The Geometry of Creative Variability: How Credal Sets Expose Calibration Gaps in Language Models

Esteban Garces Arias[1,2], Julian Rodemann[1,3], Christian Heumann[1]

[1]Department of Statistics, LMU Munich, Germany
[2]Munich Center for Machine Learning (MCML)
[3]CISPA Helmholtz Center for Information Security, Saarbrücken, Germany
Correspondence: Esteban.GarcesArias@stat.uni-muenchen.de

Abstract

Understanding uncertainty in large language models remains a fundamental challenge, particularly in creative tasks where multiple valid outputs exist. We present a geometric framework using credal sets—convex hulls of probability distributions—to quantify and decompose uncertainty in neural text generation, calibrated against human creative variation. Analyzing 500 creative writing prompts from the WRITINGPROMPTS dataset with 10 unique human continuations each, we evaluate four language models across five decoding strategies, generating 100,000 stories. Our credal set analysis reveals substantial gaps in capturing human creative variation, with the best model-human calibration reaching only 0.434 (Gemma-2B with temperature 0.7). We decompose total uncertainty into *epistemic* and *aleatoric* components, finding that the choice of decoding strategy contributes 39.4% to 72.0% of total epistemic uncertainty. Model scale shows weak correlation with calibration quality and no significant difference exists between base and instruction-tuned models in calibration quality. Our geometric framework provides actionable insights for improving generation systems for human-AI creative alignment. We release our complete experimental framework at `https://github.com/EstebanGarces/uncertainHuman`.

1 Introduction

The deployment of large language models in creative and open-ended applications demands not merely generating plausible text, but understanding and calibrating the uncertainty inherent in these generations. While uncertainty quantification has been extensively studied in discriminative tasks (Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017; Ovadia et al., 2019), the challenge becomes substantially more complex in generative settings where no single ground truth exists and quality itself becomes a multidimensional construct (Garces Arias et al., 2025b,c). This complexity is particularly acute in creative writing, where the same prompt can inspire substantially different narratives, styles, and interpretations (cf. Figure 1).

Prompt: *"The last person on Earth sits alone. There is a knock on the door."*

Human continuations:
- "My heart stopped. After three years of silence..."
- "I laughed. The universe's final joke..."
- "Pizza delivery,' a voice called out..."

Model continuations (Instruct):
- "The survivor cautiously approached the door..."
- "They slowly walked to the door, heart pounding..."
- "With trembling hands, the survivor reached..."

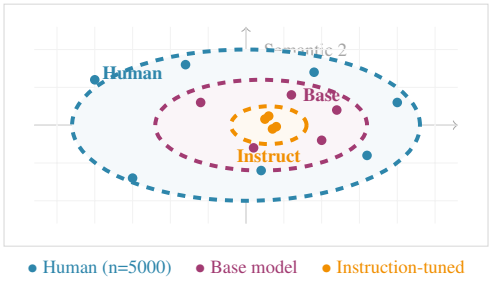

Figure 1: Examples of human versus model creative variation. **Top:** Continuations show diverse human interpretations versus convergent model responses. **Bottom:** Credal sets (dashed ellipses) represent convex hulls of diversity distributions in semantic, lexical, and syntactic space.

Current approaches to uncertainty quantification in language models predominantly focus on token-level probabilities or computationally expensive ensemble methods (Ling et al., 2024; Zhang et al., 2025). These methods, while valuable, fail to capture the semantic, lexical, and syntactic-level uncertainty that determines whether a model appropriately captures the breadth of human creative expression. More fundamentally, existing frameworks lack principled methods for distinguishing between

Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025), pages 36–50
November 9, 2025

aleatoric uncertainty—the irreducible variation inherent in creative tasks—and *epistemic uncertainty* arising from model limitations. This distinction proves crucial for both improving model design and establishing appropriate deployment boundaries. In this work, we address these limitations through a novel framework that leverages human variation as a natural calibration target for model uncertainty. Our key insight is that multiple human responses to the same creative prompt provide a direct empirical measure of aleatoric uncertainty. By representing both human and model variation as credal sets—convex hulls of probability distributions over textual characteristics—we can geometrically analyze whether models exhibit appropriate uncertainty: high variation when humans disagree, and convergent outputs when humans reach consensus. This credal set approach offers several theoretical and practical advantages over existing methods. Theoretically, it provides a rigorous framework for uncertainty decomposition that respects the inherently distributional nature of creative variation. Each prompt induces its own distribution over possible continuations, and the collection of these distributions across many prompts forms a credal set that fully characterizes the uncertainty landscape. Practically, this framework enables direct comparison between human and the model's uncertainty through geometric measures such as overlap coefficients, Hausdorff distance (Huttenlocher et al., 1993), and volume ratios. Our empirical investigation analyzes 500 carefully selected prompts from the WRITINGPROMPTS dataset, each accompanied by 10 verified unique human continuations totaling 5,000 human-written stories. We evaluate four language models—GPT2-XL (Radford et al., 2019), Gemma-2B (Gemma-Team et al., 2024), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), and Llama-3.1-8B-Instruct (Dubey et al., 2024)—generating 10 samples per configuration across five decoding strategies, yielding 100,000 model-generated stories. Through comprehensive analysis of semantic, lexical, and syntactic diversity, we construct and compare credal sets that reveal systematic patterns in how models capture or fail to capture human-like variation.

Contributions

- We introduce credal sets—convex hulls of diversity distributions—as a geometric framework for quantifying uncertainty in open-ended text generation.
- We analyze 100,000 generated stories, finding that the best model-human calibration reaches only 0.434 (Gemma-2B with temperature 0.7), revealing substantial gaps in creative variation.
- We show weak correlation between model scale and calibration (Spearman's $\rho = 0.400$, $p = 0.600$) and no significant difference between base and instruction-tuned models ($t = -0.712$, $p = 0.486$).
- We decompose uncertainty to reveal that decoding strategy choice contributes 39.4-72.0% of total epistemic uncertainty, with base models showing higher sensitivity.
- We release our complete experimental framework and datasets for reproducible research.[1]

2 Related Work

Uncertainty quantification in language models has emerged as a critical research area, particularly as these models are deployed in high-stakes applications. We organize our discussion around three main themes: theoretical frameworks for uncertainty decomposition, practical estimation methods, and uncertainty-aware generation strategies.

2.1 Theoretical Frameworks for Uncertainty Decomposition

The foundational challenge lies in decomposing total predictive uncertainty into meaningful components. Ling et al. (2024) address this for in-context learning scenarios: They derive total predictive uncertainty through the classical additive information-theoretic decomposition, where the first term captures aleatoric uncertainty (inherent randomness in the task) and the second represents epistemic uncertainty (model uncertainty). They propose entropy estimators based on variational bounds on mutual information for practical approximation. However, Wimmer et al. (2023) note that this distinction can become ambiguous in pre-trained models where the training distribution itself is uncertain.

The use of credal sets for uncertainty representation has been common in classification tasks (Zaffalon and Fagiuoli, 2003), but, to the best of our knowledge, our work is the first to apply this framework to open-ended generation. Credal sets provide

[1] https://github.com/EstebanGarces/uncertainHuman

a natural representation for situations where a single probability distribution is insufficient to capture uncertainty, instead maintaining a set of plausible distributions (Levi, 1980).

2.2 Practical Methods for Uncertainty Estimation

Various practical approaches for uncertainty estimation have been recently proposed: Lin et al. (2022); Xiong et al. (2024) explore methods to verbalize uncertainty, Kadavath et al. (2022); Liu et al. (2024); Ulmer et al. (2024) focus on probes for LLM calibration, while Pitis et al. (2023); Hou et al. (2024) have focused on self-consistency approaches.
Recent work has developed various approaches to estimate uncertainty without expensive ensemble methods. Zhang et al. (2025) introduce a training-free method injecting low-rank random weight perturbations during decoding to estimate token-level uncertainties. These are aggregated into sequence-level measures that correlate strongly with correctness on mathematical reasoning benchmarks, with epistemic uncertainty effectively identifying incorrect reasoning paths. While this perturbation approach elegantly estimates model uncertainty, it focuses on uncertainty from a single fixed model. Our work examines uncertainty arising from different decoding strategies and model architectures, providing a complementary perspective on variation sources in language model outputs. Yadkori et al. (2024) propose an information-theoretic metric based on mutual information over iteratively prompted responses, interpreting heavy dependencies between subsequent responses as indicators of high epistemic uncertainty and potential hallucination, though requiring computationally expensive multiple inference passes. Aichberger et al. (2024) pursue efficiency with a single-pass approximation using negative log-likelihood of greedy outputs, proving that high NLL correlates with high epistemic uncertainty under certain assumptions.

2.3 Uncertainty-Aware Generation and Human Baselines

Garces Arias et al. (2024); Ding et al. (2025) propose uncertainty-aware decoding that dynamically adjusts generation parameters based on local uncertainty. They compute entropy $H(p_t)$ of the token probability distribution p_t at each generation step t and adjust the truncation threshold dynamically, demonstrating that uncertainty signals can improve generation quality in real-time. Most directly related to our work, Giulianelli et al. (2023) evaluate uncertainty in neural text generators against human production variability, arguing that well-calibrated models should exhibit similar variation to humans. They analyze GPT-2 on story generation with limited prompts, finding that it under-produces diversity relative to human baselines. Our work substantially extends this research by: (1) scaling to 500 prompts with 10 unique continuations each, (2) including contemporary instruction-tuned models, (3) evaluating five decoding strategies systematically, (4) explicitly decomposing uncertainty into aleatoric and epistemic components, and (5) providing quantitative calibration metrics based on credal set overlap coefficients.

3 Methodology

3.1 Dataset Construction and Human Baselines

The WRITINGPROMPTS dataset (Fan et al., 2018) provides naturalistic creative writing data from Reddit's r/WritingPrompts community. We implement rigorous selection criteria to ensure data quality:

1. **Uniqueness verification**: We compute MD5 hashes for all stories and select only prompts with exactly 10 unique continuations, eliminating duplicates that could bias diversity measurements.

2. **Length filtering**: We retain prompts between 20-500 characters and stories between 52-987 tokens (mean: 312.4, std: 148.2), ensuring sufficient content for meaningful analysis while avoiding outliers.

3. **Quality scoring**: We prioritize prompts by the diversity of story lengths they elicit (measured by standard deviation), selecting those that inspire varied responses rather than formulaic continuations.

This process yields 500 high-quality prompts with 5,000 unique human stories, providing a robust baseline for calibration analysis.

3.2 Model Selection and Configuration

Our model selection explores the calibration landscape across different architectures and training paradigms:

Base models: GPT2-XL (1.5B) (Radford et al., 2019) serves as a canonical autoregressive baseline, while Gemma-2B (Gemma-Team et al., 2024) represents modern architectural improvements at comparable scale. These models, trained on diverse internet text without explicit instruction following, potentially preserve more natural variation patterns.

Instruction-tuned models: Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) and Llama-3.1-8B-Instruct (Dubey et al., 2024) represent strong open-source models with instruction tuning and alignment. While offering improved controllability, we investigate whether alignment training constrains creative exploration[2].

3.3 Decoding Strategy Design

We systematically evaluate five decoding strategies that control output diversity through different mechanisms:

- **Temperature scaling** ($\tau \in \{0.7, 1.2\}$): Directly modulates the entropy of the output distribution (Ackley et al., 1985)
- **Nucleus sampling** ($p = 0.9$): Dynamically adjusts the token consideration set based on cumulative probability (Holtzman et al., 2020)
- **Top-k sampling** ($k = 40$): Maintains a fixed-size token pool (Fan et al., 2018)
- **Typical sampling** ($p = 0.95$): Selects tokens based on expected information content (Meister et al., 2023)

Each configuration generates 10 independent samples with different random seeds, totaling 100,000 model-generated stories for analysis.

3.4 Diversity Metrics

Our metric suite captures multiple dimensions of textual variation through pairwise distance-based measures following Giulianelli et al. (2023):

3.4.1 Semantic Diversity

We compute semantic diversity as the mean pairwise cosine distance between Sentence-BERT embeddings (Reimers and Gurevych, 2019):

$$D_{\text{sem}}(\mathcal{S}) = \frac{2}{|\mathcal{S}|(|\mathcal{S}|-1)} \sum_{i<j} (1 - \cos(e_i, e_j)),$$

where e_i represents the embedding of story i using the all-MiniLM-L6-v2 model (Wang et al., 2020). This captures high-level narrative and thematic variation.

3.4.2 Lexical Diversity

We measure lexical diversity using Jaccard distance between word unigrams:

$$D_{\text{lex}}(\mathcal{S}) = \frac{2}{|\mathcal{S}|(|\mathcal{S}|-1)} \sum_{i<j} \left(1 - \frac{|V_i \cap V_j|}{|V_i \cup V_j|}\right),$$

where V_i represents the vocabulary set of story i. This captures variation in word choice and vocabulary richness.

3.4.3 Syntactic Diversity

We measure syntactic variation through Jaccard distance of part-of-speech (POS) bigrams:

$$D_{\text{syn}}(\mathcal{S}) = \frac{2}{|\mathcal{S}|(|\mathcal{S}|-1)} \sum_{i<j} \left(1 - \frac{|P_i \cap P_j|}{|P_i \cup P_j|}\right),$$

where P_i represents the set of POS bigrams extracted using spaCy's en_core_web_sm model (Honnibal and Montani, 2017). This captures stylistic and structural variation in the generated text.

3.5 Theoretical Framework: Credal Sets

Our methodology rests on the principle that uncertainty in creative text generation should be understood relative to the natural variation exhibited by humans facing the same creative task. We formalize this through a credal set framework that captures uncertainty as a set of plausible probability distributions rather than a single distribution.

For a given prompt p, let $\mathcal{H}_p = \{h_1, ..., h_{10}\}$ denote the set of human continuations and $\mathcal{M}_{p,m,d} = \{s_1, ..., s_{10}\}$ denote the set of model continuations for model m using decoding strategy d. For any set of continuations $\mathcal{S}$, we compute a diversity vector $\mathbf{v}_p = [D_{\text{sem}}(\mathcal{S}), D_{\text{lex}}(\mathcal{S}), D_{\text{syn}}(\mathcal{S})]$.

The human credal set for a collection of prompts $\mathcal{P}$ is then defined as:

$$\mathcal{C}_H = \text{ConvexHull}\left(\{\mathbf{v}_p^H : p \in \mathcal{P}\}\right),$$

where each $\mathbf{v}_p^H$ is the diversity vector computed from human continuations for prompt p.
Similarly, the model credal set for a specific configuration (m, d) is:

$$\mathcal{C}_{M,d} = \text{ConvexHull}\left(\{\mathbf{v}_p^{M,d} : p \in \mathcal{P}\}\right).$$

[2]All models use appropriate prompt formatting with careful post-processing to remove prompt artifacts from generations, ensuring fair comparison across architectures.

The convex hull is computed using the Quickhull algorithm (Barber et al., 1996) after standardizing the diversity vectors. This representation enables geometric analysis of uncertainty relationships through set operations and distance metrics.

3.6 Calibration Analysis

Calibration quality is assessed through the overlap coefficient of credal sets:

$$\text{Calibration}(M, d) = \text{Overlap}(\mathcal{C}_H, \mathcal{C}_{M,d}),$$

where overlap is computed using nearest-neighbor distances between credal set vertices. The overlap coefficient is calculated as:

$$\text{Overlap} = \frac{1}{2}\left(\frac{|\{v \in V_M : d(v, V_H) < \theta\}|}{|V_M|} + \frac{|\{v \in V_H : d(v, V_M) < \theta\}|}{|V_H|}\right),$$

where V_M and V_H are the vertex sets of the model and human credal sets respectively, $d(v, V)$ is the minimum distance from point v to set V, and θ is an adaptive threshold set to half the mean variance scale. Values range from 0 (disjoint sets) to 1 (perfect overlap).

3.7 Uncertainty Decomposition

To decompose uncertainty, we leverage variation across decoding strategies. For a given model M, we collect all diversity vectors across different strategies and compute:

Strategy centroids :
$\mathbf{c}_d = \text{mean}(\{\mathbf{v}_p^{M,d} : p \in \mathcal{P}\})$ for each strategy d

Between-strategy variance :
$\sigma^2_{\text{between}} = \text{Var}(\{\mathbf{c}_d : d \in \mathcal{D}\})$

Within-strategy variance :
$\sigma^2_{\text{within}} = \text{mean}_d[\text{Var}(\{\mathbf{v}_p^{M,d} : p \in \mathcal{P}\})]$
The epistemic ratio is then:

$$\text{Epistemic}_M = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}}.$$

This quantifies the proportion of uncertainty arising from configuration choices rather than inherent task ambiguity.

4 Results

4.1 Human Variation as Calibration Baseline

Analysis of 5,000 human-written stories reveals structured patterns of creative variation that establish our calibration baseline (Table 1).

Diversity Type	Mean	Std Dev
Semantic	0.645	0.066
Lexical	0.328	0.035
Syntactic	0.315	0.044

Table 1: Human diversity baselines across 500 prompts with 10 unique continuations each, computed using pairwise distance metrics.

The distribution of semantic diversity across prompts shows moderate variation with most prompts (62%) eliciting medium diversity (0.6-0.7), while 19% show high diversity (>0.7) and 19% show low diversity (<0.6). This suggests fundamental differences in prompt interpretability that models must capture.

4.2 Credal Set Geometry and Calibration

The human credal set $\mathcal{C}_H$ occupies a volume of 2.25 in the PCA-transformed diversity space, serving as the baseline for model comparison. Analysis reveals a clear distinction between model types: base models (GPT2-XL, Gemma-2B) produce compact credal sets with mean volume 1.10 ± 0.56, representing 48.9% of the human volume. In contrast, instruction-tuned models (Mistral-7B-Instruct, Llama-3.1-8B-Instruct) generate significantly larger credal sets with mean volume 3.87 ± 1.78, corresponding to 172.1% of the human baseline. The difference in credal set volumes between base and instruction-tuned models is statistically significant (Mann-Whitney U = 2.00, $p < 0.001$).

Principal component analysis of the diversity vectors reveals strong coupling between diversity dimensions. PC1 explains 85.8% of variance with nearly equal positive loadings across semantic (0.569), lexical (0.565), and syntactic (0.597) dimensions, indicating that these diversity types covary systematically. The dominance of PC1 suggests that models exhibiting high diversity in one dimension tend to show proportionally high diversity in all dimensions, as illustrated in Figure 3.

The expanded credal sets of instruction-tuned models indicate broader exploration of the diversity space compared to base models. However, larger volume does not directly correspond to better calibration, as shown in Table 2 and Figure 6, in the Appendix. This suggests that alignment with human diversity patterns depends more on the location and shape of the credal set than its absolute size.

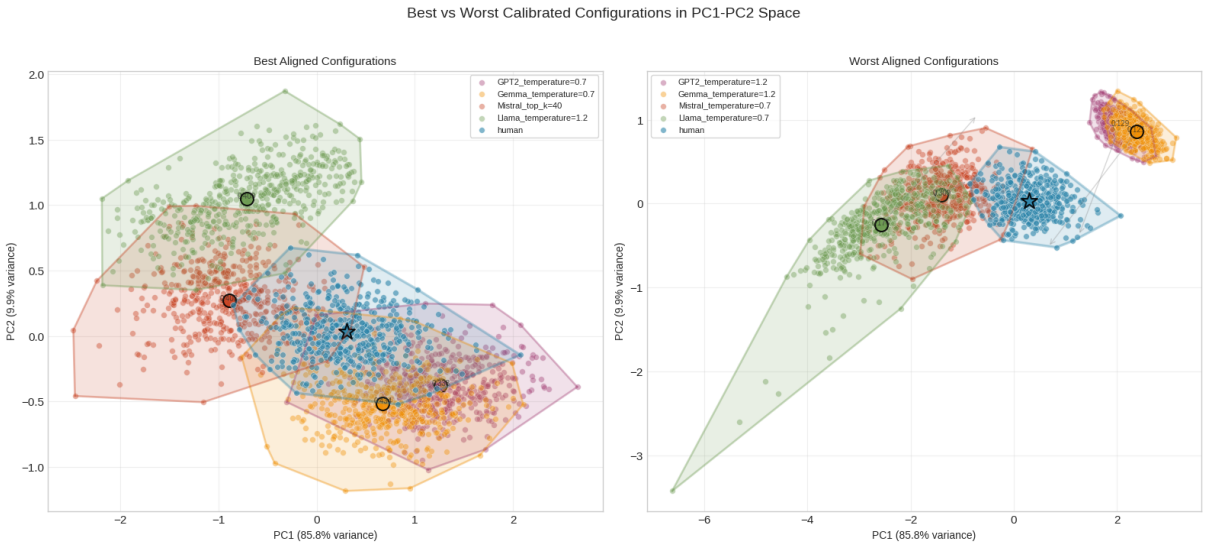

Figure 2: Credal sets visualization in principal component space. Human creative variation (blue) and model-generated variation exhibit different geometric patterns and a high sensitivity with respect to the decoding configuration. Points represent diversity vectors from individual prompts; convex hulls indicate credal set boundaries. PC1 explains 85.8% of the variance, suggesting a strong correlation between diversity dimensions. Best (left) and worst aligned configurations (right), measured by the overlap of the credal sets, are presented.

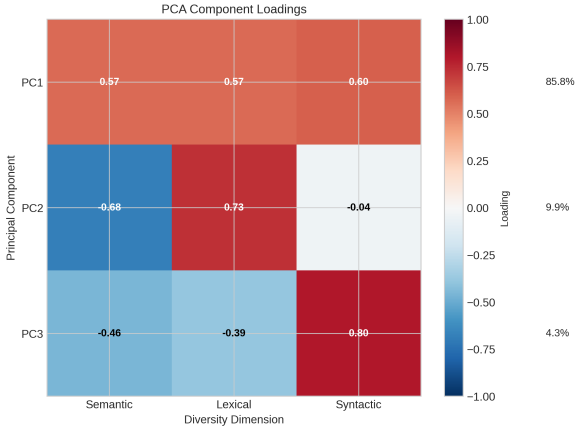

Figure 3: Overview of PCA loadings, displaying a balanced contribution of semantic, lexical, and syntactic patterns on the first principal component, which explains a large proportion of the total variance.

4.3 Distributional Analysis via Wasserstein Distance

Complementary analysis using Wasserstein distance at the prompt level corroborates the credal set findings. The Wasserstein distance measures the average distributional difference between human and model-generated diversity patterns across all prompts. The best configuration by Wasserstein distance (Gemma-2B with temperature=0.7, distance=0.065) coincides with the best-calibrated credal set, providing independent validation of the geometric approach. The moderate negative correlation between Wasserstein distance and calibration score ($\rho = -0.411$, $p = 0.072$) indicates that while both methods capture aspects of human-model alignment, they emphasize different characteristics: Wasserstein distance weights all prompts equally in measuring average distributional differences, while credal sets capture the geometric envelope of diversity behaviors. A visualization of this comparison is presented in Figure 7.

4.4 Model Calibration Patterns

Calibration analysis reveals that no model effectively reproduces human variation patterns, with best overlap coefficients reaching only 0.434 (Table 2). Figure 5 illustrates these key findings:

Model architecture effects: Gemma-2B achieves the best single configuration calibration (0.434 with temperature 0.7), though Mistral-7B-Instruct shows the highest average calibration across all strategies (0.371). Statistical analysis reveals weak positive correlation between model size and calibration (Spearman's $\rho = 0.400$, $p = 0.600$), suggesting model scale has limited influence on calibration quality. Further, base models (mean calibration: 0.274 ± 0.095) show no significant difference from instruction-tuned models (mean: 0.305 ± 0.093) in calibration quality ($t = -0.712$, $p = 0.486$, Cohen's $d = -0.336$). Despite similar calibration scores, base and

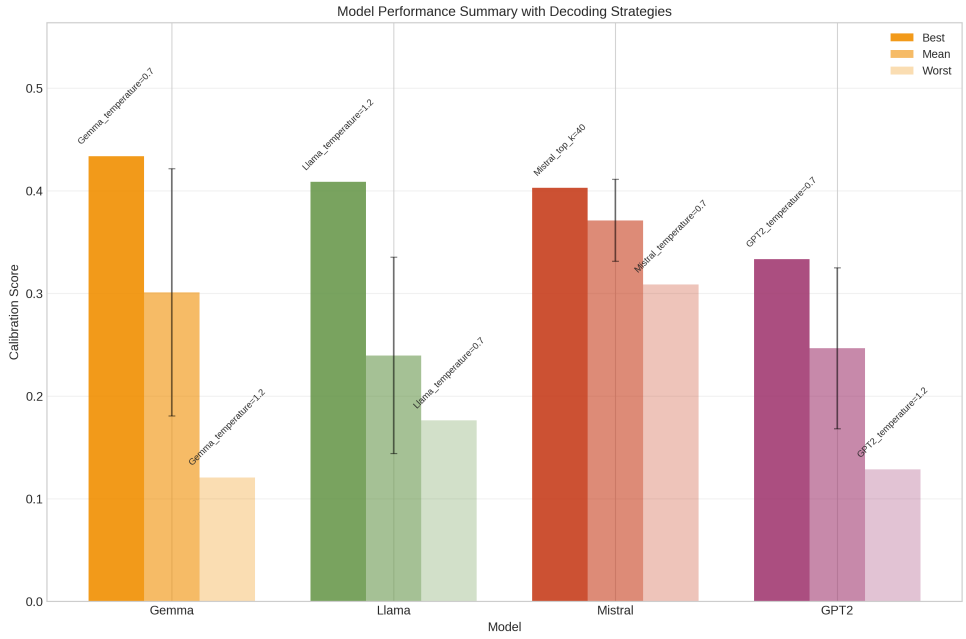

Figure 4: Overview of model performance across varying decoding strategies. Here, performance is to be understood in terms of calibration scores with respect to human credal sets. Top-k sampling provides the highest mean calibration, while Gemma-2B with temperature set to 0.7 achieves the best overall calibration.

Model	Strategy	Value	Overall Cal.	Overlap	Centroid Dist.	Volume Ratio
Gemma-2B	temperature	0.7	**0.434**	0.033	1.096	0.924
Llama-3.1-8B-Instruct	temperature	1.2	0.409	0.032	1.488	0.918
Mistral-7B-Instruct	top_k	40	0.403	0.000	1.502	1.060
Mistral-7B-Instruct	temperature	1.2	0.399	0.000	0.956	0.820
Mistral-7B-Instruct	top_p	0.9	0.391	0.000	1.721	1.070
Gemma-2B	top_k	40	0.386	0.033	1.189	0.785
Mistral-7B-Instruct	typical	0.95	0.354	0.000	1.604	1.258
GPT2-XL	temperature	0.7	0.333	0.000	1.386	0.692
Mistral-7B-Instruct	temperature	0.7	0.309	0.000	1.945	1.450
GPT2-XL	top_k	40	0.300	0.033	1.244	0.509

Table 2: Calibration metrics for top configurations. Higher values indicate better alignment with human variation. Gemma-2B with temperature 0.7 achieves best overall calibration (0.434).

instruction-tuned models differ significantly in their exploration of the diversity space, with instruction-tuned models producing credal sets 3.5× larger on average ($p < 0.001$).

Strategy effectiveness: Top-k sampling achieves the highest mean performance (0.323 ± 0.092), followed by temperature scaling (0.289 ± 0.129). Analysis of variance across all 20 model-strategy combinations reveals no significant main effect of strategy type ($F(3, 16) = 0.200$, $p = 0.895$), suggesting that strategy effectiveness depends on the specific model architecture.

4.5 Uncertainty Decomposition

Decomposition analysis reveals the relative contributions of epistemic and aleatoric uncertainty (Table 3). Base models (GPT2-XL, Gemma-2B) exhibit higher epistemic ratios (64.9-72.0%), indicating that decoding strategy choice contributes more than half of their total uncertainty. Instruction-tuned models show lower epistemic ratios (39.4-50.5%), suggesting more stable behavior across decoding strategies but potentially at the cost of reduced overall variation.

The within-strategy variance (aleatoric component) remains substantial across all models (0.091-0.224), confirming that models can generate diverse outputs for individual prompts. However, the between-strategy variance (epistemic component) highlights that generation configuration remains a critical factor in uncertainty quantification, particularly for base models.

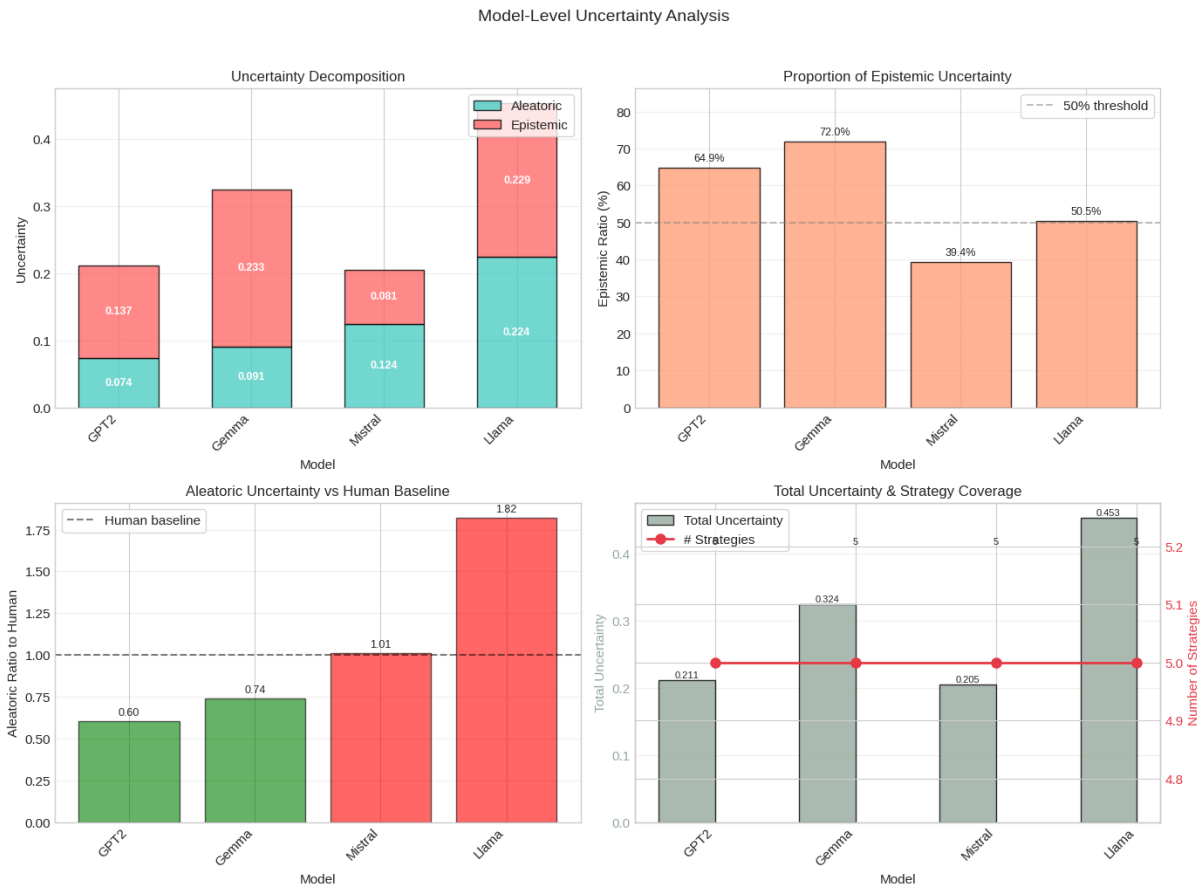

Figure 5: Uncertainty analysis and model performance overview. **Top Left:** Uncertainty decomposition showing epistemic and aleatoric components. **Top Right:** Epistemic ratio by model. **Bottom Left:** Aleatoric uncertainty vs. human baseline. **Bottom Right:** Estimated total uncertainty per model, measured over five decoding strategies.

Model	Epistemic	Aleatoric	Total	Ratio
Gemma-2B	0.233	0.091	0.324	72.0%
GPT2-XL	0.137	0.074	0.211	64.9%
Llama-3.1-8B-Instruct	0.229	0.224	0.453	50.5%
Mistral-7B-Instruct	0.081	0.124	0.205	39.4%

Table 3: Uncertainty decomposition showing absolute values and epistemic ratios. All models show substantial epistemic uncertainty, indicating sensitivity to decoding strategies.

5 Discussion

5.1 Theoretical Implications

Our credal set framework advances uncertainty quantification theory for generative models in several ways. By treating uncertainty as inherently distributional and prompt-dependent, we move beyond scalar measures that collapse rich variation patterns. The geometric interpretation through credal set operations provides intuitive understanding of miscalibration modes: models can fail through incorrect positioning (wrong variation type), volume (over/under-exploration), or shape (wrong dimensions).

The finding that best calibration reaches only 0.434 reveals fundamental gaps in how current models capture human creative variation. The notably low overlap coefficients (maximum 0.033) indicate that model and human credal sets occupy largely disjoint regions in diversity space, suggesting that current models operate in fundamentally different creative regimes than humans. The high PC1 dominance (85.8% variance) with syntactic diversity as the primary driver indicates that current models treat diversity dimensions as tightly coupled, potentially missing independent variation patterns that humans explore.

5.2 Implications for Model Development

The weak positive correlation between model scale and calibration quality ($\rho = 0.400$, $p = 0.600$) suggests that while larger models may have slight advantages, scale alone is not a determining factor for calibration quality. Our results indicate that

training objectives and data distributions likely matter more than parameter count for uncertainty calibration. The lack of significant difference between base and instruction-tuned models ($t = -0.712$, $p = 0.486$, Cohen's $d = -0.336$) with a small effect size indicates that alignment training has minimal impact on creative diversity calibration. Interestingly, instruction-tuned models showed slightly higher mean calibration (0.305 vs 0.274), though this difference was not statistically significant. The substantial epistemic uncertainty across all models (39.4-72.0%) highlights that decoding strategy choice remains a dominant source of variation. Notably, Gemma-2B shows the highest epistemic ratio (72.0%), suggesting extreme sensitivity to decoding configuration despite achieving the best single-configuration performance. This paradox suggests that optimal calibration may require careful strategy selection rather than robust performance across strategies.

5.3 Practical Deployment Considerations

For practitioners deploying language models in creative applications, our findings offer concrete guidance:

- **Model selection**: Mistral-7B-Instruct offers the most consistent performance across strategies (mean calibration: 0.371), while Gemma-2B with temperature 0.7 provides the best single configuration (0.434).
- **Strategy optimization**: Top-k sampling provides the highest mean calibration (0.323), though all models show substantial epistemic uncertainty (39-72%), making careful tuning essential.
- **Baseline expectations**: With maximum calibration at 0.434 and overlap coefficients of at most 0.033, expect substantial divergence from human creative patterns.
- **Multi-strategy ensemble**: Given high epistemic ratios, combining outputs from multiple decoding strategies is crucial for approximating human creative diversity.
- **Model-specific tuning**: In terms of calibration, base models (especially Gemma-2B at 72% epistemic) require more careful strategy selection than instruction-tuned models like Mistral-7B-Instruct (39.4% epistemic).
- **Calibration vs. quality**: Calibration along semantic, lexical, and syntactic dimensions does not necessarily indicate qualitative alignment between model-generated and human-produced text. Future work will investigate this relationship comprehensively using both human evaluations and LLM-as-a-Judge scores.
- **Generalizatbility**: Our findings are specific to *storytelling*—an open-ended task prioritizing communicative goals such as creativity, fluency, and engagement. To extend this analysis to other Natural Language Generation (NLG) research areas, we suggest task-specific calibration analyses, as different tasks involve distinct communicative objectives and varying degrees of human production variability that serve as calibration benchmarks.

6 Conclusion

This work establishes credal sets as a rigorous framework for uncertainty quantification in open-ended text generation, enabling principled geometric comparison between human and model variation patterns. Through comprehensive analysis of 100,000 generated stories calibrated against 5,000 human-written stories, we demonstrate substantial gaps in how current language models capture human creative variation, with the best calibration reaching only 0.434 (Gemma-2B with temperature 0.7) and overlap coefficients at most 0.033.

Our decomposition reveals that epistemic uncertainty from decoding strategy choice contributes 39.4-72.0% of total uncertainty across models, with base models showing higher sensitivity to configuration choices. The weak correlation between model scale and calibration ($\rho = 0.400$, $p = 0.600$) and lack of significant difference between base and instruction-tuned models ($p = 0.486$) challenge common assumptions about model development priorities. The credal set framework provides actionable insights for deploying language models in creative contexts and establishes quantitative benchmarks for evaluating progress toward human-AI creative alignment. As language models increasingly engage in open-ended generation tasks, our findings highlight the critical importance of decoding strategy selection and the need for architectural or training innovations specifically targeting uncertainty calibration.

Limitations

Several limitations warrant consideration:

- Our analysis uses convex hulls which may not capture non-convex uncertainty regions or multimodal distributions within credal sets.
- The 500-prompt sample from a single domain may not generalize to other creative writing contexts or languages.
- Decoding strategies evaluated prioritize high-probability tokens, whereas humans often select surprising, low-probability tokens for creative effect—a mismatch that may constrain achievable calibration.
- Human baselines include natural skill variation beyond pure creativity, potentially inflating aleatoric uncertainty estimates.
- Computational constraints limited us to 10 samples per configuration; larger samples might reveal finer-grained patterns.
- Statistical variance alone cannot distinguish creative quality from random variation—validating the relationship between our metrics and perceived creative quality is essential future work.

Ethics Statement

We affirm that our research adheres to the ACL Ethics Policy. This work uses publicly available datasets and involves no human subjects or personally identifiable information. We acknowledge potential biases in the Reddit-sourced dataset and encourage diverse dataset development. Our framework could help identify and mitigate generation biases by comparing model variation patterns across different demographic or cultural contexts. All code and data are released to enable reproducible research and further investigation of these important issues.

Acknowledgments

Esteban Garces is sponsored by the Munich Center for Machine Learning (MCML) and the LMU Mentoring Program. Julian Rodemann acknowledges support by the Bavarian Institute for Digital Transformation (bidt) within the Bavarian Academy of Sciences (BAS) and the LMU Mentoring Program.

References

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169.

Lukas Aichberger, Kajetan Schweighofer, and Sepp Hochreiter. 2024. Rethinking uncertainty estimation in natural language generation. *Preprint*, arXiv:2412.15176.

C. Bradford Barber, David P. Dobkin, and Hannu Huhdanpaa. 1996. The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.*, 22(4):469–483.

Yuanhao Ding, Esteban Garces Arias, Meimingwei Li, Julian Rodemann, Matthias Aßenmacher, Danlu Chen, Gaojuan Fan, Christian Heumann, and Chongsheng Zhang. 2025. GUARD: Glocal uncertainty-aware robust decoding for effective and efficient open-ended text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China. Association for Computational Linguistics. Equal contribution: Yuanhao Ding, Esteban Garces Arias, Meimingwei Li.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 516 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR.

Esteban Garces Arias, Hannah Blocher, Julian Rodemann, Matthias Aßenmacher, and Christoph Jansen. 2025c. Statistical multicriteria evaluation of LLM-generated text. In *Proceedings of the 18th International Natural Language Generation Conference (INLG 2025)*, Hanoi, Vietnam. Preprint available at arXiv:2506.18082.

Esteban Garces Arias, Hannah Blocher, Julian Rodemann, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. 2025b. Towards better open-ended text generation: A multicriteria evaluation framework. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 631–654, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Esteban Garces Arias, Julian Rodemann, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher.

2024. Adaptive contrastive search: Uncertainty-guided decoding for open-ended text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15060–15080, Miami, Florida, USA. Association for Computational Linguistics.

Gemma-Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2024. Decomposing uncertainty for large language models through input clarification ensembling. *Preprint*, arXiv:2311.08718.

D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge. 1993. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30.

Isaac Levi. 1980. *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Chance*. MIT Press.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Preprint*, arXiv:2205.14334.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. 2024. Uncertainty quantification for in-context learning of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3357–3370, Mexico City, Mexico. Association for Computational Linguistics.

Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. 2024. Uncertainty estimation and quantification for llms: A simple supervised approach. *Preprint*, arXiv:2404.15993.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121.

Yaniv Ovadia, Emily Fertig, Jie Jessie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Neural Information Processing Systems*.

Silviu Pitis, Michael R. Zhang, Andrew Wang, and Jimmy Ba. 2023. Boosted prompt ensembles for large language models. *Preprint*, arXiv:2304.05970.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Joon Oh. 2024. Calibrating large language models using their generations only. *Preprint*, arXiv:2403.05973.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep

self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. 2023. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Uncertainty in artificial intelligence*, pages 2282–2292. PMLR.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *Preprint*, arXiv:2306.13063.

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. 2024. To believe or not to believe your llm. *arXiv preprint arXiv:2406.02543*.

Marco Zaffalon and Enrico Fagiuoli. 2003. Tree-based credal networks for classification. *Reliable computing*, 9(6):487–509.

Tunyu Zhang, Haizhou Shi, Yibin Wang, Hengyi Wang, Xiaoxiao He, Zhuowei Li, Haoxian Chen, Ligong Han, Kai Xu, Huan Zhang, and 1 others. 2025. Token-level uncertainty estimation for large language model reasoning. *arXiv preprint arXiv:2505.11737*.

A Extended Results

A.1 Credal Volume Analysis

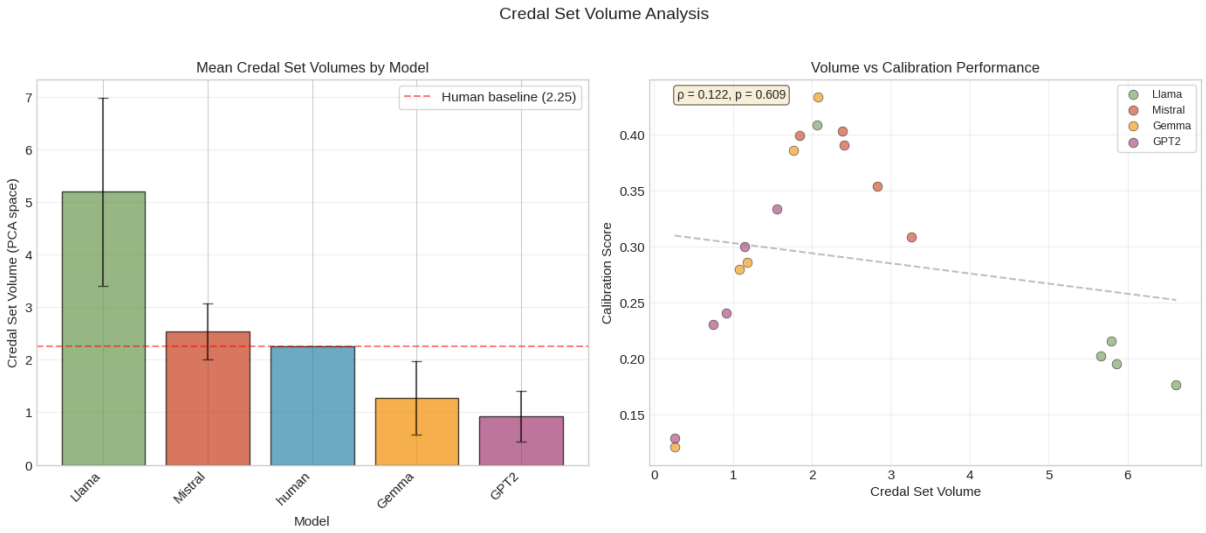

Figure 6: Analysis of credal set volumes for human and language models. **Left:** Mean credal set volumes by model (in PCA space). **Right:** Relationship between calibration score and credal set volume. A positive trend for base models (GPT2-XL and Gemma) is observed, while a negative trend is observed for instruct models (Mistral and Llama).

A.2 Wasserstein Distance Analysis

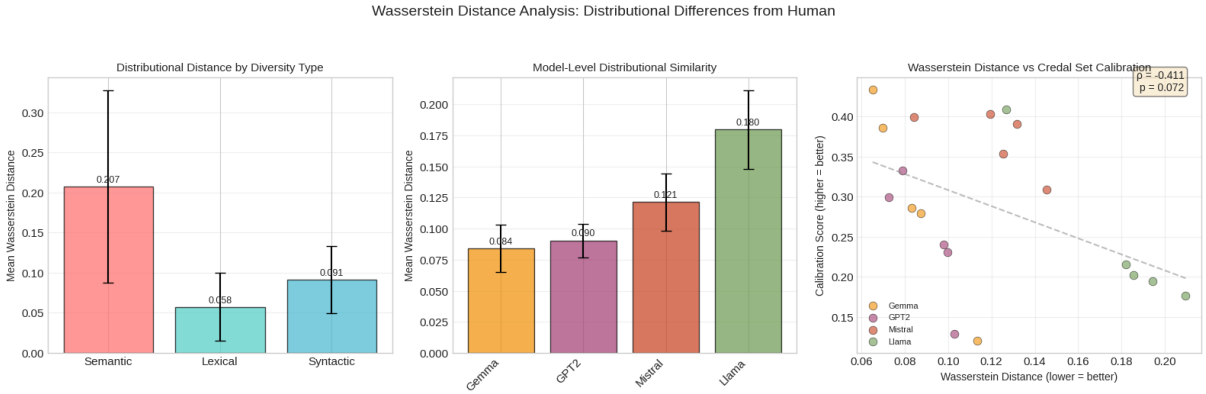

Figure 7: Distributional differences between model and human productions measured by Wasserstein distances. **Left:** Mean Wasserstein distances across semantic, lexical, and syntactic dimensions. Semantic features show the largest divergence from human distributions, followed by syntactic and lexical features. **Middle:** Model-specific distributional similarity. Gemma-2B achieves the lowest Wasserstein distances (closest to human distributions), while Llama models exhibit the highest distances. **Right:** Inverse relationship between calibration scores and Wasserstein distances (moderate negative correlation). Gemma-2B and Mistral appear in the upper-left section (high calibration, low distance), while Llama appears in the lower-right quadrant (low calibration, high distance).

A.3 Complete Calibration Results

Table 4 presents calibration coefficients for all 20 model-strategy combinations evaluated in our experiments.

Model	Strategy	Calibration
Gemma-2B	temperature_0.7	0.434
Llama-3.1-8B-Instruct	temperature_1.2	0.409
Mistral-7B-Instruct	top_k_40	0.403
Mistral-7B-Instruct	temperature_1.2	0.399
Mistral-7B-Instruct	top_p_0.9	0.391
Gemma-2B	top_k_40	0.386
Mistral-7B-Instruct	typical_0.95	0.354
GPT2-XL	temperature_0.7	0.333
Mistral-7B-Instruct	temperature_0.7	0.309
GPT2-XL	top_k_40	0.300
Gemma-2B	top_p_0.9	0.286
Gemma-2B	typical_0.95	0.279
GPT2-XL	top_p_0.9	0.240
GPT2-XL	typical_0.95	0.231
Llama-3.1-8B-Instruct	typical_0.95	0.215
Gemma-2B	temperature_1.2	0.212
Llama-3.1-8B-Instruct	top_k_40	0.199
Llama-3.1-8B-Instruct	temperature_0.7	0.196
GPT2-XL	temperature_1.2	0.188
Llama-3.1-8B-Instruct	top_p_0.9	0.175

Table 4: Complete calibration results for all model-strategy combinations, sorted by calibration coefficient.

A.4 Statistical Tests

We conducted comprehensive statistical analyses to validate our findings:

- **Model size vs calibration**: Spearman's $\rho = 0.400$ ($p = 0.600$), indicating weak positive correlation without statistical significance.
- **Base vs instruction-tuned**: Two-sample t-test: $t = -0.712$ ($p = 0.486$), no significant difference. Cohen's $d = -0.336$ (small effect size).
- **Strategy comparison**: ANOVA across strategies: $F(3, 16) = 0.200$ ($p = 0.895$), no significant differences between strategies.
- **Best performing model**: Mistral-7B-Instruct showed highest mean calibration (0.371) across all strategies.
- **Best performing strategy**: Top-k sampling achieved highest mean calibration (0.323 ± 0.092) across all models.

B Implementation Details

B.1 Computational Resources

All experiments were conducted on Google Colab with the following specifications:

- GPU: NVIDIA A100 (40GB) or V100 (16GB)
- RAM: 25-50GB depending on instance
- Storage: Google Drive for persistent storage
- Total compute time: Approximately 8 hours for generation, 1 hour for analysis

B.2 Model Configurations

Models were loaded with the following optimizations:

- 4-bit quantization for models >3B parameters using BitsAndBytes
- Flash Attention 2 where supported
- Batch sizes optimized per model (8-25 samples)
- Automatic mixed precision (AMP) with fp16

B.3 Diversity Metric Computation

Semantic embeddings were computed using Sentence-BERT (all-MiniLM-L6-v2) with the following parameters:

- Maximum sequence length: 512 tokens
- Batch size: 64 for encoding
- Pooling: Mean pooling over token embeddings

POS tagging was performed using spaCy's en_core_web_sm model with a maximum text length of 5000 characters for efficiency.