset (including the 3 passages used for few-shot examples), using the rest for evaluation.

RESULTS We plot the F1 scores with 95% confidence intervals for few-shot and finetuned models in Figure 4. There is little difference among models in each task, as well as little improvement from few-shot prompting. The non-pragmatic expression identification task appears to be significantly harder, with F1 scores below 20, whereas model performance on the novel expression task was above 40. This aligns with our finding earlier that LLMs tend to produce non-pragmatic expressions at high *n*-gram novelty level—this result suggests it may be stemming from failure to recognize non-pragmatic expressions. These findings also caution against using LLM-as-a-Judge for identification of writing issues given the lower performance on the non-pragmatic expression identification task. All finetuned models lag behind large reasoning models, suggesting that close reading task is very difficult even with task-specific adaptation. Overall, given the high number of expressions that the models could have chosen from, the performance is quite impressive: choosing at random, the model's precision would be less than 1%.

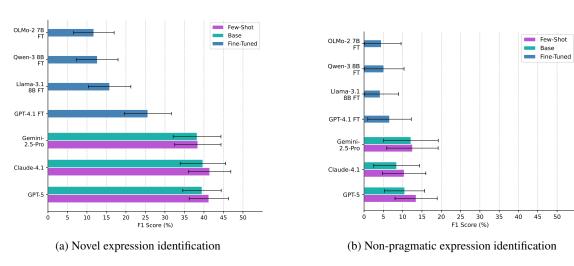


Figure 4: Comparison of finetuned, zero-shot and few-shot model performance across (a) novel and (b) non-pragmatic expression identification tasks (F1 scores with 95% CIs computed with confidenceinterval package (Gildenblat, 2023) using the Takahashi et al. (2022) method.

ALIGNMENT WITH EXPERT AND CROWD PREFERENCES We use one of the best performing models on our close reading tasks, few-shot GPT-5, to explore how well crowd preferences can approximate creativity judgments. Given a dataset of preferences (passageA, passageB, $\mathbb{1}[A \text{ preferred over B}])$, we obtain the number of novel (Nov_A, Nov_B) and non-pragmatic (Prag_A, Prag_B) expressions identified by few-shot GPT-5 for each passage. We normalize each score by the number of words⁷ in the passage and take the difference between them to obtain $\Delta \text{Nov}_{AB} = \frac{\text{Nov}_A}{\# \text{words in } A} - \frac{\text{Nov}_B}{\# \text{words in } B}$, $\Delta \text{Prag}_{AB} = \frac{\text{Prag}_A}{\# \text{words in } A} - \frac{\text{Prag}_B}{\# \text{words in } A}$

We then fit a hierarchical logistic regression model on the differences between the scores for every pair of passages to validate whether GPT-5 novelty and non-pragmaticality judgments align with expert preferences on writing. We use the Style Mimic dataset containing preferences of three expert writer annotators comparing MFA authors' imitations of famous authors against LLM-generated imitations (Chakrabarty et al., 2025a). We include random intercepts for each annotator, for seed famous authors, and for MFA authors nested within seed authors. Group-level coefficients are assumed normally distributed with their respective variances.

⁷Not the number of atomic expressions, since the model is allowed to choose expressions regardless of punctuation to mirror the creative highlight setting in the annotation.

The model showed that novelty score differences were a significant positive predictor of preference ($\beta_{\Delta \text{Nov}_{AB}} = 0.63$, SE= 0.26, OR = 1.88, p=0.014), while pragmaticality scores were not predictive ($\beta_{\Delta \text{Prag}_{AB}} = -0.05$, p=0.832), as expected due to weaker LLM-as-a-Judge performance on this task. We also find alignment between both pragmaticality and novelty scores in crowd preferences for writing from a dataset based on LMArena evaluations (see Appendix D). We also find evidence that novelty score differences are more predictive of expert preferences than differences in Creativity Index in Appendix E.

6 RELATED WORK

McCoy et al. (2023) investigates LLM copying from training data, finding that GPT-2 tends to copy large chunks of training corpora and has issues with coherence at high n-gram novelty, aligning with some of our findings. Recent advances in training data exploration (Elazar et al., 2024; Merrill et al., 2024; Liu et al., 2024; Xu et al., 2025) have enabled textual creativity metrics based on n-gram novelty with respect to trillion-token sized corpora (Lu et al., 2025). Our study cautions against using n-gram novelty alone as a metric for creativity based on the standard definition of creativity in psychology requiring the text not only to be novel but also pragmatic. In addition, we make a distinction between human-judged novelty and n-gram novelty.

There has been many efforts in quantifying issues in AI writing. Chakrabarty et al. (2025b) consulted with writing experts to create taxonomy of idiosyncrasies in AI writing. They further collected span level edits from expert writers following the proposed taxonomy to improve AI writing in an automated pipeline. More recently Shaib et al. (2024) collected annotations from experts in terms what qualifies as "slop" in both human and AI text and found that it correlate with latent dimensions such as coherence and relevance. In a concurrent study Russell et al. (2025) found that people who use AI to write are better detectors of AI writing. Such users typically rely on specific lexical clues ("AI vocabulary"), as well as more complex phenomena within the text such as formality originality and clarity to identify AI from human writing.

Recent work has shown that pervasive LLM use may cause negative societal effects like homogenization (Zhang et al., 2025; Doshi & Hauser, 2024), with RLHF-trained models producing less diverse outputs (Padmakumar & He, 2024). Padmakumar et al. (2025) proposes a novelty metric balancing originality and quality, and corroborate prior finding that LLM text is less novel than human writing (Chakrabarty et al., 2024a). They also find that inference-time methods can boost novelty, though they increase originality at the expense of quality. Unlike prior work our study investigates the relationship between n-gram novelty and human-judged creativity through expert close reading annotations. We demonstrate that high n-gram novelty often correlates with reduced pragmaticality in LLM outputs, suggesting that optimizing for novelty alone may not lead to genuinely creative text.

7 Conclusion

We propose an operationalization of textual creativity beyond n-gram novelty. Rooted in the standard definition of creativity, our definition requires to assess both novelty and appropriateness (sensicality and pragmaticality) of text. We conducted a *close reading* study of human and AI-generated text collecting annotations from professional writers (n=26), obtaining a dataset of 7542 annotated expressions and 226 creative expression highlights with justifications. Our analysis reveals that as open-source LLMs generate more novel text, they tend to generate less pragmatic expressions, and that $\approx 91\%$ of top-quartile expressions by n-gram novelty are not judged as creative, cautioning against the use of n-gram novelty metric alone for creativity evaluation. We show that closed-source reasoning models can replicate some of the human judgments on creativity, and that their novelty scores are predictive of expert writing preferences.

REFERENCES

- H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19 (6):716–723, 1974. doi: 10.1109/TAC.1974.1100705.
- Teresa M Amabile. The social psychology of creativity: a componential conceptualization. *Journal of personality and social psychology*, 45(2):357, 1983.
- R Harald Baayen, Douglas J Davidson, and Douglas M Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412, 2008.
- Max Bachmann. rapidfuzz: Levenshtein.ratio levenshtein similarity ratio. https://rapidfuzz.github.io/Levenshtein/levenshtein.html#ratio, 2025. Accessed: 2025-09-07.
- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.
- Paul Brennan and Alan Silman. Statistical methods for assessing observer variability in clinical measures. *BMJ: British Medical Journal*, 304(6840):1491, 1992.
- Barry Brummett. Techniques of close reading. Sage Publications, 2018.
- Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024a. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642731. URL https://doi.org/10.1145/3613904.3642731.
- Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–34, 2024b.
- Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. Ai-slop to ai-polish? aligning language models through edit-based writing rewards and test-time computation. *arXiv* preprint arXiv:2504.07532, 2025a.
- Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. Can ai writing be salvaged? mitigating idiosyncrasies and improving human-ai alignment in the writing process through edits. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–33, 2025b.
- Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. Technical report, National Bureau of Economic Research, 2025.
- Mihaly Csikszentmihalyi et al. Flow and the psychology of discovery and invention. *HarperPerennial, New York*, 39:1–16, 1997.
- Anil R Doshi and Oliver P Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024.
- Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. What's in my big data? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024.* OpenReview.net, 2024. URL https://openreview.net/forum?id=RvfPnOkPV4.
- Bradley Emi and Max Spero. Technical report on the pangram ai-generated text classifier, 2024. URL https://arxiv.org/abs/2402.14873.

- Barbara Di Eugenio and Michael Glass. The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101, 03 2004. ISSN 0891-2017. doi: 10.1162/089120104773633402. URL https://doi.org/10.1162/089120104773633402.
- Alvan R. Feinstein and Domenic V. Cicchetti. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543-549, 1990. ISSN 0895-4356. doi: https://doi.org/10.1016/0895-4356(90)90158-L. URL https://www.sciencedirect.com/science/article/pii/089543569090158L.
- John Fox and Sanford Weisberg. An R Companion to Applied Regression. Sage, Thousand Oaks CA, third edition, 2019. URL https://www.john-fox.ca/Companion/.
- Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2007.
- Jacob Gildenblat. A python library for confidence intervals. https://github.com/jacobgil/confidenceinterval, 2023.
- Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. OLMo: Accelerating the science of language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15789–15809, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.841. URL https://aclanthology.org/2024.acl-long.841/.
- Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, et al. Which economic tasks are performed with ai? evidence from millions of claude conversations. *arXiv* preprint arXiv:2503.04761, 2025.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.
- Philip W Jackson and Samuel Messick. The person, the product, and the response: conceptual problems in the assessment of creativity. *Journal of personality*, 1965.
- Talia Kaufmann, Gabriel Agostini, Trivik Verma, and Daniel T. O'Brien. Inequalities in mobility patterns: Reconciling access and travel in american cities. In Susannah Cramer-Greenbaum, Adam Dennett, and Chen Zhong (eds.), *The 19th International Conference on Computational Urban Planning and Urban Management*, London, UK, June 2025. ACM Press. doi: doi.org/10.17605/OSF.IO/ABYQH.
- Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. Delving into Ilm-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27):eadt3813, 2025.
- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. *NA*, 2024.

- Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. Can language models learn from explanations in context? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 537–563, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.38. URL https://aclanthology.org/2022.findings-emnlp.38/.
- VI Levenshtein. Binary coors capable or 'correcting deletions, insertions, and reversals. In *Soviet physics-doklady*, volume 10, 1966.
- Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *First Conference on Language Modeling*, 2024.
- Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Raghavi Chandu, Nouha Dziri, and Yejin Choi. AI as humanity's salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025.* OpenReview.net, 2025. URL https://openreview.net/forum?id=iloEolgolQ.
- R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven. *Transactions of the Association for Computational Linguistics*, 11:652–670, 06 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00567. URL https://doi.org/10.1162/tacl_a_00567.
- William Merrill, Noah A. Smith, and Yanai Elazar. Evaluating *n*-gram novelty of language models using rusty-DAWG. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14459–14473, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main. 800. URL https://aclanthology.org/2024.emnlp-main.800/.
- Meta. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.
- Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Feiz5HtCD0.
- Vishakh Padmakumar, Chen Yueh-Han, Jane Pan, Valerie Chen, and He He. Beyond memorization: Mapping the originality-quality frontier of language models, 2025. URL https://arxiv.org/abs/2504.09389.
- Justus J Randolph. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*, 2005.
- Mark A Runco and Garrett J Jaeger. The standard definition of creativity. *Creativity research journal*, 24 (1):92–96, 2012.
- Jenna Russell, Marzena Karpinska, and Mohit Iyyer. People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5342–5373, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.267. URL https://aclanthology.org/2025.acl-long.267/.

- Robert Keith Sawyer and Danah Henriksen. Explaining creativity: The science of human innovation. Oxford university press, 2024.
- Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C Wallace. Detection and measurement of syntactic templates in generated text. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6416–6431, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.368. URL https://aclanthology.org/2024.emnlp-main.368/.
- Chantal Shaib, Tuhin Chakrabarty, Diego Garcia-Olano, and Byron C. Wallace. Measuring ai "slop" in text, 2025. URL https://arxiv.org/abs/2509.19163.
- Seabold Skipper and Perktold Josef. statsmodels: Econometric and statistical modeling with python. 9th Python in Science Conference, 2010.
- Barbara Herrnstein Smith. What was "close reading"? a century of method in literary studies. *the minnesota review*, 2016(87):57–75, 2016.
- Kanae Takahashi, Kouji Yamamoto, Aya Kuchiba, and Tatsuki Koyama. Confidence interval for micro-averaged f 1 and macro-averaged f 1 scores. *Applied Intelligence*, 52(5):4961–4972, 2022.
- Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2025. URL https://arxiv.org/abs/2501.00656.
- Hao Xu, Jiacheng Liu, Yejin Choi, Noah A. Smith, and Hanna Hajishirzi. Infini-gram mini: Exact n-gram search at the internet scale with fm-index. *ArXiv*, abs/2506.12229, 2025. URL https://api.semanticscholar.org/CorpusID:279402964.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.
- Tal Yarkoni. The generalizability crisis. Behavioral and Brain Sciences, 45:e1, 2022.
- Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. Noveltybench: Evaluating language models for humanlike diversity. *arXiv preprint arXiv:2504.05228*, 2025.

A DATA

A.1 SAMPLING EXPRESSIONS FOR ANNOTATION

To select pre-highlighted expressions, a passage was divided into atomic expressions via splitting by punctuation and heuristic rules. Then, we computed % of novel n-grams for each expression as follows:

- E: the expression
- $\mathcal{G}_n(E)$: all *n*-grams of length *n* in *T*
- \mathcal{D} : reference set of n-grams
- 1{·}: indicator function

$$n^* = \min \Big\{ n \ge 1 : \exists g \in \mathcal{G}_n(E) \text{ with } g \notin \mathcal{D} \Big\}$$

NovelPct
$$(n^*) = \frac{\sum_{g \in \mathcal{G}_{n^*}(E)} \mathbf{1}\{g \notin \mathcal{D}\}}{|\mathcal{G}_{n^*}(E)|}$$

That is, we compute the percentage of novel n-grams for n s.t. there is at least one novel n-gram. Among the various metrics we tested, this was the most interpretable and had the best balance between correlation with expression length (as we do not want to select only long expressions) and perplexity. We set a threshold for novelty very liberally at 15% which allowed us to get a large number of expressions annotated (roughly 50%).

A.2 SELECTING PASSAGES FOR ANNOTATION

We ensured that passages were not present in the respective OLMo model training corpora by manually checking that large n-grams from the passage as well as seemingly rare n-grams (e.g., unique proper nouns) had zero counts in the pre-training corpora. In addition, we had an automatic verification mechanism where a sample of 5 15-grams from the beginning, 5 from middle, and 5 from the end of the text were checked to have a 0 count in the pretraining data.

A.3 Annotation Instructions

We provide the annotation instructions in Figure 5.

A.4 DATASET EXAMPLES

Table 3 shows additional examples of low n-gram novelty creative expressions.

B LINEAR MODELS

B.1 N-GRAM NOVELTY AND CREATIVITY

We incorporate creativity highlights in the following manner: we add all highlighted expressions and their perplexity scores to the pre-highlighted ratings data, exclude expressions that were subsets, supersets or over 90% similar by Levenstein distance. Given that highlights allowed the annotators to select any expression,

Expression	Justification	PPL
disappearing from one room and reappearing in another	As in the previous clause, this metaphor defamiliarizes a mundane experience comparing it to the magical, heightening the surreality/drama of the narrator watching their daughter.	-1.6
you're the ugliest woman I've ever seen	This sentence marks a surprising turn from what has been an innocuous description of a date up to this point.	-1.5
the days of the week whisk by like panties	Inventive simile; surprising and humorous, central to the thematic conceit of the passage. Strong voice.	-1.4
lay across the roof of the house like the severed hand of a giant	Describing a felled pine bough as similar to "the severed hand of a giant" is delightfully fresh and original.	-1.4

Table 3: Examples of expressions judged creative with low n-gram novelty (log-standardized perplexity), along with the expert writers' justifications.

we consider expressions that were not highlighted (and were not in the pre-highlighted set) as non-novel by any annotator that did not highlight it. As a result, our dataset contains an enlarged number of 15,982 annotations.

We fit the model in Table 4 using the lmer package in R.

	creative ~ ppl_log_std + (1 gen_source) + (1 annotator) + (1 seed_passage_id)
(Intercept) ppl_log_std	$-3.68 [-4.16; -3.20]^* 0.67 [0.59; 0.75]^*$
AIC	5150.42
BIC	5188.82
Log Likelihood	-2570.21
Num. obs.	15982
Num. groups: seed_passage_id	50
Num. groups: annotator	26
Num. groups: gen_source	3
Var: seed_passage_id (Intercept)	0.37
Var: annotator (Intercept)	0.55
Var: gen_source (Intercept)	0.09

^{* 0} outside the confidence interval.

Table 4: Mixed-effects logistic regression predicting creativity from standardized PPL

B.2 N-GRAM NOVELTY AND PRAGMATICALITY

To focus on pragmaticality, we exclude expressions rated as non-sensical (so we know they do not make sense in the context, not just because they are ill-formed). The model is available in Table 5.

For frontier model study, we did not exclude these expressions since the sample size of non-pragmatic expressions was already very small (12 and 43 for Claude and GPT-5). In addition, since our research

	pragmatic ~ ppl_log_std * gen_source
	+ (1 annotator) +
	(1 seed_passage_id)
(Intercept)	3.88 [3.26; 4.49]*
ppl_log_std	0.01 [-0.17; 0.18]
gen_sourceolmo1	-1.47 [-1.78; -1.16]*
gen_sourceolmo2	$0.45[0.09; 0.81]^*$
ppl_log_std:gen_sourceolmo1	-0.18[-0.42;0.05]
ppl_log_std:gen_sourceolmo2	$-0.49 \ [-0.80; -0.18]^*$
AIC	2735.38
BIC	2790.53
Log Likelihood	-1359.69
Num. obs.	7290
Num. groups: seed_passage_id	50
Num. groups: annotator	26
Var: seed_passage_id (Intercept)	0.33
Var: annotator (Intercept)	1.98
·	·

^{* 0} outside the confidence interval.

Table 5: Mixed-effects logistic regression predicting creativity from standardized PPL

question concerns a head-to-head comparison of humans and frontier LLMs, we exclude passages generated by OLMo and OLMo-2. The model is available in Table 6.

We tested the following null hypotheses regarding the effect of perplexity (ppl_log_std) on pragmatic judgments within each generator:

Human (reference group): $H_0: \beta_{\text{ppl.log_std}} = 0$ LLM: $H_0: \beta_{\text{ppl.log_std}} + \beta_{\text{ppl.log_std:gen_sourceLLM}} = 0$

The tests were carried out with the function linear Hypothesis () from the car package Fox & Weisberg (2019), applied to the fitted glmer model.

B.3 CREATIVITY AND GENERATION SOURCE

To compare creativity along all generation sources, we add the frontier model annotations to the original data, adding the highlights in the same manner as above. The model is available in Table 7.

B.4 WRITING QUALITY REWARD MODELS AND CREATIVITY

The model for creativity is available in Table 8, and for pragmaticality in Table 9.

B.5 AI-LIKELIHOOD SCORES

The model for creativity is available in Table 10, and for pragmaticality in Table 11.

	pragmatic ~ ppl_log_std * gen_source
	+ (1 annotator) +
	(1 seed_passage_id)
(Intercept)	2.64 [1.63; 3.64]*
ppl_log_std	0.11 [-0.25; 0.48]
gen_sourceclaude	$1.51 \left[-0.03; 3.05\right]$
gen_sourcegpt5	0.08 [-1.39; 1.54]
ppl_log_std:gen_sourceclaude	-0.33[-1.00; 0.34]
ppl_log_std:gen_sourcegpt5	-0.28 [-0.78; 0.21]
AIC	610.09
BIC	652.27
Log Likelihood	-297.05
Num. obs.	1439
Num. groups: annotator	12
Num. groups: seed_passage_id	5
Var: annotator (Intercept)	1.54
Var: seed_passage_id (Intercept)	0.08

^{* 0} outside the confidence interval.

Table 6: Mixed-effects logistic regression predicting creativity from standardized PPL (comparing humans and frontier models)

	creative ~ gen_source + (1 annotator) +
	(1 seed_passage_id)
(Intercept)	$-3.10 [-3.44; -2.76]^*$
gen_sourceclaude	$-0.66 [-1.14; -0.19]^*$
gen_sourcegpt5	-0.81 [-1.32; -0.30]*
gen_sourceolmo1	$-0.65 [-0.88; -0.41]^*$
gen_sourceolmo2	$-0.47 [-0.71; -0.24]^*$
AIC	5946.38
BIC	6001.07
Log Likelihood	-2966.19
Num. obs.	18268
Num. groups: seed_passage_id	50
Num. groups: annotator	26
Var: seed_passage_id (Intercept)	0.34
Var: annotator (Intercept)	0.51

^{* 0} outside the confidence interval.

Table 7: Mixed-effects logistic regression predicting creativity from generation source (comparing humans and all models)

	creative ~ WQRM_score + (1 annotator) +
	(1 seed_passage_id)
(Intercept)	$-5.19[-5.84; -4.54]^*$
WQRM_score	$0.24 [0.17; 0.31]^*$
AIC	5418.08
BIC	5448.80
Log Likelihood	-2705.04
Num. obs.	15982
Num. groups: seed_passage_id	50
Num. groups: annotator	26
Var: seed_passage_id (Intercept)	0.35
Var: annotator (Intercept)	0.50

^{* 0} outside the confidence interval.

Table 8: Mixed-effects logistic regression predicting creativity from passage-level WQRM scores

	pragmatic ~ WQRM_score
	+ (1 annotator) +
	(1 seed_passage_id)
(Intercept)	2.63 [0.52; 4.74]*
WQRM_score	0.12 [-0.14; 0.38]
AIC	2758.21
BIC	2792.68
Log Likelihood	-1374.10
Num. obs.	7290
Num. groups: seed_passage_id	50
Num. groups: annotator	26
Num. groups: gen_source	3
Var: seed_passage_id (Intercept)	0.34
Var: annotator (Intercept)	2.01
Var: gen_source (Intercept)	0.56

^{* 0} outside the confidence interval.

Table 9: Mixed-effects logistic regression predicting pragmaticality from passage-level WQRM scores

	creative ~ std_log_ai_likelihood + (1 annotator) + (1 seed_passage_id)
(Intercept) std_log_ai_likelihood	-3.10 [-3.44; -2.76]* 0.06 [-0.16; 0.27]
AIC BIC Log Likelihood Num. obs. Num. groups: seed_passage_id Num. groups: annotator Var: seed_passage_id (Intercept)	3144.00 3171.76 -1568.00 7630 50 26 0.30
Var: annotator (Intercept)	0.53

^{* 0} outside the confidence interval.

Table 10: Mixed-effects logistic regression predicting creativity from passage-level AI likelihood scores

	pragmatic ~ std_log_ai_likelihood
	+ (1 annotator) +
	(1 seed_passage_id)
(Intercept)	3.82 [3.23; 4.40]*
std_log_ai_likelihood	-0.24 [-0.53; 0.04]
AIC	1041.98
BIC	1066.50
Log Likelihood	-516.99
Num. obs.	3394
Num. groups: seed_passage_id	50
Num. groups: annotator	26
Var: seed_passage_id (Intercept)	0.37
Var: annotator (Intercept)	1.50

* 0 outside the confidence interval.

Table 11: Mixed-effects logistic regression predicting pragmaticality from passage-level WQRM scores

C LLM Prompts and Hyperparameters

C.1 OLMO VERSIONS, PROMPTS AND GENERATION HYPERPARAMTERS

To generate LLM passages for annotation, we use the following simple prompts for OLMo (OLMo-7B-0724-Instruct) and OLMo-2 (OLMo-2-0325-32B-Instruct) in Table 12.

Row	Prompt
Summary Prompt	{story}\n\nSummarize in one sentence.
Passage Generation Prompt	${summary}\n\n\overline{c}$ a fragment of a story around ${n_words}$ words long in the style of ${author}$ based on the summary above.

Table 12: Prompts for summary and passage generation.

We sample with temperature =1 and truncate the output by paragraphs if they were too long, regenerating if they were too short (not within 10% of the original passage). For summaries, we manually ensured the overall topic was consistent with the passage (in 2-3 rare cases, the summaries were completely off-topic). For stories, we manually examined the outputs and regenerated the story in case of an output that did not at all adhere to the prompt (for example, sometimes OLMo model would generate a a play script instead of a passage, and once the model generated a story fully in Spanish).

We use the same generation prompt for Claude-4.1 and GPT-5 in the exploratory frontier model study, and use default API inference hyperparameters. We used the same summaries as for the OLMo models to exclude summary content effects.

C.2 LLM-AS-A-JUDGE VERSIONS AND GENERATION HYPERPARAMTERS

• GPT-5 Hyperparameters:

- Model version: gpt-5-2025-08-07

- verbosity: high

- reasoning effort: high

• Claude-4.1 Hyperparameters:

- Model version: claude-opus-4-1-20250805

- temperature = 1

• Gemini 2.5 Pro Hyperparameters: Default hyperparameters

C.3 PROMPTS FOR MODEL EVALUATION

We based our prompts very closely on annotator instructions (see Figure 5). Figure 6 shows the prompts used for zero-shot and few-shot model evaluation.

C.4 FINETUNING HYPERPARAMETERS

Below are the hyperparameters used for model fine-tuning:

• GPT-4.1

- Model version gpt-4.1-2025-04-14
- Fine-tuning type: full parameter (API-based)
- Batch size: 2
- Learning rate multiplier: 2
- Epochs: 3
- Seed: 42

• OLMo-2-7B

- Model version: OLMo-2-1124-7B-Instruct
- Fine-tuning type: LoRA
 - * LoRA rank: 8
 - * LoRA alpha: 8
 - * lora dropout 0.2
- per device batch size: 1
- gradient accumulation steps: 2
- learning rate 2e-4
- scheduler type linear
- warmup ratio 0.03
- weight decay 0.01
- epochs 3
- max seq length 4096

• Qwen3-8B

- Model version: Qwen3-8B
- Fine-tuning parameters: same as above

• LLama-3.1

- Model version: Llama-3.1-8B-Instruct
- Fine-tuning parameters: same as above

All open model fine-tuning was ran on 2 A100 40GB NVIDIA GPUs. We used the Open-Instruct codebase for fine-tuning script Lambert et al. (2024).

D CROWD PREFERENCE ALIGNMENT

We explore whether similar association is present in the LMArena dataset of crowd-sourced writing preferences released by Chakrabarty et al. (2025a). Each pair contains text generated from 2 language models (model A and model B). Data was sampled such that both model A and model B are in the top 15 most popular models, and so that it contains the same number of comparisons as the Style Mimic data (450). Similarly, we fit a logistic regression model adding random intercepts for the model A and model B. We find that novelty score differences were marginally predictive of the crowd preference ($\beta_{\Delta Nov_{AB}} = 0.21$, SE= 0.11, OR = 1.24, p = 0.054), and pragmaticality scores had a significant negative effect ($\beta_{\Delta Prag_{AB}} = -0.26$, SE= 0.11, OR = 0.77, p = 0.020). Together with Style Mimic findings, this indicates that while LLM-as-a-Judge novelty scores align with both expert and crowd preferences, pragmaticality scores only align with crowd preferences. This may indicate a misalignment between expert and crowd preferences on non-pragmatic expressions in writing. Figure 7 visualizes the difference of the creativity score effect on expert vs. crowd preferences.

E COMPARISON WITH CREATIVITY INDEX

We investigate how well do novelty scores from the LLM-as-a-Judge model align with expert preferences compared to Creativity Index scores. We compute the Creativity Index for each paragraph and use the difference between each pair as a predictor for preference (similarly to the ΔNov_{AB} , ΔPrag_{AB} scores). We find that Creativity Index is predictive of expert preferences ($\beta_{\Delta \text{CI}_{AB}} = 0.51$, SE= 0.24, OR = 1.66, p = 0.038). In comparison to the LLM-as-a-Judge novelty scores, novelty scores had a slightly stronger effect ($\beta_{\Delta \text{Nov}_{AB}} = 0.63$, p = 0.014). In a model including both predictors, novelty scores showed a trend-level effect ($\beta_{\Delta \text{Nov}_{AB}} = 0.48$, p = 0.010), while Creativity Index scores did not contribute significantly ($\beta_{\Delta \text{CI}_{AB}} = 0.26$, p = 0.358). Likelihood ratio tests confirmed that adding Creativity Index predictors to a model containing novelty scores did not improve model fit (p = 0.35), whereas adding novelty scores to a model containing Creativity Index showed a trend toward improved fit (p = 0.096). Model comparison using AIC (Akaike information criterion, a measure used to compare statistical models) Akaike (1974) also favored novelty: the model with only novelty had the lowest AIC (471.16) compared with the Creativity Index-only model (473.07) and the joint model (472.30), indicating that novelty scores provide the most parsimonious explanation of expert preference.

F ADDITIONAL FIGURES

Annotation Instructions

Hide Instructions ▲

Thank you for participating in our study! Below you will find a passage for creativity annotation. Your goal is to identify expressions that contribute to this text's uniqueness and creative value through close reading of the passage. The task consists of two parts:

Part 1: Pre-highlight Rating

Rate whether the already highlighted expressions adhere to the following three criteria of creativity:

VERY IMPORTANT: For each expression, evaluate: does it makes sense? does it flow naturally? it is novel? Provide judgements on all three of these categories, not just one of them.. When you turn the toggle on, e.g. for the Making Sense question, it means you annotate the expression as Making Sense (answering YES).

Makes Sense On Its Own

Whether the expression makes sense by itself in terms of grammar and overal fluency. If the expression could make sense in some context, mark it as making sense by itself, there is no context in which the expression makes sense, mark it as not making sense by itself, and also as not making sense in context.

For example, "conversation veering towards the topic" makes sense, but "conversation veering towards the talk of the topic" might not.

Similarly, "from their apartment on East Tenth Street" should be marked as making sense whereas "from it cloud on East Tenth Street", i.e. something completely ungrammatical / nonsensical, no matter what context you put it in, should be marked as not.

Makes Sense Contextually (Flows Naturally)

Whether expression makes sense in the context of the text, has no logical or continuity issues, does not sound awkward, odd, or incoherent with the rest of the text.

For example, if the previous text describes that "Alice's life became sad and dull" and the expression following it is "And so Alice flourished", describing how happy she is, it could be considered incoherent since it logically contradicts the previous passage.

Another example: "her fingers tended to his fire" may sound awkward and unnatural.

Note: Generally, if the expression does not make sense on its own, it also does not make sense in context.

Novel

Whether the expression is unusual, surprising, or original given the context of the text.

A powerful metaphor, an interesting detail about a character, or an insightful literary technique could all contribute to contextual novelty. For example, a personification "as if the tables have become heavier since she walked in" could be considered surprising and unusual.

Important: You will be asked to provide a justification for why you find the expression novel if you check this box.

Important: The expression you select has to be novel, unusual, surprising in some way, not simply "strong".

Example (expression pre-highlighted in yellow could be marked as novel):

Then, on the sixth of November, 2015, my dad had a sudden heart attack, the result of a hereditary disease that had already claimed five or six people in my family, the first of them at the end of the Qing dynasty

Justification: the juxtaposition of "first" and "end" in this fragment makes the reader think about the cyclical nature of time, which is a big theme in the rest of the passage. The author hints that while some events (the fall of a dynasty here, later on a disease or a heart attack) may be seen as the decline or final stage of something, they are also new beginnings.

Part 2: Creativity Highlights

Using your cursor, select any expressions that were not highlighted, but which you **find creative** according to the definition below. You can highlight creative expressions **at the same time** as you are completing Part 1.

Selection Criteria

The expressions you select must meet ALL THREE criteria simultaneously:

- Make sense
- How natural
 Be novel

Important: Do not select "bad" expressions! Focus on identifying the creative ones.

Important: Creative expressions have to satisfy ALL THREE criteria. Importantly, they have to be perceived as novel by you.

You may select expressions concurrently with rating the highlighted expressions from Part 1.

You can also highlight parts of the pre-highlighted expressions.

Example of a highlighted expression:

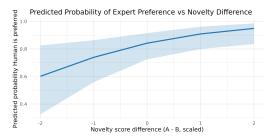
| "while standing in other corners, watching other people move similar furniture around similar indoor spaces"

Justification: The parallel and anti-parallel structure of the sentence (other/other, similar/similar) works to characterize the situation Ivan finds himself simultaneously novel and familiar. There is also a break in the linear flow of what can and should differ in each of the situation: corners, people, furniture, and spaces do not nest themselves in this way and the idea that similar furniture can occur along with other people—in Ivan's head—helps me understand the conceptualization, universalist process that the character is applying to the scene.

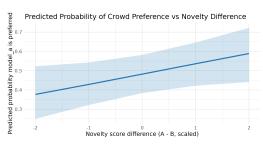
Figure 5: Annotation Instructions

Novel Expression Judge Prompt	Below you will find a passage for creativity annotation. Your goal is to identify expressions that contribute to this text's uniqueness and creative value through close reading of the passage.
	Find all expressions that are novel in the passage. A novel expression is unusual, surprising, or original given the context of the text. A powerful metaphor, an interesting detail about a character, or an insightful literary technique could all contribute to contextual novelty. For example, a personification "as if the tables have become heavier since she walked in" could be considered surprising and unusual. Find as many expressions as you can but be sure to select only those that are truly novel. Important: You will be asked to provide a justification for why you find the expression novel. Important: The expression you select has to be novel, unusual, surprising in some way, not simply "strong". Example: Passage: Then, on the sixth of November, 2015, my dad had a sudden heart attack, the result of a hereditary disease that had already claimed five or six people in my family, the first of them at the end of the Qing dynasty. Output: {"expression": "the first of them at the end of the Qing dynasty", "justification": "the juxtaposition of "first" and "end" in this fragment makes the reader think about the cyclical nature of time, which is a big theme in the rest of the
	passage. The author hints that while some events (the fall of a dynasty here, later on a disease or a heart attack) may be seen as the decline or final stage of something, they are also new beginnings."} Passage: {passage}
Non-pragmati Expression Judge Promp	and the expression following it is "And so Alice flourished", describing how happy

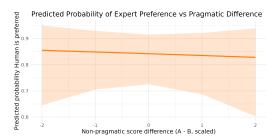
Figure 6: Prompts used for few-shot and zero-shot model evaluation.



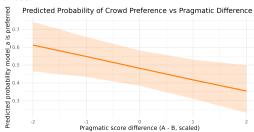
(a) Effect of novelty scores on expert preferences.



(c) Effect of novelty scores on expert preferences.



(b) Effect of pragmaticality scores on expert preferences.



(d) Effect of pragmaticality scores on crowd preferences.

Figure 7: Comparison of the impact of creativity on expert vs. crowd preferences, bands indicate 95% confidence intervals, group-level intercepts are population-level fixed effects.

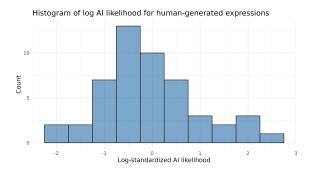


Figure 8: passage-level log AI-likelihood distribution for human-written expressions.

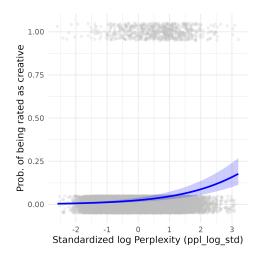


Figure 9: Effect of n-gram novelty (log-standardized perplexity) on probability of the expression being rated as novel.

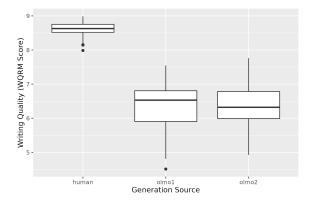
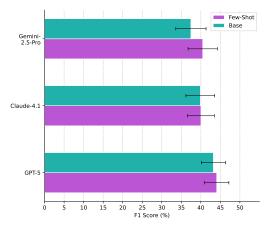
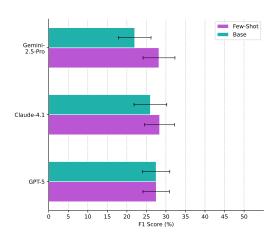


Figure 10: Writing quality reward model distribution by generation source.





(a) **Novel Expression Identification:** few-shot and zero-shot results.

(b) **Non-Pragmatic Expression Identification:** few-shot and zero-shot results.

Figure 11: Comparison of zero-shot and few-shot model performance across (a) novel and (b) non-pragmatic expression identification tasks on the full test set of 97% of the data excluding 3 few-shot passages only (F1 scores with 95% CIs computed with confidenceinterval package (Gildenblat, 2023) using the Takahashi et al. (2022) method.