A Credal-Set Perspective on Task-Induced Distributional Drift in Text Generation

Esteban Garcés Arias*
Department of Statistics
Ludwig Maximilian University
Munich Center for Machine Learning (MCML)
Ludwigstraße 33, 80539 München
Esteban.GarcesArias@stat.uni-muenchen.de

Abstract

When a text generation system is applied under different task constraints, the distribution of its outputs may change even if model parameters remain fixed. Characterizing such task-induced drift is important for understanding and monitoring generative systems in deployment. In this work, we study how within-input production variability shifts when humans and four large language models move from open-ended storytelling to constrained translation. Rather than relying solely on scalar variability summaries, we represent each input's pairwise distance profile as a probability vector on the simplex and analyze the resulting distributions geometrically using credal sets—convex regions that capture the shape and spread of variability distributions. Across semantic, lexical, and syntactic distance measures, both human and model output distributions exhibit systematic variability contraction under translation constraints. At the population level, the human task-conditioned distributions show the largest observed cross-task displacement in this study and approximately preserve overall variability-distribution shape, whereas model distributions tend to redistribute variability unevenly across dimensions under the present operationalization, with lexical contraction accompanied by syntactic expansion. Drift directions are moderately aligned at the semantic level but less consistent for surface-level variability. These findings suggest that geometric variability signatures provide an informative, reference-free signal for characterizing how generative systems respond to changing task conditions. We release our full experimental framework at `https://github.com/EstebanGarces/human_drift`.

1 Introduction

Machine learning systems deployed in practice are routinely exposed to distributional changes. Detecting and characterizing such *drift* is important for maintaining system reliability, even when model parameters remain fixed (Quiñonero-Candela et al., 2008; Moreno-Torres et al., 2012). While most prior work focuses on temporal, domain, or covariate shift, changes in *task constraints* constitute another natural source of distributional change: the same generation system may behave differently when operating under open-ended versus constrained conditions.

In text generation, open-ended tasks such as storytelling permit a wide range of valid outputs for a given input, whereas constrained tasks such as translation impose strong semantic requirements. These differences are reflected in the *within-input variability* of outputs—the dispersion among multiple valid completions for the same prompt or source sentence. Changes in this variability across tasks provide a natural signal of distributional drift that does not rely on external labels or reference outputs.

*https://www.misoda.statistik.uni-muenchen.de/personen/mitarbeiter/garcesarias/index.html

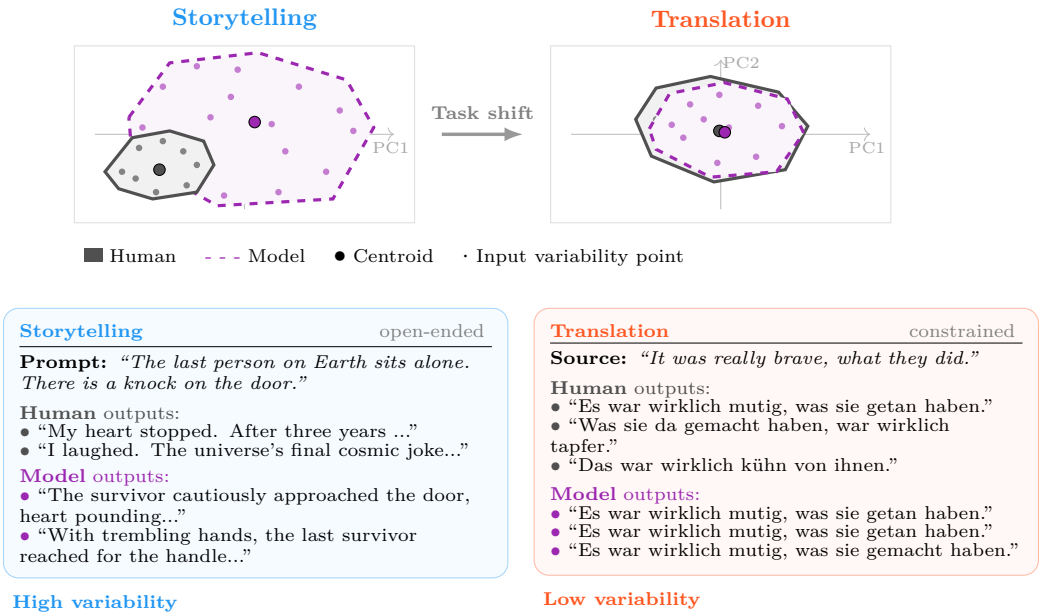

Figure 1: **Top:** Each input's pairwise distance distribution maps to a point on the probability simplex (shown in PCA projection). The convex hull of these points defines a credal set. The schematic illustrates broader separation under storytelling and greater convergence under translation. **Bottom:** Illustrative outputs for storytelling and translation. Open-ended generation permits diverse human responses, while model outputs tend to cluster more tightly. Under translation constraints, both become more similar.

This paper studies task-induced drift through a geometric analysis of within-input variability. Rather than summarizing each input by a single variability score, we represent the full distribution of pairwise output distances as a probability vector on the simplex. Across inputs, these vectors form clouds whose convex hulls define *credal sets*—closed convex sets of distributions capturing the shape, spread, and directional structure of variability profiles. Comparing credal sets across tasks enables a multi-level characterization of drift that complements scalar summaries. This perspective shifts the analysis from asking whether variability changes on average to asking how full variability-profile distributions move and reshape under changing task constraints.

We compare humans with four open-weight language models (LLaMA, Mistral, Qwen, Gemma) across two tasks: storytelling and translation (Figure 1). First, we compute per-input variability summaries using semantic, lexical, and syntactic distance measures and estimate population-level drift using bootstrap confidence intervals. Second, we analyze task-induced changes in the geometry of credal sets, quantifying drift magnitude, asymmetry, spread change, and directional alignment.

Research questions.

RQ1. How does within-input production variability shift when the generation task changes from storytelling to translation?

RQ2. How do the magnitude and geometric structure of task-induced drift differ between human and model output distributions?

RQ3. Do models and humans adapt to task constraints in similar geometric directions across semantic, lexical, and syntactic variability dimensions?

2 Related Work

Distribution shift and drift detection. Distributional change between training and deployment has been studied extensively (Quiñonero-Candela et al., 2008; Moreno-Torres et al., 2012), with surveys categorizing drift by its temporal profile (Gama et al., 2014). Most of this literature addresses temporal, domain, or covariate drift. Our work studies *task-induced* drift: the distributional change arising when the same system operates under different task constraints (Hinder et al., 2023).

Uncertainty in neural text generation. Tools for decomposing predictive uncertainty include Monte Carlo dropout (Gal & Ghahramani, 2016) and deep ensembles (Lakshminarayanan et al., 2017; Kendall & Gal, 2017; Hüllermeier & Waegeman, 2021). Kuhn et al. (2023) propose semantic entropy for language modeling, and Ling et al. (2024) decompose uncertainty in in-context learning. Our work takes a distributional rather than per-token perspective, measuring variability among multiple outputs for the same input.

Credal sets and imprecise probabilities. Credal sets—closed convex sets of probability distributions—were formalized by Walley (1991). Recent applications include credal Bayesian deep learning (Caprio et al., 2024) and conformalized credal sets (Javanmardi et al., 2024). Garces Arias et al. (2025) used credal sets from pairwise distance distributions to analyze variability in open-ended generation within a single task. We extend this framework to the *cross-task* setting.

Human production variability and LLM homogeneity. Giulianelli et al. (2023) study how human production variability manifests across NLG tasks. A growing body of work documents the tendency of language models toward lower diversity: Jiang et al. (2025) find less creative variability than humans, Padmakumar & He (2024) show instruction tuning increases homogenization, and Doshi & Hauser (2024) observe that diversity gaps widen at the collective level. The credal set framework provides a geometric characterization of these gaps beyond scalar statistics.

3 Methodology

3.1 Per-Input Variability

For an input x (a storytelling prompt or source sentence), let $\mathcal{Y}(x) = \{y_1, \ldots, y_n\}$ denote the $n = 10$ outputs produced by a given source under a given task. We compute the multiset of pairwise distances $\mathcal{D}(x; d) = \{d(y_i, y_j) : 1 \leq i < j \leq n\}$ under three complementary metrics: *semantic distance* (cosine distance on L2-normalized SBERT embeddings (Reimers & Gurevych, 2019)), *lexical distance* (Jaccard distance on word bigram sets), and *syntactic distance* (Jaccard distance on POS trigram sets). Formal definitions are provided in Appendix A.

We summarize within-input variability by the mean pairwise distance:

$$\mu(x; d) = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} d(y_i, y_j). \tag{1}$$

Task-induced drift is defined as $\Delta(d) = \mathbb{E}[\mu_T(x; d)] - \mathbb{E}[\mu_S(x; d)]$, estimated by the difference of sample means with 95% bootstrap confidence intervals ($B = 10{,}000$). We additionally compare variability distributions via the 1-Wasserstein distance between empirical CDFs of per-input means.

3.2 Credal Set Construction

The scalar summary $\mu(x; d)$ discards the shape of the distance distribution $\mathcal{D}(x; d)$. To preserve this shape, we map each input's full pairwise-distance distribution to a point on the probability simplex.

Rank transform and binning. Because metrics differ in scale (semantic distance lies in $[0, 2]$; Jaccard distances in $[0, 1]$), we apply a metric-wise pooled rank transform: for each metric, all pairwise distances across sources and tasks are pooled, and each value is replaced by its percentile rank in $[0, 1]$. Rank-transformed distances are binned into $K = 15$ uniform bins on $[0, 1]$, and counts are converted to a probability vector $\mathbf{p}(x; d) \in \Delta^{K-1}$ with Laplace smoothing ($\alpha = 1$). Each input thus maps to a point on the $(K-1)$-simplex.

Credal set. For a fixed source s, task τ, and metric d, the credal set is the convex hull of simplex points: $\mathcal{C}_{s,\tau,d} = \mathrm{conv}(\{\mathbf{p}(x; d) : x \in \mathcal{X}_{s,\tau}\})$. Prior to construction, we remove degenerate inputs where all outputs are effectively identical (maximum raw pairwise distance below 10^{-6} across all metrics). Full construction details are in Appendix B.

3.3 Geometric Drift Measures

We quantify drift between storytelling ($\mathcal{C}_S$) and translation ($\mathcal{C}_T$) credal sets using four complementary measures. **(1) Drift magnitude**: the Hausdorff distance $d_H(\mathcal{C}_S, \mathcal{C}_T)$, approximated via support functions over $N_{\mathrm{dir}} = 5{,}000$ random directions. **(2) Drift asymmetry**: the directed Hausdorff components $d(\mathcal{C}_S \to \mathcal{C}_T)$ and $d(\mathcal{C}_T \to \mathcal{C}_S)$, revealing whether one task's region extends further beyond the other. **(3) Spread change**: the ratio of mean directional widths $\mathrm{MW}(\mathcal{C}_T)/\mathrm{MW}(\mathcal{C}_S)$, indicating contraction (< 1) or expansion (> 1). **(4) Drift direction**: the cosine similarity between model and human centroid shift vectors $\Delta\mathbf{c} = \bar{\mathbf{c}}_T - \bar{\mathbf{c}}_S$, measuring directional alignment.

All measures are embedded in a bootstrap procedure ($B = 200$, $N_{\mathrm{sub}} = 100$ inputs per source–task combination). We report bootstrap means and 95% percentile confidence intervals. Formal definitions of all geometric measures are provided in Appendix C.

4 Experimental Setup

We analyze four open-weight language models: LLaMA 3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, Qwen2.5-7B-Instruct, and Gemma-2-9b-it (see Table 3, Appendix D.2). All models use fixed decoding parameters: temperature 0.7, top-p 0.95, 10 samples per input. Task differences are induced exclusively via natural-language system instructions (Appendix D.1). Full decoding hyperparameters are listed in Table 4, Appendix D.2.

The study uses two datasets: **Storytelling** (100 open-ended English prompts, 10 continuations per source) and **Translation** (500 English–German sentence pairs, 10 translations per source). Human outputs are produced by independent annotators; annotator pools differ across tasks, so the two tasks are treated as distinct task-conditioned populations. Accordingly, human cross-task comparisons should be interpreted as comparisons between task-conditioned populations rather than within-person adaptation to task change.

Semantic embeddings use `paraphrase-multilingual-MiniLM-L12-v2` (Reimers & Gurevych, 2019) with L2 normalization. POS tagging uses spaCy's language-specific pipelines (`en_core_web_sm`, `de_core_news_sm`). Stories are truncated to 200 words for POS processing.

5 Results

5.1 Per-Input Variability Shifts (RQ1)

We begin with the most direct question: how within-input production variability changes when the task shifts from open-ended storytelling to constrained translation. Figure 6 (Appendix) shows the distribution of per-input mean pairwise distances across sources, tasks, and variability metrics. Across all three metrics, both human and model outputs exhibit lower variability under translation than under storytelling, consistent with the stronger constraints imposed by meaning preservation. The contraction is most pronounced for semantic variability, where translation outputs cluster near zero distance. Lexical and syntactic

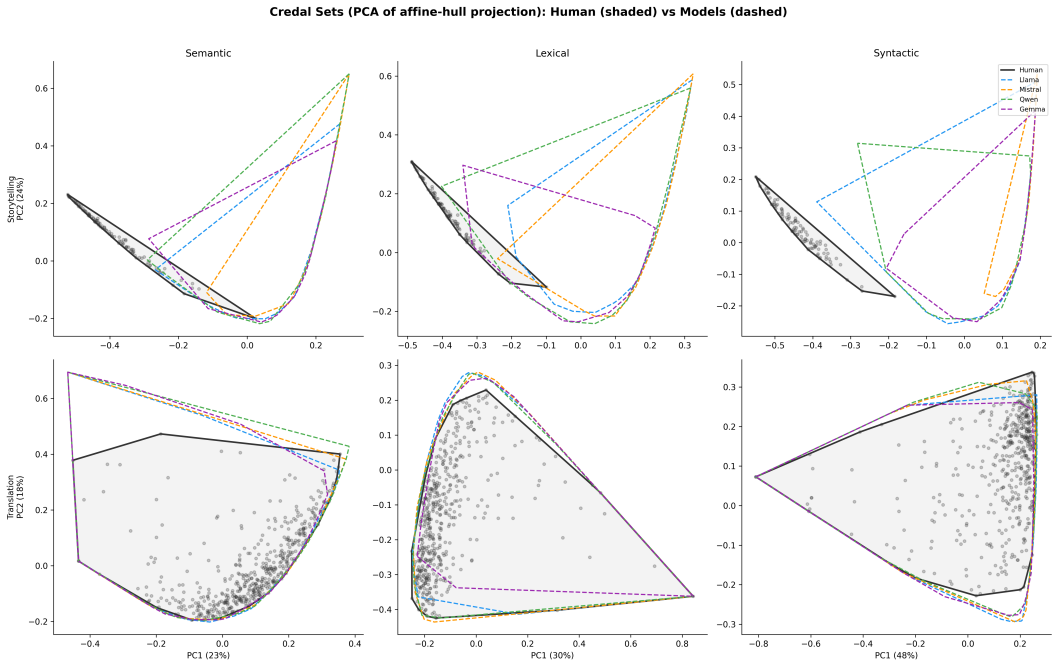

Figure 2: Credal sets projected into PCA space. Rows correspond to tasks and columns to variability metrics. Human credal sets are shown with solid hulls and model credal sets with dashed hulls. Large markers with black edges indicate centroids. Storytelling sets are generally broader and more separated, whereas translation sets are typically more compact.

variability also decrease overall, although model distributions retain more spread than the human task-conditioned distributions in these dimensions.

Figure 3 quantifies this shift using bootstrap estimates of absolute and relative drift (translation − storytelling). All sources show negative drift across all metrics. Semantic variability contracts by the largest proportion, followed by lexical and syntactic variability. Confidence intervals are narrow, indicating that these shifts are stable under resampling. Taken together, these results provide clear evidence that changing the task from storytelling to translation induces systematic within-input variability contraction.

5.2 Credal-Set Geometry: Global Overview

Having established that variability contracts at the per-input level, we now examine how the full distributions of pairwise-distance profiles change geometrically across tasks. Figure 2 projects the credal sets into PCA space for each variability metric. Two broad patterns emerge. First, storytelling credal sets generally occupy broader and more separated regions, whereas translation credal sets are typically more compact. Second, humans and models differ not only in how far their task-conditioned regions move, but also in how those regions are shaped and positioned relative to one another. This global view motivates the more specific geometric analyses below, which quantify drift magnitude, asymmetry, spread change, and directional alignment.

5.3 Geometric Drift Magnitude and Asymmetry (RQ2)

Drift magnitude. Table 7 (Appendix) reports Hausdorff distances between storytelling and translation credal sets, with bootstrap confidence intervals visualized in Figure 9 (Appendix). The human task-conditioned distributions show the highest average observed drift magnitude in this study (0.597), exceeding the four model distributions (range: 0.538–0.572). This pattern is most consistent for semantic and lexical variability. For syntactic variability,

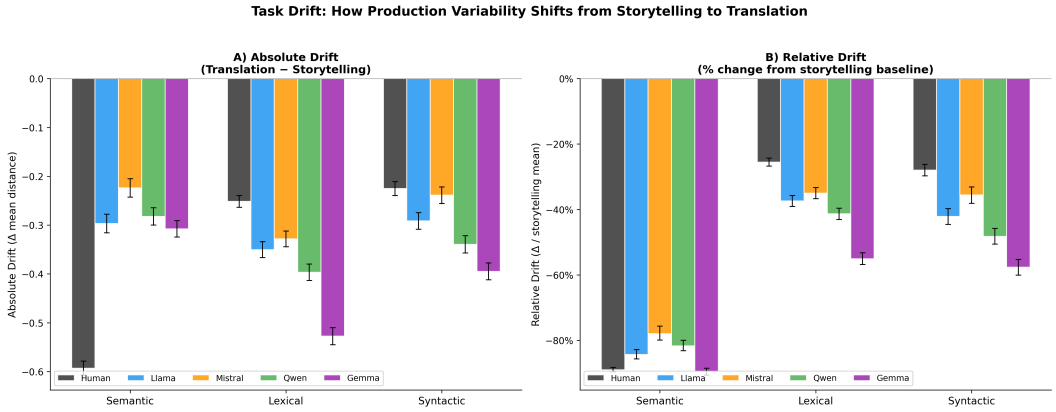

Figure 3: Bootstrap drift estimates (translation − storytelling) with 95% confidence intervals. Left: absolute drift. Right: relative drift (proportion of the storytelling baseline).

drift magnitudes are more similar across sources (0.600–0.633), suggesting weaker source dependence in this dimension. Because several confidence intervals overlap, these differences should be interpreted as moderate rather than categorical.

Drift asymmetry. Figure 10 (Appendix) decomposes this displacement into directed Hausdorff components. Across all sources and metrics, translation credal sets extend further into regions not covered by storytelling than the reverse. This asymmetry is weakest for semantic variability, where the two directions are closer in magnitude, and more pronounced for lexical and syntactic variability. One plausible interpretation is that translation variability is more heterogeneous across inputs: some source sentences yield near-degenerate outputs, whereas others admit multiple valid reformulations, producing extremes less prominent in storytelling. Overall, the geometric drift is therefore not only substantial, but also directionally asymmetric. Because these measures are hull-based, they should be interpreted as geometry-sensitive summaries of task-conditioned variability regions; bootstrap subsampling improves stability but does not fully remove sensitivity to extreme simplex points.

5.4 Spread Change and Directional Alignment (RQ3)

Spread change. Table 1 reports mean-width ratios between translation and storytelling credal sets, with Figure 7 (Appendix) visualizing the corresponding confidence intervals. Human task-conditioned distributions show ratios close to one across all three metrics, indicating approximate preservation of overall variability-distribution spread under task change. Model distributions exhibit a different pattern: lexical variability consistently contracts, while syntactic variability expands. Ratios whose bootstrap confidence intervals exclude one are marked with an asterisk. Mean-width ratios are computed within each bootstrap iteration and then averaged; accordingly, the reported ratio need not equal the quotient of the mean width estimates shown in the same row.

Drift direction and alignment. Figure 4 visualizes centroid shift vectors from storytelling to translation in the PCA projection. The human task-conditioned distributions exhibit the largest average centroid displacement in this study, whereas model shifts are generally shorter and vary more in orientation across metrics. The bootstrap confidence ellipses indicate that centroid locations are stable under subsampling.

Table 2 and Figure 8 (Appendix) quantify this comparison by reporting cosine similarity between each model's centroid shift vector and the corresponding human centroid shift vector. Alignment is moderate for semantic variability (0.39–0.59) but markedly weaker for lexical and syntactic variability. Mistral shows near-zero lexical alignment (0.03) and slightly negative syntactic alignment (−0.08), indicating drift that is nearly orthogonal to,

Table 1: Spread change: mean-width ratio (Translation / Storytelling) with bootstrap 95% CIs. Values below 1.0 indicate contraction; values above 1.0 indicate expansion. ∗: CIs exclude 1.0.

Source	Metric	MW_S	MW_T	Ratio	CI
Human	Sem.	0.349	0.322	1.02	[0.56, 1.83]
	Lex.	0.312	0.328	1.09	[0.75, 1.62]
	Syn.	0.327	0.360	1.10	[0.97, 1.25]
LLaMA	Sem.	0.394	0.367	0.93	[0.82, 1.06]
	Lex.	0.567	0.361	0.64*	[0.56, 0.72]
	Syn.	0.354	0.387	1.10	[0.99, 1.27]
Mistral	Sem.	0.416	0.374	0.90*	[0.81, 0.99]
	Lex.	0.593	0.366	0.62*	[0.55, 0.71]
	Syn.	0.302	0.398	1.32*	[1.16, 1.52]
Qwen	Sem.	0.409	0.372	0.91	[0.82, 1.03]
	Lex.	0.506	0.383	0.76*	[0.64, 0.91]
	Syn.	0.310	0.369	1.20*	[1.09, 1.33]
Gemma	Sem.	0.382	0.390	1.02	[0.90, 1.14]
	Lex.	0.409	0.363	0.89*	[0.75, 1.00]
	Syn.	0.321	0.388	1.21*	[1.10, 1.32]

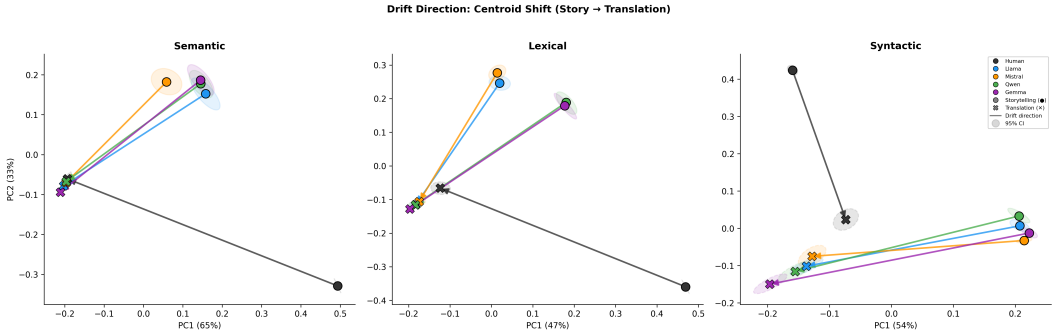

Figure 4: Drift direction: centroid shift vectors from storytelling (filled circle) to translation (cross) in PCA space. Shaded ellipses show 95% bootstrap confidence regions around each centroid. Human task-conditioned distributions exhibit larger displacement magnitudes, while model drift directions vary more across metrics.

or slightly opposed to, the human pattern in those dimensions. Across all models, alignment is strongest in the semantic dimension, suggesting greater agreement at the level of meaning than at the lexical or syntactic levels.

6 Discussion, Limitations, and Conclusion

Addressing the research questions. Regarding **RQ1**, both the per-input analysis and the credal-set framework show that within-input variability contracts when moving from storytelling to translation, with the largest reduction at the semantic level (Figure 3). Regarding **RQ2**, the human task-conditioned populations exhibit the largest observed cross-task drift magnitude on average in this study (Table 7 (Appendix); Figure 4), although this comparison should be interpreted cautiously because the human outputs were collected from different annotator pools across tasks. Regarding **RQ3**, the geometric analysis reveals qualitative differences not captured by scalar summaries. Human credal sets do not show strong evidence of systematic spread expansion or contraction across tasks, whereas model credal sets tend to compress lexical range while expanding syntactic range (Table 1). Drift direction alignment is moderate at the semantic level but weaker at the lexical and syntactic

Table 2: Drift direction alignment: cosine similarity between model and human centroid shift vectors, with bootstrap 95% CIs.

Model	Metric	Cosine	CI_{lo}	CI_{hi}
LLaMA	Semantic	0.586	0.513	0.665
	Lexical	0.065	0.003	0.121
	Syntactic	0.099	0.045	0.170
Mistral	Semantic	0.394	0.300	0.475
	Lexical	0.026	−0.019	0.069
	Syntactic	−0.076	−0.133	−0.016
Qwen	Semantic	0.546	0.489	0.611
	Lexical	0.365	0.310	0.430
	Syntactic	0.175	0.107	0.242
Gemma	Semantic	0.484	0.401	0.560
	Lexical	0.367	0.309	0.435
	Syntactic	0.074	0.018	0.126

levels (Table 2), indicating that models and humans do not adapt to task constraints in the same way.

Implications for drift monitoring. These findings suggest that variability-based distributional signatures can provide informative signals for characterizing task-induced drift. The credal set framework operates without labels or reference outputs, captures distributional shape beyond scalar summaries, and enables multi-level comparison. The observation that agreement at one level (e.g., semantics) can mask disagreement at others (e.g., syntax) underscores the value of multi-level monitoring.

Limitations. We analyze two tasks (storytelling and translation), so generalizability to other task pairs remains to be established. The three variability metrics are complementary but not exhaustive: discourse structure, pragmatic appropriateness, and stylistic variation are not explicitly represented. Translation is limited to English–German, and typologically more distant language pairs may yield different patterns. The human outputs serve as a reference distribution rather than a normative standard. Because the human outputs for storytelling and translation were produced by different annotator pools, cross-task human comparisons should be interpreted as comparisons between task-conditioned populations rather than paired within-annotator drift. Sample sizes also differ across tasks (100 vs. 500 inputs), although the bootstrap subsampling procedure partially addresses this imbalance for the credal-set analysis. Because credal sets are constructed as convex hulls, geometric quantities such as Hausdorff distance and mean width can be influenced by extreme simplex points, even though bootstrap subsampling helps stabilize these estimates. In particular, the observed pattern of lexical contraction paired with syntactic expansion should be interpreted relative to the specific tokenization, POS-tagging, and truncation choices used in the present operationalization. Our analysis is descriptive rather than causal, so we do not identify the mechanisms underlying the observed differences between human and model distributions. Further details on design choices and hyperparameters are provided in Appendix D.

Conclusion. We presented a distributional analysis of task-induced drift in text generation, combining per-input variability summaries with a geometric framework based on credal sets. Comparing storytelling and translation across human and model output distributions, we found that (i) both human and model outputs exhibit systematic variability contraction under stronger task constraints, (ii) the human task-conditioned distributions show the largest observed cross-task displacement on average, and (iii) model drift directions are only moderately aligned with the human pattern, with the weakest alignment at the lexical and syntactic levels. Future work may extend this framework to temporal drift, incorporate additional representational levels, and develop online monitoring procedures based on geometric drift signatures.

Acknowledgments

The author sincerely thanks the Mentoring Program of the Faculty of Mathematics, Statistics and Informatics at LMU Munich for its support. He is also grateful to Prof. Dr. Eyke Hüllermeier, Prof. Dr. Christian Heumann, Yusuf Sale, and Julian Rodemann for stimulating discussions and helpful suggestions that contributed to this work.

References

Michele Caprio, Souradeep Dutta, Kuk Jin Jang, Vivian Lin, Radoslav Ivanov, Oleg Sokolsky, and Insup Lee. Credal bayesian deep learning, 2024. URL https://arxiv.org/abs/2302.09656.

Anil R. Doshi and Oliver P. Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024. doi: 10.1126/sciadv.adn5290. URL https://www.science.org/doi/abs/10.1126/sciadv.adn5290.

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

João Gama, Indrundefined Žliobaitundefined, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4), March 2014. ISSN 0360-0300. doi: 10.1145/2523813. URL https://doi.org/10.1145/2523813.

Esteban Garces Arias, Julian Rodemann, and Christian Heumann. The geometry of creative variability: How credal sets expose calibration gaps in language models. In Bryan Eikema, Raúl Vázquez, Jonathan Berant, Marie-Catherine de Marneffe, Barbara Plank, Artem Shelmanov, Swabha Swayamdipta, Jörg Tiedemann, Chrysoula Zerva, and Wilker Aziz (eds.), *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pp. 36–50, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-349-4. doi: 10.18653/v1/2025.uncertainlp-main.5. URL https://aclanthology.org/2025.uncertainlp-main.5/.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. What comes next? evaluating uncertainty in neural text generators against human production variability. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14349–14371, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.887. URL https://aclanthology.org/2023.emnlp-main.887/.

Fabian Hinder, Valerie Vaquet, and Barbara Hammer. One or two things we know about concept drift – a survey on monitoring evolving environments, 2023. URL https://arxiv.org/abs/2310.15826.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

Alireza Javanmardi, David Stutz, and Eyke Hüllermeier. Conformalized credal set predictors, 2024. URL https://arxiv.org/abs/2402.10723.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond), 2025. URL https://arxiv.org/abs/2510.22954.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 2017.

Lorenz Kuhn, Yarin Gal, and Mark Gales. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *International Conference on Learning Representations*, 2023.

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

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. Uncertainty quantification for in-context learning of large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3357–3370, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.184. URL `https://aclanthology.org/2024.naacl-long.184/`.

Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45 (1):521–530, 2012. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2011.06.019. URL `https://www.sciencedirect.com/science/article/pii/S0031320311002901`.

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity?, 2024. URL `https://arxiv.org/abs/2309.05196`.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2008.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL `https://aclanthology.org/D19-1410/`.

Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, 1991.

Appendix

A Metric Definitions

Semantic distance. Let $\mathbf{e}(y) \in \mathbb{R}^m$ be the L2-normalized SBERT embedding of output y (Reimers & Gurevych, 2019), so $\|\mathbf{e}(y)\|_2 = 1$. The cosine distance is

$$d_{\text{sem}}(y_i, y_j) \;=\; 1 - \mathbf{e}(y_i)^\top \mathbf{e}(y_j),$$

which lies in $[0, 2]$ for unit vectors. We use `paraphrase-multilingual-MiniLM-L12-v2` with `normalize_embeddings=True`; the SBERT encoder truncates inputs to 128 tokens.

Lexical distance. Let $B_{\text{word}}(y)$ be the set of word bigrams in y after lowercased tokenization. The Jaccard distance is

$$d_{\text{lex}}(y_i, y_j) \;=\; 1 - \frac{|B_{\text{word}}(y_i) \cap B_{\text{word}}(y_j)|}{|B_{\text{word}}(y_i) \cup B_{\text{word}}(y_j)|},$$

which lies in $[0, 1]$.

Syntactic distance. Let $T_{\text{POS}}(y)$ be the set of POS trigrams extracted with spaCy. The syntactic Jaccard distance is

$$d_{\text{syn}}(y_i, y_j) \;=\; 1 - \frac{|T_{\text{POS}}(y_i) \cap T_{\text{POS}}(y_j)|}{|T_{\text{POS}}(y_i) \cup T_{\text{POS}}(y_j)|},$$

also in $[0, 1]$. Stories are truncated to the first 200 words before POS tagging.

B Credal Set Construction Details

Rank transform. For each metric d, let $\mathcal{A}_d$ be the multiset of all pairwise distances across all sources and both tasks. For any observed distance value z, its percentile rank is

$$r_d(z) \;=\; \frac{1}{|\mathcal{A}_d|} \sum_{a \in \mathcal{A}_d} \mathbf{1}\{a \le z\} \;\in [0,1].$$

We replace each $z \in \mathcal{D}(x;d)$ by $r_d(z)$, placing all distances on a common scale.

Histogram-to-simplex mapping. Fix $K = 15$ uniform bins on $[0,1]$ with edges $0 = b_0 < b_1 < \cdots < b_K = 1$. Let $c_k(x;d)$ be the count of rank-transformed distances in bin k. With Laplace smoothing $\alpha = 1$:

$$p_k(x;d) \;=\; \frac{c_k(x;d) + \alpha}{\sum_{\ell=1}^{K} \left(c_\ell(x;d) + \alpha\right)}.$$

Degeneracy filter. We remove any input x such that

$$\max\Big(\max \mathcal{D}(x;d_{\text{sem}}), \max \mathcal{D}(x;d_{\text{lex}}), \max \mathcal{D}(x;d_{\text{syn}})\Big) < 10^{-6},$$

applied uniformly across all sources and tasks.

C Geometric Measure Definitions

For a source s and metric d, let $\mathcal{C}_S$ and $\mathcal{C}_T$ denote the storytelling and translation credal sets. All points are projected to their affine hull via SVD; random directions are sampled in that reduced dimension.

Hausdorff distance.

$$d_H(\mathcal{C}_S, \mathcal{C}_T) \;=\; \max\Big\{\sup_{x \in \mathcal{C}_S} \inf_{y \in \mathcal{C}_T} \|x - y\|_2,\; \sup_{y \in \mathcal{C}_T} \inf_{x \in \mathcal{C}_S} \|x - y\|_2\Big\},$$

approximated via support functions $h_{\mathcal{C}}(u) = \sup_{x \in \mathcal{C}} u^\top x$ over $N_{\text{dir}} = 5{,}000$ random directions.

Directed Hausdorff components. We define

$$d(\mathcal{C}_S \to \mathcal{C}_T) = \max_u \left[h_{\mathcal{C}_S}(u) - h_{\mathcal{C}_T}(u)\right]_+ ,$$

and

$$d(\mathcal{C}_T \to \mathcal{C}_S) = \max_u \left[h_{\mathcal{C}_T}(u) - h_{\mathcal{C}_S}(u)\right]_+ .$$

Mean directional width. For direction u, the width is

$$w_{\mathcal{C}}(u) = h_{\mathcal{C}}(u) - h_{\mathcal{C}}(-u).$$

Mean width $\mathrm{MW}(\mathcal{C})$ averages $w_{\mathcal{C}}(u)$ over random directions. Spread-change ratios are computed within each bootstrap iteration and then averaged across iterations; thus, the reported mean ratio need not equal the quotient of the reported mean widths.

Centroid shift and alignment. The empirical centroid is

$$\bar{\mathbf{c}}_\tau = \frac{1}{|\mathcal{P}_{s,\tau,d}|} \sum_{\mathbf{p} \in \mathcal{P}_{s,\tau,d}} \mathbf{p}.$$

The drift vector is $\Delta\mathbf{c} = \bar{\mathbf{c}}_T - \bar{\mathbf{c}}_S$. Alignment between a model and the human reference is measured by cosine similarity:

$$\cos(\Delta\mathbf{c}_{\text{model}}, \Delta\mathbf{c}_{\text{human}}) = \frac{\Delta\mathbf{c}_{\text{model}}^\top \Delta\mathbf{c}_{\text{human}}}{\|\Delta\mathbf{c}_{\text{model}}\|_2 \, \|\Delta\mathbf{c}_{\text{human}}\|_2}.$$

Model key	HuggingFace identifier	Notes
`llama`	`meta-llama/Llama-3.1-8B-Instruct`	listed in config
`mistral`	`mistralai/Mistral-7B-Instruct-v0.3`	listed in config
`qwen`	`Qwen/Qwen2.5-7B-Instruct`	listed in config
`gemma`	`google/gemma-2-9b-it`	listed in config

Table 3: Models used for generation.

Setting	Storytelling	Translation
Decoding mode	sampling (`do_sample=True`)	sampling (`do_sample=True`)
Temperature	`0.7`	`0.7`
Top-p (nucleus)	`0.95`	`0.95`
# generations per input	`10`	`10`
Max new tokens	`2048`	`256`
RNG seed	`42` (set before each call)	same
Padding	`pad_token_id = eos_token_id`	same
Precision / device	`bfloat16`, `device_map="cuda"`	same

Table 4: Decoding hyperparameters used to generate 10 samples per input for each task.

D Experimental Details and Hyperparameters

D.1 Task Instructions

Task differences are induced exclusively via natural-language system instructions; decoding hyperparameters are held fixed across tasks.

Storytelling.

> *You are a creative fiction writer. Write a story continuation based on the given writing prompt. Output only the story text. Do not repeat the prompt, do not add titles, and do not add commentary.*

The user message consists of a WritingPrompts-style prompt with leading tags (e.g., `[WP]`) removed.

Translation.

> *You are a professional English-to-German translator. Translate the given English sentence into German. Output only the German translation and nothing else.*

Inputs are formatted using each model's chat template with system and user roles. If system messages are unsupported, the instruction is prepended to the user message.

D.2 Generation Hyperparameters

D.3 Analysis Hyperparameters

Tables 5 and 6 summarize the parameters for the per-input variability and credal-set analyses, respectively.

D.4 Reproducibility

All analyses were run in Google Colab with GPU acceleration (SBERT encoding on `cuda`). We fix the global NumPy RNG seed at 42 and use a dedicated generator with seed 2025 for bootstrap subsampling and per-iteration Monte Carlo direction seeds. All confidence intervals are 95% percentile bootstrap intervals.

Category	Parameter	Value / description
General	Random seed	`np.random.seed(42)`
General	Sources	{`human, llama, mistral, qwen, gemma`}
General	Tasks	{`storytelling, translation`}
General	Metrics	{`semantic, lexical, syntactic`}
Filtering	Min. gens/input	≥ 2
SBERT	Model	`paraphrase-multilingual-MiniLM-L12-v2`
SBERT	Device / batch	`cuda` / `256`
SBERT	Max tokens	128
spaCy	Models	`en_core_web_sm` / `de_core_news_sm`
spaCy	Story truncation	200 words before POS tagging
Bootstrap	Iterations	$B = 10{,}000$; CI: percentile 2.5/97.5
Wasserstein	Distance	`scipy.stats.wasserstein_distance`

Table 5: Hyperparameters for per-input variability and bootstrap drift analysis.

Category	Parameter	Value / description
Degeneracy	Threshold	10^{-6} on max pairwise distance across metrics
Binning	Bins / smoothing	$K = 15$ uniform; Laplace $\alpha = 1$
Binning	Min. bin width	$0.05\times$ data range (quantile fallback)
Bootstrap	Iterations / subsample	$B = 200$; $N_{\text{sub}} = 100$ per (s, t)
Monte Carlo	Directions / seed	$N_{\text{dir}} = 5{,}000$; master seed 2025

Table 6: Hyperparameters for credal-set drift analysis.

D.5 Additional Notes on Robustness

Because credal sets are constructed as convex hulls, geometric quantities such as Hausdorff distance and mean width can be influenced by extreme simplex points. The bootstrap subsampling procedure reduces instability by recomputing the geometric summaries across repeated subsets of inputs, but it does not eliminate sensitivity to hull-defining extremes. Robust alternatives, such as trimmed-hull constructions or other set summaries based on central regions, remain important directions for future work.

E Supplementary Figures and Tables

Figure 5 shows the credal-set centroids in shared rank-transform space for all source–task combinations, providing a direct view of where distributional mass concentrates in each bin of the distance histogram.

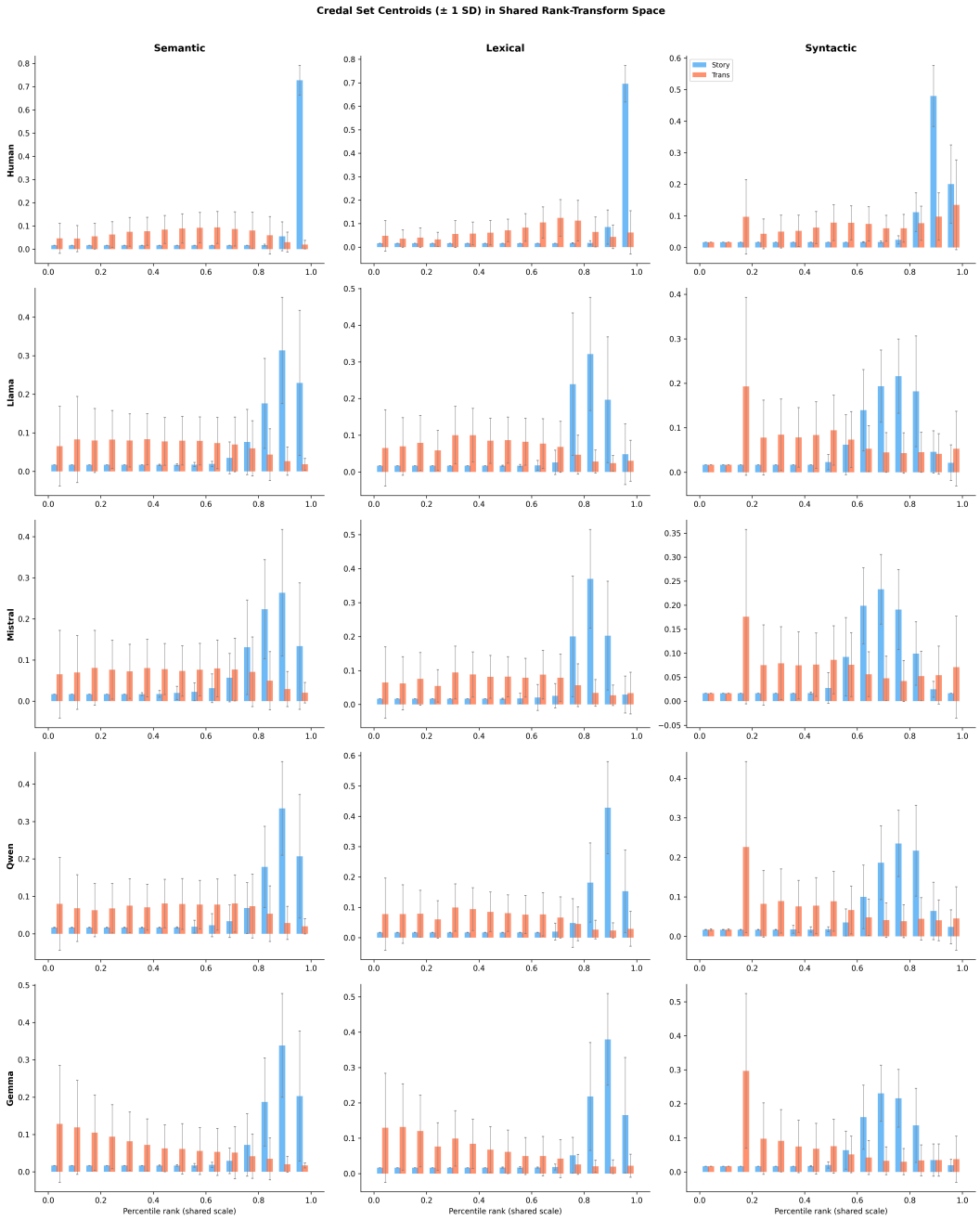

Figure 5: Credal-set centroids (± 1 SD) in shared rank-transform space. Rows correspond to sources and columns to metrics. Blue bars denote storytelling and orange bars denote translation.

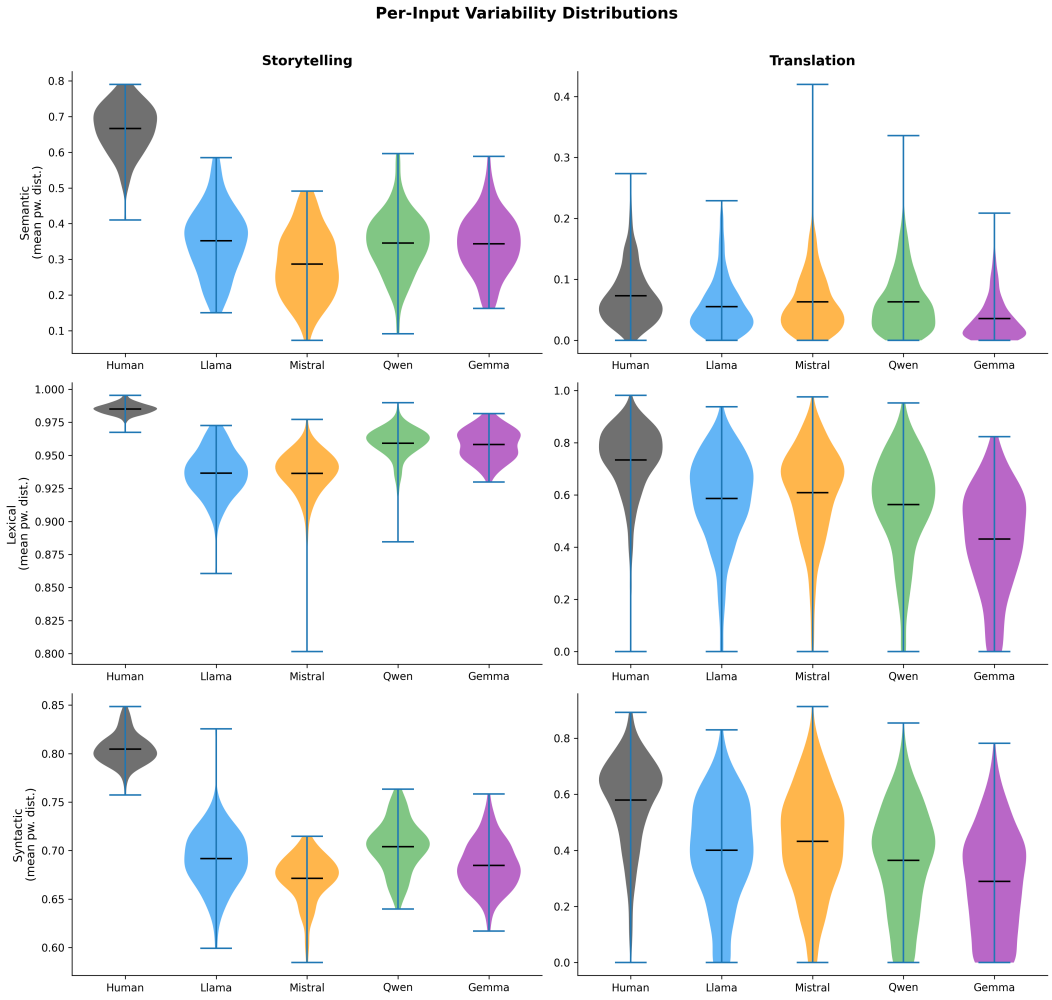

Figure 6: Per-input variability distributions across sources and tasks. Each violin summarizes the distribution of mean pairwise distances over all inputs for a given source, task, and metric.

Table 7: Drift magnitude: Hausdorff distance d_H between storytelling and translation credal sets per source and metric, with bootstrap 95% CIs.

Source	Metric	d_H	$\mathrm{CI}_{\mathrm{lo}}$	$\mathrm{CI}_{\mathrm{hi}}$
Human	Semantic	0.604	0.530	0.682
	Lexical	0.575	0.494	0.682
	Syntactic	0.610	0.495	0.726
LLaMA	Semantic	0.537	0.449	0.650
	Lexical	0.516	0.421	0.640
	Syntactic	0.617	0.557	0.681
Mistral	Semantic	0.488	0.415	0.586
	Lexical	0.493	0.418	0.582
	Syntactic	0.633	0.547	0.695
Qwen	Semantic	0.496	0.429	0.570
	Lexical	0.523	0.444	0.603
	Syntactic	0.601	0.539	0.685
Gemma	Semantic	0.549	0.482	0.643
	Lexical	0.548	0.493	0.604
	Syntactic	0.620	0.555	0.707

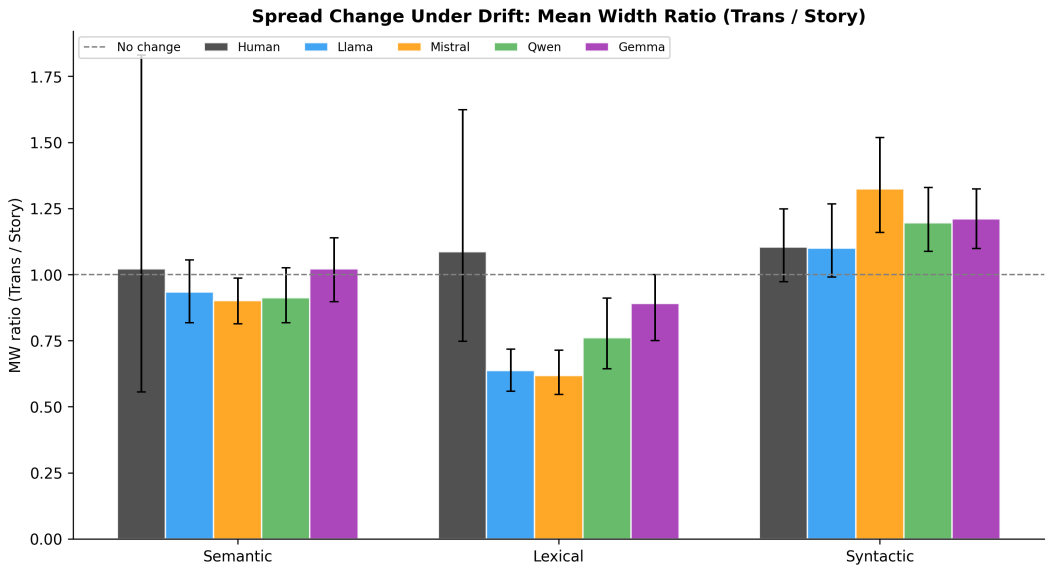

Figure 7: Spread change: mean-width ratio per source and metric. The dashed line at 1.0 indicates no change. Error bars show bootstrap 95% CIs.

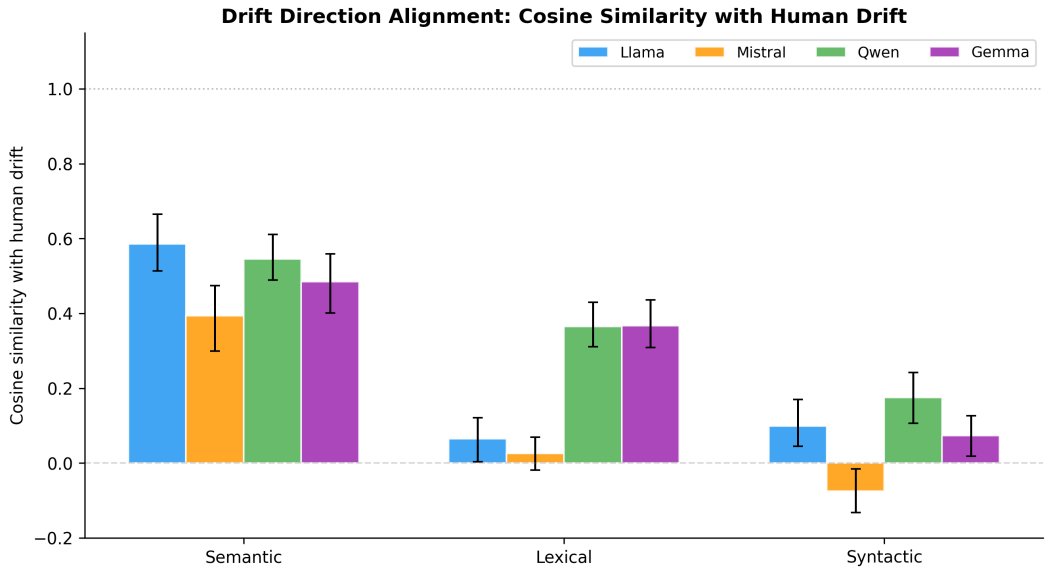

Figure 8: Cosine similarity between each model's drift vector and the human drift vector, per metric. Error bars show bootstrap 95% CIs.

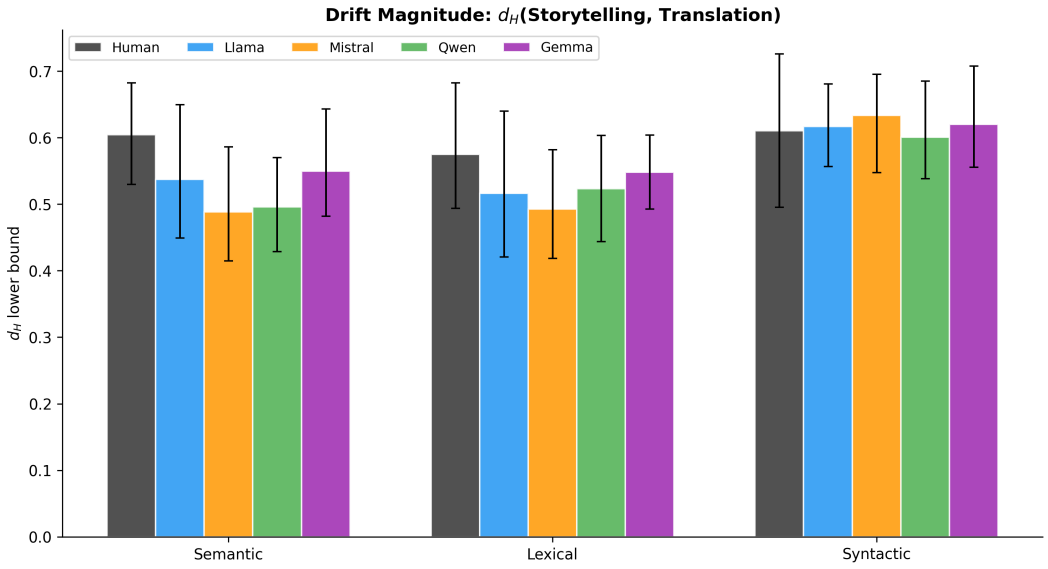

Figure 9: Drift magnitude per source and metric (bootstrap 95% CIs).

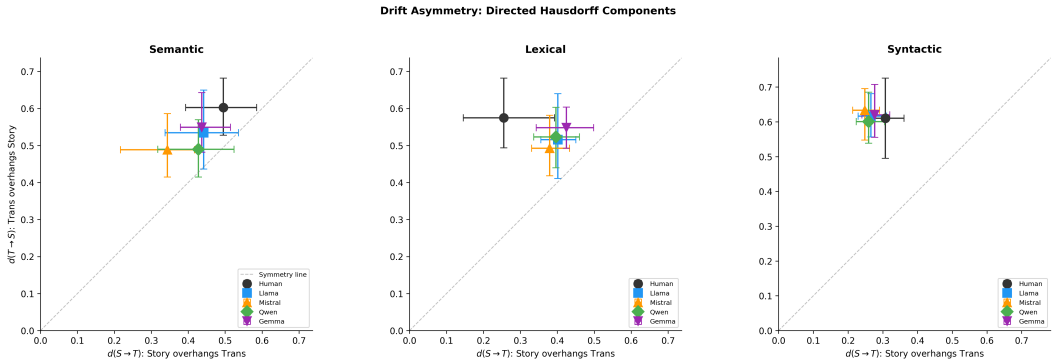

Figure 10: Drift asymmetry: directed Hausdorff components with bootstrap 95% CIs. Points above the diagonal indicate that translation extends further beyond storytelling than the reverse.