Curiosity-Driven LLM-as-a-judge for Personalized Creative Judgment

Vanya Bannihatti Kumar[1], Divyanshu Goyal[2], Akhil Eppa[3], and Neel Bhandari[4]

Adobe

Abstract. Modern large language models (LLMs) excel at objective tasks such as evaluating mathematical reasoning and factual accuracy, yet they falter when faced with the nuanced, subjective nature of assessing creativity. In this work, we propose a novel curiosity-driven LLM-as-a-judge for evaluating creative writing which is personlized to each individual's creative judgments. We use the Torrance Test of Creative Thinking(TTCW) benchmark introduced in (6), which has stories annotated by expert humans across various subjective dimensions like *Originality*, to test our hypothesis. We show that our method enables models across various sizes, to learn the nuanced creative judgments of different individuals, by showing improvements over baseline supervised finetuning(SFT) method across various evaluation metrics like Pearson correlation, Cohen's κ and F1 values. Our method is especially useful in subjective evaluations where not all the annotators agree with each other.

Keywords: Curiosity-driven learning · Creativity Evaluation · Personalisation

1 Introduction

Rigorous, standardized evaluation has repeatedly catalyzed progress in machine learning, from ImageNet(27) and GLUE(33), driving leaps in the fields of computer vision and Natural Language Processing, respectively. The same effect is evident in objective math reasoning, where benchmarks like GSM8K(7), together with RL-trained reasoning models such as OpenAI's o1(20) and DeepSeek-R1(10) have obtained strong results on hard contests like AIME and IMO.

While robust evaluation metrics exist for objective tasks such as mathematical reasoning and factual verification, subjective tasks like creativity remain difficult to assess reliably. There are many previous works (23; 35) which show that using Large Language Models(LLM) as a judge prefer their own generations making them unreliable. Despite the success of LLMs on objective benchmarks, they still struggle to evaluate creativity in a manner aligned with human judgment. As shown in (6) and Table 12 and Table 2, even state-of-the-art models fall short in consistently evaluating the subjective dimensions of the story as well as a human expert. This can be attributed to the fact that individual preferences shape creativity and rarely align uniformly across people.

To address this gap, we present an enhanced LLM-as-a-judge that not only learns from a diverse pool of annotations but also adapts its scoring to align with individual annotators or experts. This allows for more faithful and preference-aware evaluation of creativity. We emphasize personalization in our framework because the task of assessing subjective criteria is inherently variable across individuals. To this end, we propose a curiosity-driven LLM-as-a-judge for evaluating creativity in text generation, drawing inspiration from the curiosity-based Reinforcement Learning (RL) framework of (25). However, unlike the RL setting in (25), we reinterpret curiosity as an *belief-shift signal* for creative evaluation. Specifically, when the model is "surprised" by an expert's explanation, it signals a mismatch between the LLM's prior belief and the expert's preference; conversely, low surprise indicates alignment between the LLM and the expert (see Fig 5. To implement this, we first train an Intrinsic Curiosity Model (ICM) that measures the LLM's surprise at a given explanation while simultaneously predicting which expert or annotator produced the explanation. The intuition behind predicting the annotator is that the model can learn which annotator caused the belief shift, allowing it to calibrate the curiosity signal for each annotator individually, thereby improving personalization. The resulting *curiosity score* is then fed as an auxiliary, self-supervised signal to improve a supervised fine-tuning (SFT) model (see Fig 1).

In our experiments, we establish a baseline using an SFT model that predicts annotators' binary judgments from the story and question (see Fig 3a). To evaluate the effect of curiosity, we enhance this baseline with an ICM-derived curiosity score. More concretely we append the curiosity score to story and question in the baseline model. This helps us do a fair comparison on effect of curiosity signal on the final judgment and thereby measure the lift in performance our methodology provides over the baseline.

We conduct extensive experiments across various model sizes to ensure our method scales well with model size. Since the TTCW dataset size is extremely small, we do a 5-fold cross validation in order to ensure that our results are statistically significant. We also test our method in out-of-distribution scenarios to ensure that our method generalizes well. Averaged across model sizes, ICM significantly improves Pearson correlation and F1 scores. More details about the results can be found in Fig 4.

2 Methodology

In this section, we describe our curiosity-driven LLM-as-a-judge for evaluating creativity in text generation, which combines belief shift estimation with expert attribution. Our method leverages the TTCW dataset (6), which is based on the Torrance Test of Creative Thinking (30) but adapted for LLMs. We focus on a subset of five creativity dimensions particularly relevant for evaluating the creative judgments of generative language models. We detail the dataset structure, model architecture, loss functions, and the formulation of our curiosity signal.

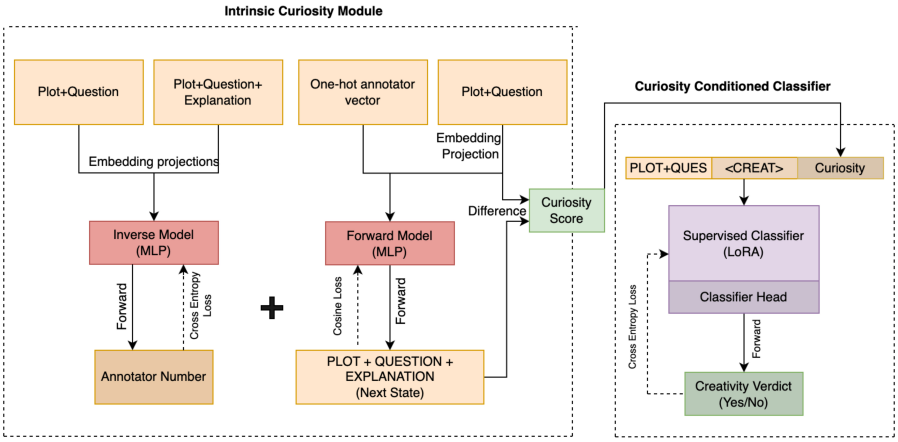


Fig. 1: Overview of Architecture during training for Curiosity-Driven LLM-as-a-judge

2.1 Dataset

The TTCW dataset[1] provides expert human-annotated creativity judgments across 14 distinct dimensions. All the distinct dimensions in the TTCW dataset are mentioned in Appendix A.1. For this study, we focus on five dimensions, 3 of which are categorised under Originality and 2 under flexibility: *Originality in Thought*, *Originality in Form*, *Originality in Theme and Content*, *Structural Flexibility*, and *Perspective and Voice Flexibility*. Our analysis is restricted to these five dimensions, encompassing all dimensions under *Originality* and two representative dimensions from *Flexibility*. We picked these 5 dimensions among the 14(Table 4) as these are more subjective in nature and hence the most ideal to evaluate our methodology. We defer exploration of the remaining dimensions to future work. Questions associated with each dimension can be found in appendix 6.

2.2 Data Format and Task Setup

Each example in the dataset consists of a story S, a creativity-focused question Q_d specific to dimension d, an expert ID z_i where $i \in \{1, 2, 3\}$ for each annotation by an expert, three expert-provided explanations $\mathcal{E} = \{e_1, e_2, e_3\}$, and corresponding binary verdicts $V_i \in \{\texttt{yes}, \texttt{no}\}$ for each explanation.

The task is to improve the model's performance on producing judgments similar to that of a particular expert when the model is presented with the story and the creative question

[1] Huggingface TTCW dataset

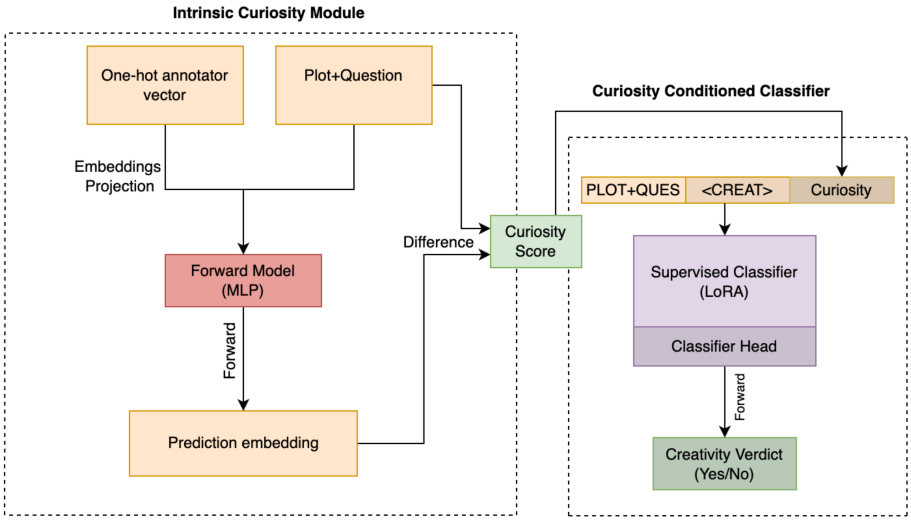


Fig. 2: Overview of Architecture during inference for Curiosity Driven LLM-as-a-judge

2.3 Intrinsic Curiosity Model Overview

Our model operates in two stages:

1. **Belief Shift Estimation (Forward Score)**: The model measures the impact of an expert explanation on their prediction of creativity.
2. **Expert Attribution (Backward Score)**: The model identifies which expert wrote a given explanation.

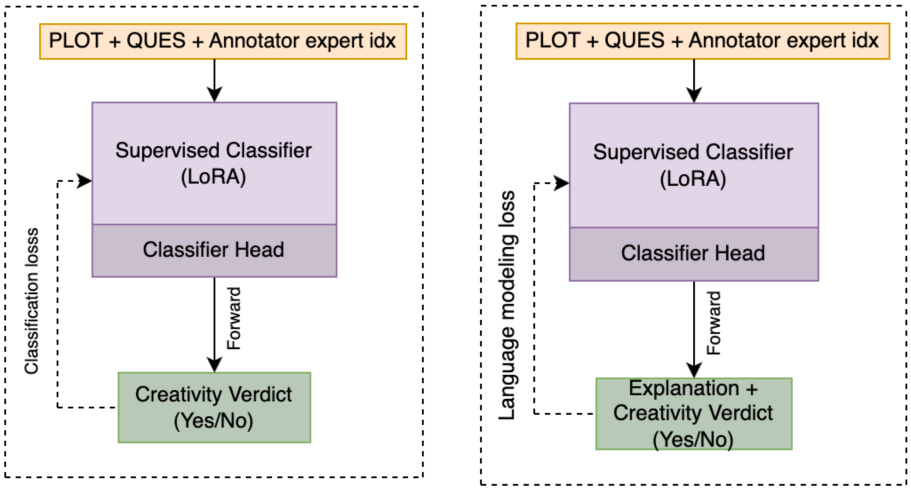
Forward Score: Belief Shift via Cosine Loss We define two states:

- **State A**: Input consisting of the story and question and one-hot vector of the expert ID z_i represented as $(S, Q_d, onehot(z_i))$ where $i \in \{1, 2, 3\}$ as each story-question pair is annotated by 3 experts.
- **State B**: Input augmented with one expert explanation (S, Q_d, e_i) where $i \in \{1, 2, 3\}$.

Let $f_\theta^{(A)} = f_\theta(S, Q_d, onehot(z_i))$ and $f_\theta^{(B)} = f_\theta(S, Q_d, e_i)$, where f_θ denote the judge's scoring function (logit head) with parameters θ that maps the input to a scalar judgment logit.

The forward loss is defined as the cosine loss between these two predictions:

$$\mathcal{L}_{\text{forward}} = 1 - \frac{f_\theta^{(A)} \cdot f_\theta^{(B)}}{\|f_\theta^{(A)}\| \|f_\theta^{(B)}\|}$$



(a) Baseline without using explanations (b) Baseline using explanations

Fig. 3: Comparison of baselines with and without using explanations.

This loss captures how much the model's belief about creativity of the story shifts when it incorporates the explanation by the annotator, which we define as the intrinsic curiosity measure.

Backward Score: Expert Attribution via Cross-Entropy To help the model to understand the distinct reasoning styles of different experts, we introduce an auxiliary classification task. Given (S, Q_d, e_i), the model predicts the identity of the expert $z_i \in \{1, 2, 3\}$ who authored explanation e_i:

$$p_\phi(z_i \mid S, Q_d, e_i) = \text{softmax}(g_\phi(S, Q_d, e_i))$$

The backward loss is the cross-entropy between the predicted and true expert label:

$$\mathcal{L}_{\text{backward}} = -\log p_\phi(z_i \mid S, Q_d, e_i)$$

Loss function of Intrinsic curiosity model(ICM) We define the ICM model's loss as a weighted combination of the forward and backward components:

$$\mathcal{L}_{\text{curiosity}} = \mathcal{L}_{\text{forward}} + \lambda \cdot \mathcal{L}_{\text{backward}}$$

where λ is a tunable hyperparameter that balances the two objectives. In our experiments we set λ as 1.

Incorporating the Curiosity Signal to SFT To evaluate the utility of the learned curiosity signal, we use it as a conditioning input to a supervised fine-tuning (SFT) model trained to predict expert verdicts. For each instance, we append the scalar curiosity score to the original input using a special delimiter token `<CREAT>`, resulting in the following input format:

$$\begin{aligned} \text{Input:} \quad & Q_d + S + \texttt{<CREAT>} \\ & + Curiosity_{\text{Score}} \quad \longrightarrow \quad \text{Target: } V_i \end{aligned} \tag{1}$$

$$\begin{aligned} \text{Curiosity}_{\text{score}} = & f_\theta\big(S,\, Q_d,\, e_i\big) \\ & - f_\theta\big(S,\, Q_d,\, \text{onehot}(\text{expert\_idx})\big) \end{aligned} \tag{2}$$

$V_i \in \{\texttt{yes}, \texttt{no}\}$ is the binary verdict associated with explanation e_i. The model uses the $Curiosity_{\text{Score}}$ as a signal to predict the verdict of the given annotator. We use cross-entropy loss for training this classifier model

2.4 Inference

During inference(see Fig 2), the story and creativity-focused questions are first passed through the intrinsic curiosity model (ICM) to compute a curiosity score. This score reflects the model's internal belief shift in response to the input for that particular annotator. The resulting curiosity score is then appended to the original input, using a special delimiter token `<CREAT>`—and passed to the SFT classifier model. This classifier then predicts the binary creativity verdict (`yes` or `no`) for the given story-question pair. .

2.5 Baseline with explanations

For the baseline comparison , we use a standard SFT model that produces the explanation and binary verdict given the input(see fig. 3b). The model input is structured as:

$$\texttt{Input:} \quad Q_d + S + z_i \quad \longrightarrow \quad \texttt{Target:}\{V_i, e_i\}$$

At inference time, we provide Q_d, S, and z_i as input, and the model outputs a JSON structure, from which the predicted verdict is parsed and compared to the ground truth. This baseline is trained using language modeling loss.

2.6 Baseline without explanations

We ensure to compare our method against the baseline SFT in a classification setting rather than a causal language model setting to ensure fairness in comparison(see fig. 3a). Since we set up the baseline SFT in a classification setting,

we do not include the explanations as neither part of the input or the output of the classification task. In this classification setting we use the question and the story as part of input and the verdict as part of the output.

$$\texttt{Input:} \quad Q_d + S + z_i \quad \longrightarrow \quad \texttt{Target:}\{V_i\}$$

2.7 Evaluation

Evaluating subjective tasks like creativity presents unique challenges, as even human annotators often disagree on what constitutes a "correct" judgment. Rather than attempting to define a universal metric for creativity, our approach embraces this subjectivity by focusing on personalization. We aim to adapt evaluation signals to individual experts by learning from a small number of their labeled examples. This allows us to model subjective preferences more faithfully and use this personalized model to assess creativity in a user-aligned manner. To quantify model performance in capturing individual judgments, we report **Pearson Correlation** (4) and **Cohen's** κ (8), along with **Precision**, **Recall**, and **F1-score**. These metrics enable us to assess both the predictive accuracy and ranking consistency of our models in aligning with subjective human evaluations.

3 Theory: Why Curiosity Beats Using Explanation Text Directly

Let e denote the expert's explanation, $\mathrm{x} = Q_d{+}S$, $s_{\text{base}}(x) = f_\theta\big(S,\, Q_d,\, \text{onehot}(z_i)\big)$ the pre-explanation logit, and $s_{\text{expl}}(x, e_i) = f_\theta\big(S,\, Q_d,\, e_i\big)$ the post-explanation logit produced by the model when conditioned on e. The $\text{Curiosity}_{\text{Score}}$ is defined as the belief shift.

$$\text{Curiosity}_{\text{score}} \;=\; f_\theta\big(S,\, Q_d,\, e_i\big) - f_\theta\big(S,\, Q_d,\, \text{onehot}(z_i)\big),$$

and *discard* e thereafter. We train a predictor $\hat{p}_\theta(V{=}1 \mid x, \text{Curiosity}_{\text{score}}) = \sigma\big(h_\theta(x, \text{Curiosity}_{\text{score}})\big)$ where V is the verdict, h is the LLM judge model and σ represents softmax. This yields three advantages grounded in standard theory.

(1) Weight-of-evidence sufficiency. In logit/Bayesian updates, additional information acts *additively* on log-odds via a log-likelihood ratio (*weight of evidence*) (1):

$$\log \frac{\Pr(V=1 \mid x, e_i)}{\Pr(V=0 \mid x, e_i)} = \log \frac{\Pr(V=1 \mid x)}{\Pr(V=0 \mid x)} + \underbrace{\log \frac{p(e \mid V=1, x)}{p(e \mid V=0, x)}}_{\text{weight of evidence}} \tag{3}$$

In our methodology, $\text{Curiosity}_{\text{Score}} = s_{\text{expl}} - s_{\text{base}}$ is an *empirical estimate* of this increment on the log-odds scale, so it preserves the decision-relevant

Table 1: ICM method results against the SFT baseline with explanations

Model	Exp.	LoRA α/R	Pearson	Cohen's κ	F1	Precision	Recall
Qwen0.5B	SFT	256/256	0.170 ±0.049	0.155 ±0.046	0.382 ±0.049	0.452 ±0.059	0.334 ±0.060
	ICM	32/16	**0.524 ±0.092**	**0.383 ±0.076**	**0.616 ±0.048**	**0.494 ±0.046**	**0.818 ±0.067**
Qwen1.5B	SFT	256/256	0.170 ±0.048	0.155 ±0.048	0.402 ±0.049	0.432 ±0.020	0.383 ±0.083
	ICM	32/16	**0.587 ±0.061**	**0.406 ±0.065**	**0.629 ±0.045**	**0.506 ±0.045**	**0.836 ±0.056**
Qwen3B	SFT	256/256	0.113 ±0.083	0.110 ±0.081	0.339 ±0.051	0.401 ±0.067	0.298 ±0.060
	ICM	32/16	**0.540 ±0.057**	**0.356 ±0.081**	**0.598 ±0.054**	**0.481 ±0.050**	**0.794 ±0.070**
Qwen7B	SFT	128/128	0.160 ±0.050	0.168 ±0.085	0.371 ±0.021	0.443 ±0.050	0.324 ±0.038
	ICM	32/16	**0.605 ±0.083**	**0.429 ±0.082**	**0.643 ±0.053**	**0.518 ±0.051**	**0.850 ±0.072**

effect of e while removing lexical/style nuisance. Consequently, conditioning on $\text{Curiosity}_{\text{Score}}$ approximates the theoretically "right" sufficient update in a logistic decision rule (1).

(2) Variance reduction via a control-variate effect. Let Z be the random quantity we wish to estimate more stably (e.g., per-example loss), and let $C = \text{Curiosity}_{\text{Score}}$ be the control signal. With Pearson correlation

$$\rho \;=\; \text{Corr}(Z, C) \;=\; \frac{\text{Cov}(Z, C)}{\sqrt{\text{Var}(Z)\,\text{Var}(C)}} \in [-1, 1],$$

the classic control-variate construction implies that the optimally adjusted estimator $Z^\star = Z - \alpha^\star(C - \mathbb{E}[C])$ achieves

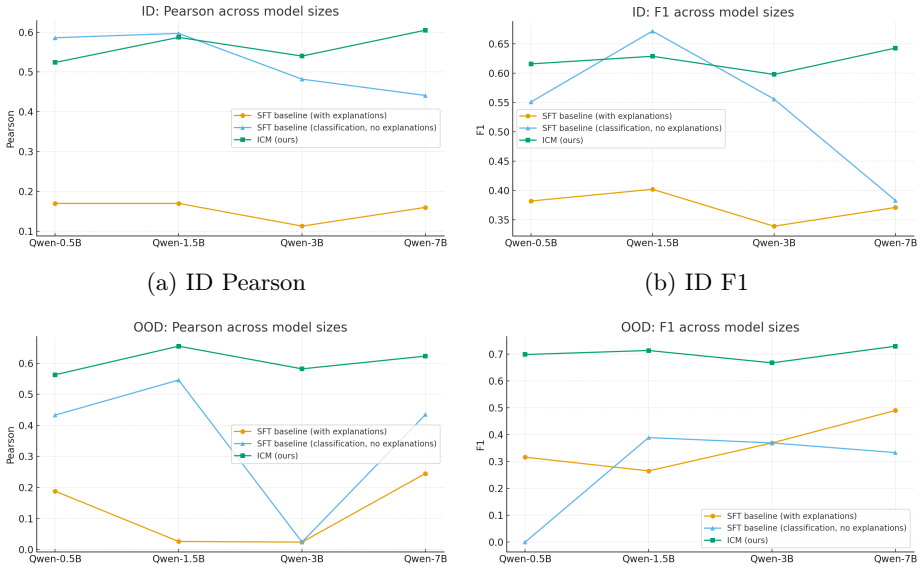
$$\text{Var}(Z^\star) \;=\; \text{Var}(Z)\,\big(1 - \rho^2\big) \quad \text{at} \quad \alpha^\star = \frac{\text{Cov}(Z, C)}{\text{Var}(C)}.$$

Thus any nonzero correlation with c strictly reduces variance (21, Ch. 8). Here, $Z = \ell_i(\theta)$ (per-example cross-entropy loss) to reduce risk variance. Lower variance improves sample efficiency and stabilizes training.

(3)Curiosity as a Model of Annotator Behaviour and Generalization Subjective labels reflect both item difficulty and rater idiosyncrasy. A classic way to formalize this is a random–effects logit (9; 1):

$$\text{logit}\ \Pr(V{=}1 \mid x, z_i) \;=\; f(x) \;+\; b_{z_i}(x), \tag{4}$$

where $f(x)$ captures item evidence and $b_a(x)$ represents the (possibly context-dependent) strictness/leniency of annotator a. Since the curiosity score is able to model the annotator behaviour without considering the idiosyncrasies of the explanation text, it is able to better generalize to out-of-distribution dimensions for that annotator.



(a) ID Pearson (b) ID F1

(c) OOD Pearson (d) OOD F1

Fig. 4: Three-way comparison across model sizes for **ICM (ours)**, **SFT baseline (classification, no explanations)**, and **SFT baseline (with explanations)**. Panels show Pearson and F1 for in-distribution (top) and out-of-distribution (bottom). For exact results of the ID and OOD experiments of *baseline without explanation(classification)*, refer to Table 13 and Table 14

4 Experiments

We evaluate our Intrinsic Curiosity Modeling (ICM) approach against a supervised fine-tuning (SFT) baseline (see Section 2) across multiple model sizes. For a fair comparison in terms of identical input and outputs, we compare the ICM setup against SFT baseline with explanations. We also compare the ICM setup against FT baseline without explanations in order to ensure the same classification loss is used.

Dataset TTCW contains 48 stories annotated on 5 dimensions with three expert judgments per story–dimension pair, yielding 720 examples. We use 5-fold cross-validation with an 80/20 split, giving approximately 576 training and 144 test items per fold. Because individual folds are small, we report means across folds for all metrics (Table 1; see also Section 2.7). Splits are stratified to preserve the positive/negative label ratio.

Training setup. The *baseline with explanations* uses a causal language modeling objective and our ICM model uses a classification objective. We align shared

Table 2: Comparison of ICM method against GPT-5 one-shot

Model	Exp.	Pearson	F1	Precision	Recall
Qwen0.5B	ICM	0.524 ±0.092	0.616 ±0.048	0.494 ±0.046	0.818 ±0.067
Qwen1.5B	ICM	0.587 ±0.061	0.629 ±0.045	0.506 ±0.045	0.836 ±0.056
Qwen3B	ICM	0.540 ±0.057	0.598 ±0.054	0.481 ±0.050	0.794 ±0.070
Qwen7B	ICM	0.605 ±0.083	0.643 ±0.053	0.518 ±0.051	0.850 ±0.072
GPT-5	ICM	0.2409 ±0.1379	0.3467 ±0.1592	0.5698 ±0.2305	0.2608 ±0.1378

hyperparameters—learning rate, LoRA (15) rank, and batch size—wherever applicable to ensure comparability. The ICM combined loss uses $\lambda = 1$. All fine-tuning (ICM and SFT baselines) uses LoRA; full details are in Table 5. For the *baseline without explanations*, which also uses a classification loss, we match all of the ICM hyperparameters.

Compute and precision. All runs use a single NVIDIA A100 (80 GB) GPU. Mixed precision with **bfloat16** is enabled when supported. When base models are loaded with 8-bit quantization, matrix multiplies in bitsandbytes execute in FP16 while LoRA heads operate in bfloat16.

Convergence and reproducibility. We train to loss convergence in all runs and fix random seeds for data splits and initialization. Hyperparameters and implementation details appear in Table 5.

5 Analysis

5.1 Effect of model scale

From Fig 4 we can see that our ICM method improves across model sizes whereas the *baseline classification method with no explanation* degrades with increase in model size for both ID and OOD settings. The reason why the *baseline classification method with no explanation* maybe degrading with scale is because this method primarily overfits on the small dataset with larger model sizes. Although the *baseline with explanation* improves with increase in model size, it remains uniformly low compared to the ICM method.

5.2 Generalization

To understand the generalization ability of the baseline and the ICM models, we use the same setup as earlier but train the model in both methods on 4 dimensions - *Originality in Form*, *Originality in Theme and Content*, *Structural Flexibility*, and *Perspective and Voice Flexibility*, and test these trained models

Table 3: ICM method results against the SFT baseline with explanations on Out-of-distribution data

Model	Experiment	LoRA α/Rank	Pearson	Cohen's κ	F1	Precision	Recall
Qwen0.5B	SFT	256/256	0.188	0.147	0.316	0.632	0.211
	ICM	32/16	**0.563**	**0.458**	**0.698**	**0.625**	**0.790**
Qwen1.5B	SFT	256/256	0.026	0.023	0.265	0.423	0.193
	ICM	32/16	**0.655**	**0.486**	**0.713**	**0.639**	**0.807**
Qwen3B	SFT	256/256	0.024	0.024	0.369	0.413	0.333
	ICM	32/16	**0.582**	**0.403**	**0.667**	**0.597**	**0.754**
Qwen7B	SFT	128/128	0.245	0.237	0.490	0.585	0.421
	ICM	32/16	**0.623**	**0.514**	**0.729**	**0.653**	**0.825**

on the held out dimension of *Originality in Thought*. In this way there is absolutely no data leakage since the dimension the model is tested on was never seen during the training. From figure 4, we can see that gains of the ICM method over both the baseline methods are much more in the OOD settings rather than ID settings. This suggests the generalizability of our method because we are essentially allowing the model to understand the user behavior before predicting which is much more generalizable as compared to both baseline SFT methods.

5.3 Comparison with GPT-5

Table 2 has the results of the ICM setup against GPT-5. We can see that even Qwen-0.5B model is able to beat GPT-5 model across all evaluation metrics except precision. The GPT-5 model was prompted with the same story, question and annotator index along with one shot example(randomly picked from training set) by the same annotator. GPT-5 model was more biased towards the answer "no" and whenever "yes" was predicted, it was almost always wrong. This further proves the effectiveness of our method.

6 Conclusion and Future Work

We introduced a curiosity-driven LLM-as-a-judge for evaluating creativity in text generation, addressing the limitations of baseline SFT for inherently subjective tasks. Our approach leverages a two-part curiosity signal, capturing belief shifts via model responses to expert explanations and incorporating expert attribution through a backward prediction task. This signal enhances a SFT setup, leading to stronger alignment with human judgments across multiple creativity dimensions in the TTCW dataset. Experiments show that incorporating curiosity-based modeling consistently improves performance across model scales, surpassing standard SFT baselines in both correlation with human ratings and classification accuracy. Not only does it scale with model size, it also improves

the performance in out-of-distribution scenarios, where we test the models on one heldout test dimension by training the models on the other 4 creativity dimension. Future work includes extending the curiosity-driven LLM-as-a-judge to other domains like marketing, evaluating novelty of scientific ideas etc,. We also plan to use the curiosity signal as a reward signal in RL setup to further improve our current results.

7 Literature Review

The evaluation of creativity in language models builds upon decades of work in creativity research, where the Torrance Tests of Creative Thinking (TTCT) assess fluency, flexibility, originality, and elaboration (30), and the Consensual Assessment Technique (CAT) uses aggregated expert judgments, a reliable but labour-intensive process (26). The authors of (6) adapted TTCT into the Torrance Tests for Creative Writing (TTCW), designing fourteen binary tests and enlisting creative-writing experts to evaluate 48 stories; their study showed that large language models pass these tests three to ten times less often than human writers (6), highlighting a sizable gap in creative competence. Alternative evaluation paradigms, such as the Leap-of-Thought (LoT) framework for humorous, associative reasoning, argue that step-by-step chain-of-thought prompting can limit creativity and instead encourage non-sequential "leaps" (41). Efforts to automate creativity scoring (e.g., distributional-semantics proxies for novelty) often align weakly with expert judgments, reinforcing the need for human-aligned signals.

Because creativity judgments are *subjective*, collapsing rater perspectives via majority vote can erase systematic, meaningful disagreement. Following work on multi-annotator modeling, we treat annotators as distributions to be modeled rather than aggregated away (18), rather than use the classical aggregation methods that infer a single latent "truth" (37; 14). In parallel, recent results caution against naïve *LLM-as-judge* usage: evaluators can recognize and prefer their own generations, introducing self-preference bias (24). Calibrated autoraters offer a partial mitigation via broad multi-task training and bias auditing (31). These findings motivate rater-aware or human-anchored evaluation signals for creativity.

Intrinsic-motivation signals from reinforcement learning offer a principled lens on novelty seeking. Information-gain and prediction-error formulations—VIME (13), ICM (25), and Random Network Distillation (5)—are effective for exploration under sparse extrinsic reward. By analogy, curiosity-style signals can inform language evaluation by rewarding "useful novelty" (divergent yet coherent), complementing semantic-distance and rater-based methods. Our work instantiates this by modeling belief shifts when a language model incorporates expert explanations (a prediction-error–like signal) and combining it with expert attribution, yielding a more interpretable and *personalized* measure of creativity.

Bibliography

[1] Agresti, A.: Categorical Data Analysis. John Wiley & Sons, 3rd edn. (2013)

[2] Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)

[3] Bellemare, M.G., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., Munos, R.: Unifying count-based exploration and intrinsic motivation (2016), https://arxiv.org/abs/1606.01868

[4] Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. In: Noise reduction in speech processing, pp. 1–4. Springer (2009)

[5] Burda, Y., Edwards, H., Storkey, A., Klimov, O.: Exploration by random network distillation. arXiv preprint arXiv:1810.12894 (2018)

[6] Chakrabarty, T., Laban, P., Agarwal, D., Muresan, S., Wu, C.S.: Art or artifice? large language models and the false promise of creativity (2024), https://arxiv.org/abs/2309.14556

[7] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., Schulman, J.: Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021)

[8] Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)

[9] Dawid, A.P., Skene, A.: Maximum likelihood estimation of observer error-rates using the em algorithm. Journal of The Royal Statistical Society Series C-applied Statistics **28**, 20–28 (1979), https://api.semanticscholar.org/CorpusID:45813168

[10] DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z.F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J.L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R.J., Jin, R.L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S.S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W.L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X.Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y.K., Wang, Y.Q., Wei, Y.X.,

Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y.X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z.Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., Zhang, Z.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (2025), https://arxiv.org/abs/2501.12948

[11] Fisch, A., Eisenstein, J., Zayats, V., Agarwal, A., Beirami, A., Nagpal, C., Shaw, P., Berant, J.: Robust preference optimization through reward model distillation (2025), https://arxiv.org/abs/2405.19316

[12] Guilford, J.P.: Creativity: Yesterday, today and tomorrow. Journal of Creative Behavior **1**, 3–14 (1967), https://api.semanticscholar.org/CorpusID:143529843

[13] Houthooft, R., Chen, X., Duan, Y., Schulman, J., Turck, F.D., Abbeel, P.: Vime: Variational information maximizing exploration (2017), https://arxiv.org/abs/1605.09674

[14] Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., Hovy, E.: Learning whom to trust with MACE. In: Vanderwende, L., Daumé III, H., Kirchhoff, K. (eds.) Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1120–1130. Association for Computational Linguistics, Atlanta, Georgia (Jun 2013), https://aclanthology.org/N13-1132/

[15] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)

[16] Jiang, D., Ren, X., Lin, B.Y.: Llm-blender: Ensembling large language models with pairwise ranking and generative fusion (2023), https://arxiv.org/abs/2306.02561

[17] Madotto, A., Namazifar, M., Huizinga, J., Molino, P., Ecoffet, A., Zheng, H., Papangelis, A., Yu, D., Khatri, C., Tur, G.: Exploration based language learning for text-based games (2020), https://arxiv.org/abs/2001.08868

[18] Mostafazadeh Davani, A., Díaz, M., Prabhakaran, V.: Dealing with disagreements: Looking beyond the majority vote in subjective annotations. Transactions of the Association for Computational Linguistics **10**, 92–110 (2022). https://doi.org/10.1162/tacl_a_00449, https://aclanthology.org/2022.tacl-1.6/

[19] Nath, A., Jung, C., Seefried, E., Krishnaswamy, N.: Simultaneous reward distillation and preference learning: Get you a language model who can do both (2025), https://arxiv.org/abs/2410.08458

[20] OpenAI, :, Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., Iftimie, A., Karpenko, A., Passos, A.T., Neitz, A., Prokofiev, A., Wei, A., Tam, A., Bennett, A., Kumar, A., Saraiva, A., Vallone, A., Duberstein, A., Kondrich, A.,

Mishchenko, A., Applebaum, A., Jiang, A., Nair, A., Zoph, B., Ghorbani, B., Rossen, B., Sokolowsky, B., Barak, B., McGrew, B., Minaiev, B., Hao, B., Baker, B., Houghton, B., McKinzie, B., Eastman, B., Lugaresi, C., Bassin, C., Hudson, C., Li, C.M., de Bourcy, C., Voss, C., Shen, C., Zhang, C., Koch, C., Orsinger, C., Hesse, C., Fischer, C., Chan, C., Roberts, D., Kappler, D., Levy, D., Selsam, D., Dohan, D., Farhi, D., Mely, D., Robinson, D., Tsipras, D., Li, D., Oprica, D., Freeman, E., Zhang, E., Wong, E., Proehl, E., Cheung, E., Mitchell, E., Wallace, E., Ritter, E., Mays, E., Wang, F., Such, F.P., Raso, F., Leoni, F., Tsimpourlas, F., Song, F., von Lohmann, F., Sulit, F., Salmon, G., Parascandolo, G., Chabot, G., Zhao, G., Brockman, G., Leclerc, G., Salman, H., Bao, H., Sheng, H., Andrin, H., Bagherinezhad, H., Ren, H., Lightman, H., Chung, H.W., Kivlichan, I., O'Connell, I., Osband, I., Gilaberte, I.C., Akkaya, I., Kostrikov, I., Sutskever, I., Kofman, I., Pachocki, J., Lennon, J., Wei, J., Harb, J., Twore, J., Feng, J., Yu, J., Weng, J., Tang, J., Yu, J., Candela, J.Q., Palermo, J., Parish, J., Heidecke, J., Hallman, J., Rizzo, J., Gordon, J., Uesato, J., Ward, J., Huizinga, J., Wang, J., Chen, K., Xiao, K., Singhal, K., Nguyen, K., Cobbe, K., Shi, K., Wood, K., Rimbach, K., Gu-Lemberg, K., Liu, K., Lu, K., Stone, K., Yu, K., Ahmad, L., Yang, L., Liu, L., Maksin, L., Ho, L., Fedus, L., Weng, L., Li, L., McCallum, L., Held, L., Kuhn, L., Kondraciuk, L., Kaiser, L., Metz, L., Boyd, M., Trebacz, M., Joglekar, M., Chen, M., Tintor, M., Meyer, M., Jones, M., Kaufer, M., Schwarzer, M., Shah, M., Yatbaz, M., Guan, M.Y., Xu, M., Yan, M., Glaese, M., Chen, M., Lampe, M., Malek, M., Wang, M., Fradin, M., McClay, M., Pavlov, M., Wang, M., Wang, M., Murati, M., Bavarian, M., Rohaninejad, M., McAleese, N., Chowdhury, N., Chowdhury, N., Ryder, N., Tezak, N., Brown, N., Nachum, O., Boiko, O., Murk, O., Watkins, O., Chao, P., Ashbourne, P., Izmailov, P., Zhokhov, P., Dias, R., Arora, R., Lin, R., Lopes, R.G., Gaon, R., Miyara, R., Leike, R., Hwang, R., Garg, R., Brown, R., James, R., Shu, R., Cheu, R., Greene, R., Jain, S., Altman, S., Toizer, S., Toyer, S., Miserendino, S., Agarwal, S., Hernandez, S., Baker, S., McKinney, S., Yan, S., Zhao, S., Hu, S., Santurkar, S., Chaudhuri, S.R., Zhang, S., Fu, S., Papay, S., Lin, S., Balaji, S., Sanjeev, S., Sidor, S., Broda, T., Clark, A., Wang, T., Gordon, T., Sanders, T., Patwardhan, T., Sottiaux, T., Degry, T., Dimson, T., Zheng, T., Garipov, T., Stasi, T., Bansal, T., Creech, T., Peterson, T., Eloundou, T., Qi, V., Kosaraju, V., Monaco, V., Pong, V., Fomenko, V., Zheng, W., Zhou, W., McCabe, W., Zaremba, W., Dubois, Y., Lu, Y., Chen, Y., Cha, Y., Bai, Y., He, Y., Zhang, Y., Wang, Y., Shao, Z., Li, Z.: Openai o1 system card (2024), https://arxiv.org/abs/2412.16720

[21] Owen, A.B.: Monte carlo theory, methods and examples (2013)

[22] Pan, S.: Tiny reward models (2025), https://arxiv.org/abs/2507.09973

[23] Panickssery, A., Bowman, S., Feng, S.: Llm evaluators recognize and favor their own generations. Advances in Neural Information Processing Systems **37**, 68772–68802 (2024)

[24] Panickssery, A., Bowman, S.R., Feng, S.: Llm evaluators recognize and favor their own generations (2024), https://arxiv.org/abs/2404.13076

[25] Pathak, D., Agrawal, P., Efros, A.A., Darrell, T.: Curiosity-driven exploration by self-supervised prediction (2017), https://arxiv.org/abs/1705.05363
[26] Patterson, J.D., Barbot, B., Lloyd-Cox, J., Beaty, R.E.: Audra: An automated drawing assessment platform for evaluating creativity. Behavior Research Methods **56**(4), 3619–3636 (2024)
[27] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge (2015), https://arxiv.org/abs/1409.0575
[28] Schmidhuber, J.: Formal theory of creativity, fun, and intrinsic motivation. IEEE Trans. on Auton. Ment. Dev. **2**(3), 230–247 (Sep 2010). https://doi.org/10.1109/TAMD.2010.2056368, https://doi.org/10.1109/TAMD.2010.2056368
[29] Sidahmed, H., Phatale, S., Hutcheson, A., Lin, Z., Chen, Z., Yu, Z., Jin, J., Chaudhary, S., Komarytsia, R., Ahlheim, C., Zhu, Y., Li, B., Ganesh, S., Byrne, B., Hoffmann, J., Mansoor, H., Li, W., Rastogi, A., Dixon, L.: Parameter efficient reinforcement learning from human feedback (2024), https://arxiv.org/abs/2403.10704
[30] Torrance, E.P.: Torrance Tests of Creative Thinking: Norms–Technical Manual (Research Edition). Personnel Press, Princeton, NJ (1966)
[31] Vu, T., Krishna, K., Alzubi, S., Tar, C., Faruqui, M., Sung, Y.H.: Foundational autoraters: Taming large language models for better automatic evaluation (2024), https://arxiv.org/abs/2407.10817
[32] Wan, Y., Wu, J., Abdulhai, M., Shani, L., Jaques, N.: Enhancing personalized multi-turn dialogue with curiosity reward (2025), https://arxiv.org/abs/2504.03206
[33] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: A multi-task benchmark and analysis platform for natural language understanding (2019), https://arxiv.org/abs/1804.07461
[34] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models (2023), https://arxiv.org/abs/2203.11171
[35] Wataoka, K., Takahashi, T., Ri, R.: Self-preference bias in llm-as-a-judge (2025), https://arxiv.org/abs/2410.21819
[36] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (2023), https://arxiv.org/abs/2201.11903
[37] Whitehill, J., Wu, T.f., Bergsma, J., Movellan, J., Ruvolo, P.: Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., Culotta, A. (eds.) Advances in Neural Information Processing Systems. vol. 22. Curran Associates, Inc. (2009), https://proceedings.neurips.cc/paper_files/paper/2009/file/f899139df5e1059396431415e770c6dd-Paper.pdf
[38] Zar, J.H.: Spearman rank correlation. Encyclopedia of Biostatistics **7** (2005)

Table 4: Dimensions of TTCW dataset

Dimension	Facets
Fluency	Understandability & Coherence
	Narrative Pacing
	Scene vs Exposition
	Literary Devices & Language Proficiency
	Narrative Ending
Flexibility	Emotional Flexibility
	Perspective & Voice Flexibility
	Structural Flexibility
Originality	Originality in Form
	Originality in Thought
	Originality in Theme & Content
Elaboration	World Building & Setting
	Character Development
	Rhetorical Complexity

[39] Zhang, Y., Wang, L., Fang, M., Du, Y., Huang, C., Wang, J., Lin, Q., Pechenizkiy, M., Zhang, D., Rajmohan, S., Zhang, Q.: Distill not only data but also rewards: Can smaller language models surpass larger ones? (2025), https://arxiv.org/abs/2502.19557

[40] Zhao, Y., Zhang, R., Li, W., Li, L.: Assessing and understanding creativity in large language models. Machine Intelligence Research **22**(3), 417–436 (Apr 2025). https://doi.org/10.1007/s11633-025-1546-4, http://dx.doi.org/10.1007/s11633-025-1546-4

[41] Zhong, S., Huang, Z., Gao, S., Wen, W., Lin, L., Zitnik, M., Zhou, P.: Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13246–13257 (June 2024)

A Appendix

A.1 Dimensions in dataset

In Table 4, all the dimensions that are part of the TTCW dataset are mentioned.

A.2 More experiment and compute details

A.3 Limitations

Our study has some limitations that we hope to address in future work. First, the empirical scope is narrow: we evaluate only on TTCW dataset. Our current

Table 5: Core hyperparameters used in all runs.

`max_length`	4096
`lora_dropout`	0.1
`target_modules`	`["q_proj","k_proj","v_proj","o_proj","gate_proj","up_proj","down_proj"]`
`lr_scheduler`	cosine (warmup_ratio = 0.1)
`per_device_train_batch_size`	4
`gradient_accumulation_steps`	8
`weight_decay`	0.01
`max_grad_norm`	0.5
`num_train_epochs`	3
`seed`	42

method is text-only; extending to richer modalities and subjective tasks beyond TTCW remains future work. In addition, the dataset is small (48 stories × 5 dimensions with three expert judgments per story–dimension, totaling 720 instances). We therefore rely on 5-fold cross-validation and report means and deviation across 5 folds. Finally, model coverage is limited to one family (Qwen2.5 0.5B–7B), leaving generalization across architectures untested, which we aim to do in future work.

A.4 Question for each dimension

Table 6: Creativity evaluation categories and questions

Category	**Question**
Originality in Thought	Is the story an original piece of writing without any cliches?
Originality in Form and Structure	Does the story show originality in its form and/or structure?
Originality in Theme and Content	Will an average reader of this story obtain a unique and original idea from reading it?
Perspective and Voice Flexibility	Does the story provide diverse perspectives, and if there are unlikeable characters, are their perspectives presented convincingly and accurately?
Structural Flexibility	Does the story contain turns that are both surprising and appropriate?

A.5 Statistical significance testing

Table 7: Statistical significance test across 5 folds for Qwen-0.5b model

Metric	SFT(with expl) (mean±SD)	ICM (mean±SD)	Δ (ICM−SFT)	p (paired t)	Statistically significant?
Pearson	0.160 ± 0.055	0.524 ± 0.092	0.364	0.002	Yes
Spearman	0.160 ± 0.055	0.484 ± 0.078	0.324	<0.001	Yes
F1	0.371 ± 0.054	0.616 ± 0.048	0.245	<0.001	Yes

Table 8: Statistical significance test across 5 folds for Qwen-1.5b model

Metric	SFT(with expl) (mean±SD)	ICM (mean±SD)	Δ (ICM−SFT)	p (paired t)	Statistically significant?
Pearson	0.170 ± 0.058	0.586 ± 0.064	0.416	<0.001	Yes
Spearman	0.170 ± 0.058	0.522 ± 0.069	0.352	<0.001	Yes
F1	0.402 ± 0.050	0.629 ± 0.045	0.227	<0.001	Yes

Table 9: Statistical significance test across 5 folds for Qwen-3b model.

Metric	SFT(with expl) (mean±SD)	ICM (mean±SD)	Δ (ICM−SFT)	p (paired t)	Statistically significant?
Pearson	0.113 ± 0.092	0.540 ± 0.074	0.427	<0.001	Yes
Spearman	0.113 ± 0.092	0.494 ± 0.091	0.381	<0.001	Yes
F1	0.339 ± 0.053	0.618 ± 0.061	0.279	<0.001	Yes

Table 10: Statistical significance test across 5 folds for Qwen-7b model.

Metric	SFT(with expl) (mean±SD)	ICM (mean±SD)	Δ (ICM−SFT)	p (paired t)	Statistically significant?
Pearson	0.170 ± 0.058	0.606 ± 0.084	0.436	<0.001	Yes
Spearman	0.170 ± 0.058	0.542 ± 0.089	0.373	<0.001	Yes
F1	0.381 ± 0.029	0.663 ± 0.058	0.282	<0.001	Yes

Table 11: Average passing rate (%) on individual TTCW, based on annotations of 10 creative writing experts across 48 stories; last column reports Fleiss' κ (expert agreement).

Dimension	Test	GPT-3.5	GPT-4	Claude v1.3	New Yorker	Expert κ
Fluency	Understandability & Coherence	22.2	33.3	55.6	91.7	0.27
	Narrative Pacing	8.3	52.8	61.1	94.4	0.39
	Scene vs Exposition	8.3	50.0	58.3	91.7	0.27
	Literary Devices & Language	5.6	36.1	13.9	88.9	0.37
	Narrative Ending	8.3	19.4	33.3	91.7	0.48
Flexibility	Emotional Flexibility	16.7	19.4	36.1	91.7	0.32
	Perspective & Voice Flexibility	8.3	16.7	19.4	72.2	0.44
	Structural Flexibility	11.1	19.4	30.6	88.9	0.39
Originality	Originality in Form	2.8	8.3	0.0	63.9	0.41
	Originality in Thought	2.8	44.4	19.4	91.7	0.40
	Originality in Theme & Content	0.0	19.4	11.1	75.0	0.66
Elaboration	World Building & Setting	16.7	41.7	58.3	94.4	0.33
	Character Development	8.3	16.7	16.7	61.1	0.31
	Rhetorical Complexity	2.8	11.1	5.6	88.9	0.66
Average		**8.7**	**27.9**	**30.0**	**84.7**	**0.41**

Table 12: Correlation between LLM-administered TTCW and expert annotations (Cohen's κ) on all 48 stories.

Dimension	Test	GPT-3.5	GPT-4	Claude
Fluency	Understandability & Coherence	-0.01	-0.01	-0.17
	Narrative Pacing	0.05	0.00	-0.22
	Scene vs Exposition	-0.03	-0.08	-0.23
	Literary Devices & Language	0.04	-0.09	-0.11
	Narrative Ending	-0.02	0.02	0.02
Flexibility	Emotional Flexibility	-0.04	0.00	0.09
	Perspective & Voice	0.00	0.26	0.14
	Structural Flexibility	-0.04	0.00	-0.07
Originality	Originality in Form	0.08	0.09	0.03
	Originality in Thought	0.19	0.31	0.15
	Originality in Theme & Content	0.06	-0.01	0.18
Elaboration	World Building & Setting	0.00	0.00	0.09
	Character Development	-0.08	0.02	0.00
	Rhetorical Complexity	0.00	0.00	0.02
Average		**0.016**	**0.035**	**-0.006**

A.6 ICM results against SFT baseline without explanations

Table 13: ICM method results against the SFT baseline without explanations (classification). Means±SD are shown where SD was available from 5-fold runs.

Model	Exp. type	Pearson	Precision	Recall	F1
Qwen-0.5B (SFT-Cls)	ID	**0.586 ±0.085**	**0.769**	0.461	0.551 ±0.198
Qwen-0.5B (ICM)	ID	0.524 ±0.092	0.494	**0.818**	**0.616 ±0.048**
Qwen-1.5B (SFT-Cls)	ID	**0.602 ±0.064**	**0.787**	0.602	**0.663 ±0.070**
Qwen-1.5B (ICM)	ID	0.586 ±0.064	0.481	**0.794**	0.629 ±0.045
Qwen-3B (SFT-Cls)	ID	0.482 ±0.160	**0.670**	0.573	0.556 ±0.094
Qwen-3B (ICM)	ID	**0.540 ±0.074**	0.481	**0.794**	**0.618 ±0.061**
Qwen-7B (SFT-Cls)	ID	0.441 ±0.130	**0.535**	0.342	0.383 ±0.251
Qwen-7B (ICM)	ID	**0.606 ±0.084**	0.518	**0.850**	**0.663 ±0.058**

Table 14: ICM method results against the SFT baseline without explanations(classification) on Out-of-distribution data

Model	Experiment type	pearson	precision	recall	f1
Qwen-0.5B(SFT-Classifcation)	OOD	0.433	0.000	0.000	0.000
Qwen-0.5B(ICM)	OOD	**0.563**	**0.625**	**0.790**	**0.698**
Qwen-1.5B(SFT-Classifcation)	OOD	0.604	**0.962**	0.439	0.602
Qwen-1.5B(ICM)	OOD	**0.655**	0.639	**0.807**	**0.713**
Qwen-3B(SFT-Classifcation)	OOD	0.546	**0.933**	0.246	0.389
Qwen-3B(ICM)	OOD	**0.582**	0.597	**0.754**	**0.667**
Qwen-7B(SFT-Classifcation)	OOD	0.435	0.800	0.211	0.333
Qwen-7B(ICM)	OOD	**0.623**	**0.653**	**0.825**	**0.729**

A.7 Curiosity scores based on non-finetuned base Qwen-0.5B model's prediction and ground truth match and mismatch

A.8 Why is inverse model necessary?

When we ablated for the inverse model in our ICM setup with the given expert annotated data we do not see any difference in the results with using the inverse model or without using it. But the inverse model becomes necessary when we have a non-expert annotator like GPT-2, since it helps to clearly distinguish such outliers. This shows that our forward model of the ICM is good enough to distinguish between multiple expert annotators but we do need the inverse model for outlier cases. The details of our experiments can be found in Table 15, we used Qwen-0.5B model for this experiment.

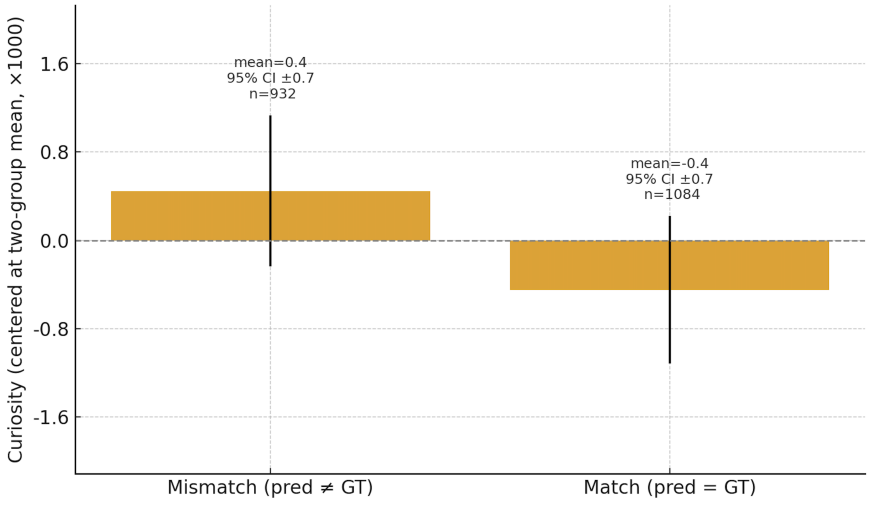


Fig. 5: Curiosity scores based on match and mismatch of predictions from Qwen-0.5B base non-finetuned model and the ground truth

Table 15: Inverse model ablations

Method	**Annotations**	**Pearson**	**Precision**	**Recall**	**F1**	**Cohen's κ**
ICM with Inverse	Without GPT-2	0.503 $\pm$0.014	0.552 $\pm$0.014	0.728 $\pm$0.017	0.628 $\pm$0.015	0.347 $\pm$0.027
ICM without Inverse	Without GPT-2	0.500 $\pm$0.027	0.551 $\pm$0.011	0.727 $\pm$0.009	0.627 $\pm$0.010	0.346 $\pm$0.017
ICM with Inverse	With GPT-2	0.151 $\pm$0.300	0.153 $\pm$0.265	0.233 $\pm$0.403	0.185 $\pm$0.320	0.093 $\pm$0.166
ICM without Inverse	With GPT-2	0.002 $\pm$0.041	0.333 $\pm$0.577	0.001 $\pm$0.002	0.002 $\pm$0.004	0.000 $\pm$0.004