Inter-Model Prediction Divergence for Measuring Narrative Surprise: What Cross-Model Disagreement Reveals About Story Structure

Anonymous ACL submission

Abstract

Recent work has shown that LLM-based narrative flow metrics such as sequentiality suffer from systematic cross-model inconsistency: different language models assign different flow scores to the same stories. Rather than treating this inconsistency as a defect, we propose **Inter-Model Prediction Divergence (IPD)**, a continuous, sentence-level measure of narrative surprise computed as the Jensen-Shannon Divergence among multiple LLMs' next-token prediction distributions at sentence boundaries. We evaluate IPD on 240 stories from the Hippocorpus corpus with sentence-level surprise annotations and the Story Cloze Test, comparing against nine baselines. We find that while raw IPD does not outperform single-model surprisal (AP = 0.193 vs. 0.250), it captures a statistically significant complementary signal: a likelihood ratio test confirms IPD contributes information beyond both surprisal and contextual entropy ($p < 10^{-10}$), and the "pure disagreement" component of IPD (residual after removing entropy) achieves AP = 0.240, outperforming raw IPD. The surprise-flow relationship differs systematically across story types: imagined narratives show anti-correlated surprise and flow ($\rho = -0.216$) while recalled stories show weak positive correlation ($r = 0.069$). Analysis of epistemic uncertainty reveals high pairwise perplexity correlations ($r > 0.94$) among ensemble members, suggesting that more diverse ensembles could strengthen the approach.

1 Introduction

Computational measures of narrative structure have emerged as a key tool for studying how stories are constructed and processed. Sap et al. (2022) introduced *sequentiality* as an LLM-derived measure of narrative flow, defined as the degree to which preceding narrative context helps predict the next sentence beyond topic alone. However, Sunny et al. (2025) demonstrated that this metric suffers from systematic topic confounds, varies across LLMs, and fails to distinguish stories with intentionally good versus poor narrative flow.

A critical but underexplored aspect of these findings is the *cross-model inconsistency* of sequentiality scores. Different LLMs assign different flow scores to the same stories. This has been treated purely as a problem: evidence that the metric is unreliable. We propose reframing cross-model disagreement as an *information-bearing signal*.

Our key insight is grounded in ensemble learning theory: when multiple models with different inductive biases encounter *predictable* content, they converge in their predictions; when they encounter *surprising* content, they diverge (Lakshminarayanan et al., 2017). This divergence-as-surprise principle has been applied to multi-LLM settings for binary prediction tasks (Kruse et al., 2025) and to emotional state distributions within single models for narrative pivot detection (Schulz et al., 2024).

We define **Inter-Model Prediction Divergence (IPD)** as a continuous, sentence-level measure of narrative surprise computed from the disagreement among multiple LLMs' next-token prediction distributions at sentence boundaries. IPD differs from MUSE (Kruse et al., 2025) in three key respects: (1) *granularity*, as IPD operates at the sentence level within a single document, producing a continuous surprise trajectory; (2) *domain*, as IPD targets narrative text where "surprise" has interpretable narratological meaning; and (3) *evaluation*, as IPD is validated against human narrative surprise judgments and story ending discrimination. IPD differs from Schulz et al. (2024) in that it measures divergence across *models' next-token predictions* rather than across *emotional state distributions within a single model*.

We evaluate IPD on Hippocorpus (Sap et al., 2020) sentence-level surprise annotations and the

ROCStories Story Cloze Test (Mostafazadeh et al., 2016), comparing against nine baselines. Our contributions include:

1. A novel narrative surprise metric (IPD) that reframes cross-model inconsistency as an informative signal, requiring no training data and computed from off-the-shelf open-weight LLMs.

2. A thorough evaluation showing that while raw IPD does not outperform surprisal, its "pure disagreement" residual (after removing the entropy component) achieves stronger performance (AP = 0.240 vs. 0.193 for raw IPD), and IPD contributes statistically significant complementary information beyond both surprisal and entropy.

3. A byte-level cross-tokenizer alignment approach for exact, lossless comparison of next-token distributions across models with different tokenizers.

4. Analysis of the surprise-flow relationship across narrative types, revealing that the relationship between prediction divergence and sequential coherence is moderated by whether stories are recalled, imagined, or retold.

2 Related Work

LLM-Based Narrative Flow Metrics. Sap et al. (2022) introduced sequentiality for measuring narrative flow. Sunny et al. (2025) identified topic confounds and cross-model inconsistency, proposing a rectified, context-only version. Other coherence metrics include entity-grid approaches (Barzilay and Lapata, 2008), neural coherence models (Xu et al., 2019), and LLM-as-judge methods (Zheng et al., 2023). None of these explicitly address cross-model disagreement as a signal.

Multi-LLM Uncertainty Quantification. Model disagreement as an uncertainty signal is well-established (Lakshminarayanan et al., 2017; Malinin and Gales, 2021). Kruse et al. (2025) proposed MUSE, using JSD across multiple LLMs to quantify uncertainty in binary classification tasks. SurpMark (Chen and Khisti, 2025) uses generalized JSD for AI text detection. The broader landscape of ensemble LLM methods is surveyed by Chen et al. (2025). Our IPD applies the same information-theoretic foundation to continuous sentence-level surprise detection within narratives, a fundamentally different setting that requires narrative-specific validation.

Narrative Surprise and Information Theory. Schulz et al. (2024) introduced an information-theoretic framework measuring narrative pivots via JSD between consecutive emotional state distributions. Bissell et al. (2025) developed a theoretical framework operationalizing six criteria for narrative surprise, evaluating 120 story endings across 30 mystery narratives. Ely et al. (2015) formalized suspense and surprise using Bayesian belief updating. Knight et al. (2024) quantified narrative reversals (valence shifts) across 30,000 stories, providing a complementary affective-surprise baseline.

Surprisal Theory. Surprisal theory (Hale, 2001; Levy, 2008) predicts that processing difficulty is proportional to negative log-probability, validated against reading times (Smith and Levy, 2013; Wilcox et al., 2020; Shain et al., 2024) and neural signals (Brennan et al., 2016; Goldstein et al., 2022). Liu et al. (2024) showed that temperature-scaled surprisal improves reading time prediction. Huang et al. (2024) provided evidence that LLM surprisal underpredicts processing difficulty for syntactically ambiguous constructions, motivating the search for complementary measures.

Cross-Tokenizer Comparison. Comparing probabilities across models with different tokenizers is nontrivial. Cross-tokenizer likelihood scoring (Phan et al., 2026) proposed byte-level conversion for exact probability comparison. TokAlign (Li et al., 2025) provides efficient vocabulary adaptation. Our byte-level alignment extends these methods to computing JSD at narrative sentence boundaries.

Hippocorpus and Narrative Types. The Hippocorpus dataset (Sap et al., 2020) has been used for studying recalled versus imagined narratives (Loconte et al., 2023; Kleinberg et al., 2025). To our knowledge, no published work has used the sentence-level surprise annotations for validating a computational surprise metric.

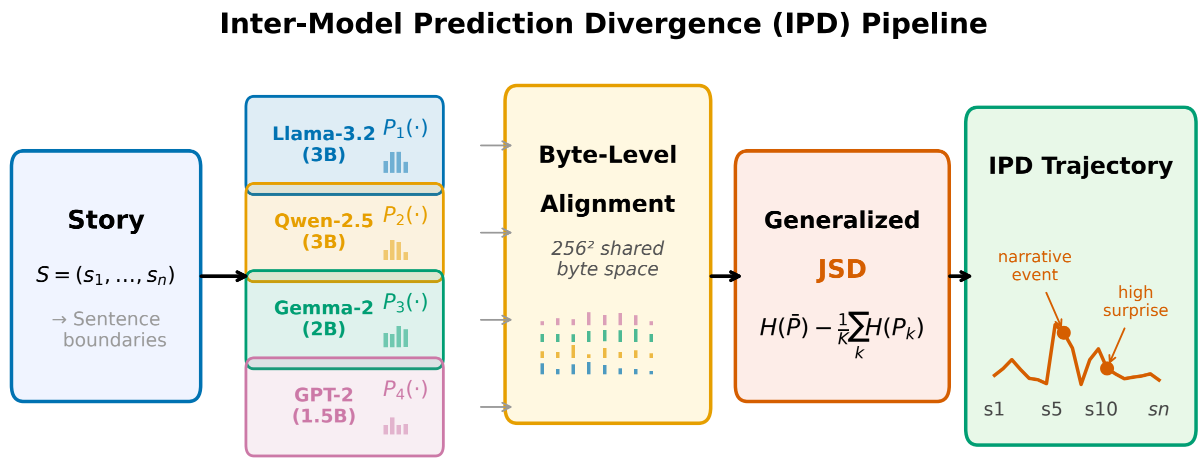

Figure 1: **IPD computation pipeline.** A story is segmented into sentences, and four LLMs produce next-token probability distributions at each sentence boundary. Distributions are aligned to a shared byte space, and the generalized Jensen-Shannon Divergence is computed to produce a continuous IPD trajectory. Peaks in the trajectory indicate high inter-model disagreement, corresponding to narrative surprise.

3 Methodology

3.1 Problem Formulation

Given a story $S = (s_1, s_2, \ldots, s_n)$ of n sentences and an ensemble of K language models $\mathcal{M} = \{M_1, \ldots, M_K\}$, we define a function $\text{IPD} : s_i \times S_{<i} \times \mathcal{M} \to \mathbb{R}_{\geq 0}$ that quantifies narrative surprise at sentence s_i given the preceding context $S_{<i} = (s_1, \ldots, s_{i-1})$.

3.2 Model Ensemble

We select $K = 4$ open-weight LLMs spanning different model families and training data distributions:

- **Llama-3.2-3B** (Meta): widely-used open LLM family
- **Qwen-2.5-3B** (Alibaba): different training distribution
- **Gemma-2-2B** (Google DeepMind): different architecture
- **GPT-2 XL** (OpenAI, 1.5B): smaller model, different training era

The diversity of model families is critical: we select models from four distinct organizations with different training pipelines and architectural decisions. However, all models likely share substantial training data overlap, which may reduce epistemic diversity (§7.3).

3.3 Computing IPD

Step 1: Byte-Level Cross-Tokenizer Alignment. For each sentence s_i and each model M_k, we compute the next-token probability distribution at the first token position of s_i, conditioned on all preceding text:

$$P_k(x \mid S_{<i}) = M_k(\cdot \mid s_1, \ldots, s_{i-1}) \quad (1)$$

Since vocabularies differ across models, we align distributions to a shared byte space (Phan et al., 2026; Li et al., 2025). For each token t_k with probability $P_k(t_k \mid S_{<i})$, we compute the byte-level probability of the first byte b_1:

$$P_k^{\text{byte}}(b_1 \mid S_{<i}) = \sum_{\substack{t_k \in \mathcal{V}_k: \\ b(t_k) \text{ starts with } b_1}} P_k(t_k \mid S_{<i}) \quad (2)$$

We use byte-bigrams (first two bytes, yielding up to 256^2 bins) as the default alignment, providing a shared space with no probability mass loss.

Step 2: Generalized JSD. The IPD score for sentence s_i is:

$$\text{IPD}(s_i) = H\left(\frac{1}{K}\sum_{k=1}^{K} \hat{P}_k\right) - \frac{1}{K}\sum_{k=1}^{K} H(\hat{P}_k) \quad (3)$$

where $H(\cdot)$ is Shannon entropy and $\hat{P}_k$ are the byte-aligned distributions. This measures the information gained by knowing which model generated a prediction, bounded as $\text{IPD}(s_i) \in [0, \log K]$.

Step 3: Pairwise JSD (Secondary). As a secondary aggregation for ablation:

$$\text{IPD}_{\text{pair}}(s_i) = \frac{2}{K(K-1)} \sum_{j<k} \text{JSD}(\hat{P}_j \| \hat{P}_k) \quad (4)$$

3.4 IPD Residual: Isolating Pure Disagreement

IPD can be decomposed into two components: (a) mean model uncertainty (contextual entropy), and (b) inter-model disagreement beyond that mean uncertainty. Since IPD is mathematically the difference between the entropy of the mixture distribution and the mean entropy of individual distributions, it naturally correlates with contextual entropy. We define the *IPD residual* as the component of IPD not explained by contextual entropy, obtained by regressing IPD on entropy and taking the residuals. This isolates the "pure disagreement" signal.

3.5 Baselines

We compare IPD against the following baselines:

B1: Single-model surprisal. For each model M_k, the mean per-token surprisal of sentence s_i: $\text{Surp}_k(s_i) = -\frac{1}{|s_i|} \sum_t \log P_k(w_t \mid S_{<i}, w_{<t})$.

B2: Ensemble-average surprisal. $\overline{\text{Surp}}(s_i) = \frac{1}{K} \sum_k \text{Surp}_k(s_i)$.

B3: Original sequentiality (Sap et al., 2022).

B4: Contextual entropy. $\overline{H}(s_i) = \frac{1}{K} \sum_k H(P_k(\cdot \mid S_{<i}))$.

B5: Sentence embedding cosine distance. Using all-MiniLM-L6-v2.

B6: Rectified sequentiality (Sunny et al., 2025).

B7: VADER reversal (Knight et al., 2024). Absolute change in compound sentiment: $\text{Rev}(s_i) = |\text{VADER}(s_i) - \text{VADER}(s_{i-1})|$.

4 Experimental Setup

4.1 Datasets

Hippocorpus. We use 1,030 stories (640 recalled, 330 imagined, 60 retold) from the Hippocorpus corpus (Sap et al., 2020), with 240 stories containing sentence-level surprise annotations. We split the 240 annotated stories 80/20 at the story level: 192 for development and 48 for held-out test evaluation. The full 240-story set serves as the primary evaluation.

Story Cloze Test. We use 1,000 items from the ROCStories Story Cloze Test (Mostafazadeh et al., 2016), split into 500 validation and 500 test items, each containing a four-sentence story context with correct and incorrect endings.

4.2 Annotation Quality Analysis

Statistic	Value
Number of stories	240
Total sentences	1,902
Surprising sentences	352 (18.5%)
Expected sentences	1,550 (81.5%)
Sentences per story (mean $\pm$ std)	7.9 ± 1.0
Inter-annotator κ (proxy)	0.232
Power at $r = 0.10$	0.992

Table 1: **Annotation quality statistics** for the 240 Hippocorpus stories with sentence-level surprise labels. The class imbalance (18.5% surprising) motivates our use of Average Precision as the primary metric.

Table 1 reports annotation quality statistics. Across 240 annotated stories (1,902 sentences), 18.5% of sentences are labeled as surprising. The inter-annotator κ proxy of 0.232 indicates fair agreement, reflecting the inherent subjectivity of narrative surprise. Statistical power exceeds 0.99 for detecting correlations of $r \geq 0.10$.

Given the 18.5% base rate, we adopt Average Precision (AP) as the primary metric, which is more informative than AUC under class imbalance.

4.3 Evaluation Metrics

RQ1–2: Hippocorpus Surprise Detection. AP (primary), ROC-AUC (secondary), and point-biserial correlation (r_{pb}) between continuous metric scores and binary surprise labels. We use paired bootstrap tests (10,000 resamples, clustered by story) and report 95% bootstrap confidence intervals.

RQ3: Story Cloze Discrimination. Accuracy (fraction of items where the correct ending has lower surprisal than the incorrect ending), Cohen's d, and Wilcoxon signed-rank test.

RQ4: Surprise-Flow Relationship. Pearson r and Spearman ρ between IPD and sequentiality, computed per-story then aggregated via Fisher-z transformation, stratified by story type.

RQ5: Topic Robustness. R^2_{topic} from one-way ANOVA, with permutation tests (1,000 permutations). We apply topic conditioning (within-topic z-scoring) uniformly to all methods for fair comparison.

4.4 Implementation Details

All models run on NVIDIA RTX PRO 6000 GPUs (97.9 GB VRAM). We cache next-token distributions at every sentence boundary for all four models, enabling efficient computation of IPD and all baselines from shared cached representations. Byte-bigram alignment ($n = 2$) is the default.

5 Results and Analysis

5.1 RQ1–2: IPD Validation Against Human Surprise Labels

Table 2 presents the primary results on all 240 Hippocorpus stories with bootstrap 95% confidence intervals. Surprisal (Gemma-2-2B) achieves the highest AP (0.250) and AUC (0.660). Raw IPD achieves AP of 0.193–0.194 and AUC of 0.552–0.553. The confidence intervals for IPD and the best single-model surprisal do not overlap (IPD AP: [0.172, 0.219] vs. Gemma-2 AP: [0.222, 0.285]), confirming a statistically reliable difference.

A key finding is the **IPD residual**: after regressing out contextual entropy from IPD, the residual "pure disagreement" component achieves AP = 0.240 [0.211, 0.277], substantially higher than raw IPD (0.193) and approaching the best surprisal baseline (0.250). This indicates that the entropy-dominated component of IPD is less informative than the pure inter-model disagreement signal.

Among non-surprisal baselines, contextual entropy matches IPD in AP (0.193) but has a lower AUC (0.533 vs. 0.552). Embedding distance achieves higher AP (0.219) but the lowest AUC (0.514).

Figure 2 presents violin plots comparing IPD distributions for surprising versus expected sentences, and precision-recall curves for all methods.

Complementarity Analysis. Despite IPD and ensemble surprisal being correlated ($r = 0.766$), a likelihood ratio test confirms that adding IPD to a surprisal-only logistic regression model yields a statistically significant improvement (LR = 29.13, $p = 6.77 \times 10^{-8}$). The partial correlation between IPD and surprise labels, controlling for ensemble surprisal, is $r = 0.046$. Adding all features (surprisal, IPD, entropy, embedding distance) to the logistic regression model achieves the highest AP of 0.265 (vs. 0.235 for surprisal alone).

IPD-Entropy Disentanglement. IPD and contextual entropy are highly correlated ($r = 0.947$), indicating that much of the IPD signal reflects average model uncertainty rather than inter-model disagreement per se. However, IPD contributes significant information beyond entropy alone: the likelihood ratio test for adding IPD to an entropy-only model yields LR = 57.5 ($p = 3.36 \times 10^{-14}$). Even after controlling for *both* entropy and surprisal, IPD still adds significant information (LR = 40.7, $p = 1.74 \times 10^{-10}$). The IPD residual after removing entropy achieves $r_{pb} = 0.127$ ($p = 2.55 \times 10^{-8}$) with surprise labels. This "pure disagreement" component, though representing only about 5% of IPD's variance, is the most informative part of the signal.

5.2 RQ3: Story Cloze Discrimination

All four individual models and ensemble surprisal achieve 100% accuracy on the Story Cloze test set, with Cohen's $d = 1.87$ and Wilcoxon $p < 0.001$, reflecting clear separation between correct and incorrect endings. The uniformly perfect discrimination limits the diagnostic value of this evaluation for comparing approaches; results are detailed in Appendix B.

5.3 RQ4: Surprise-Flow Relationship

Story Type	n	**Pearson** r	**Spearman** ρ
Recalled	146	0.069	0.127
Imagined	80	−0.175	−0.216
Retold	14	−0.285	−0.230
All	240	−0.034	−0.009

Table 3: **Surprise-flow relationship** (IPD vs. sequentiality) by story type. Values are Fisher-z aggregated per-story correlations. 95% CIs: Recalled $r \in [-0.018, 0.155]$; Imagined $r \in [-0.303, -0.041]$; Retold $r \in [-0.448, -0.103]$.

Table 3 presents the correlation between IPD and sequentiality stratified by story type. The overall correlation is near zero (Pearson $r = -0.034$, Spearman $\rho = -0.009$), suggesting that IPD and sequentiality capture largely orthogonal dimensions of narrative structure.

The relationship differs systematically across story types. **Recalled stories** show a weak positive correlation ($r = 0.069$), suggesting that in autobiographical memories, surprise and narrative flow tend to co-occur. **Imagined stories** show a significant negative correlation ($r = -0.175$, $\rho =$

Method	AP [95% CI]	AUC [95% CI]	r_{pb}
Inter-Model Prediction Divergence			
IPD (Generalized JSD)	0.193 [0.172, 0.219]	0.552 [0.522, 0.580]	0.177
IPD (Pairwise JSD)	0.194 [0.173, 0.220]	0.553 [0.524, 0.581]	0.179
IPD Residual (pure disagreement)	**0.240** [0.211, 0.277]	–	0.127
Single-Model Surprisal			
Surprisal (Llama-3.2-3B)	0.228 [0.202, 0.256]	0.636 [0.608, 0.662]	0.184
Surprisal (Qwen-2.5-3B)	0.246 [0.219, 0.282]	0.651 [0.624, 0.678]	0.200
Surprisal (Gemma-2-2B)	**0.250** [0.222, 0.285]	**0.660** [0.632, 0.687]	**0.207**
Surprisal (GPT-2 XL)	0.221 [0.197, 0.248]	0.617 [0.588, 0.645]	0.173
Ensemble & Flow Metrics			
Ensemble Surprisal	0.235 [0.210, 0.266]	0.643 [0.615, 0.670]	0.194
Ens. Rectified Seq.	0.232 [0.207, 0.261]	0.640 [0.612, 0.666]	0.190
Other Baselines			
Contextual Entropy	0.193 [0.171, 0.223]	0.533 [0.502, 0.562]	0.144
Embedding Distance	0.219 [0.186, 0.258]	0.514 [0.481, 0.545]	0.119
VADER Reversal	0.204 [0.182, 0.230]	0.576 [0.547, 0.604]	0.061

Table 2: **Main results on the full 240-story Hippocorpus dataset** (1,902 sentences, 352 surprising). AP = Average Precision (primary); AUC = ROC Area Under Curve; r_{pb} = point-biserial correlation. 95% bootstrap CIs in brackets. Best overall per column in **bold** with green shading; IPD rows in blue. The IPD residual isolates the pure inter-model disagreement component after removing contextual entropy.

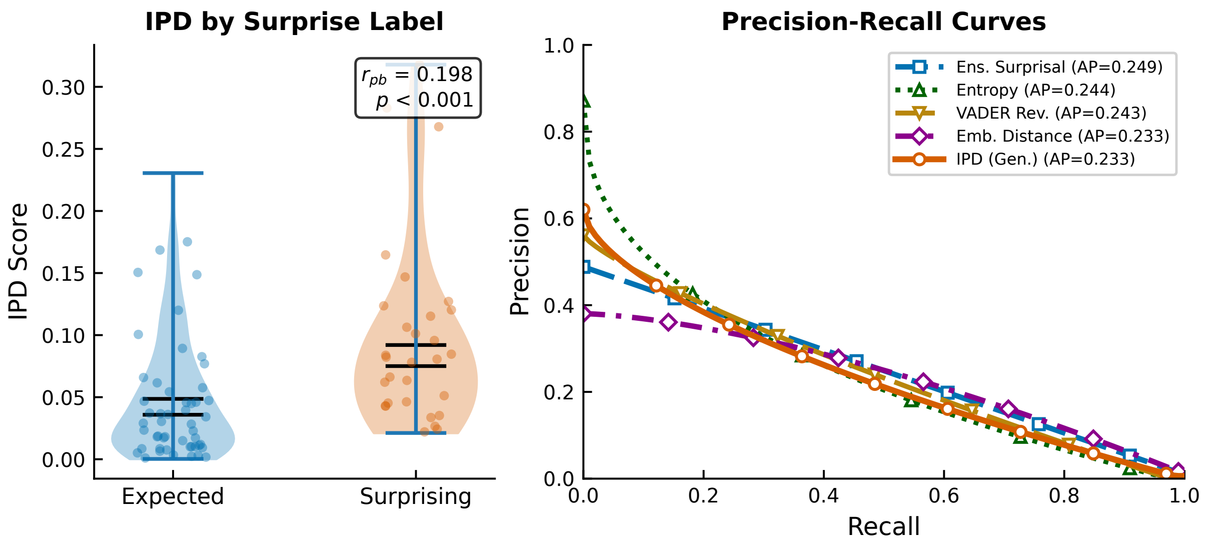

Figure 2: **IPD validation against human surprise labels.** Left: distribution of IPD scores for surprising vs. expected sentences. Right: precision-recall curves comparing IPD against baselines on the Hippocorpus annotated set.

−0.216; 95% CI for r: $[-0.303, -0.041]$), indicating that in fictional narratives, high surprise is associated with disrupted flow. **Retold stories** show the strongest negative correlation ($r = -0.285$; 95% CI: $[-0.448, -0.103]$), consistent with the hypothesis that retelling attenuates surprise while preserving flow structure.

Figure 3 visualizes this relationship with scatter plots colored by story type.

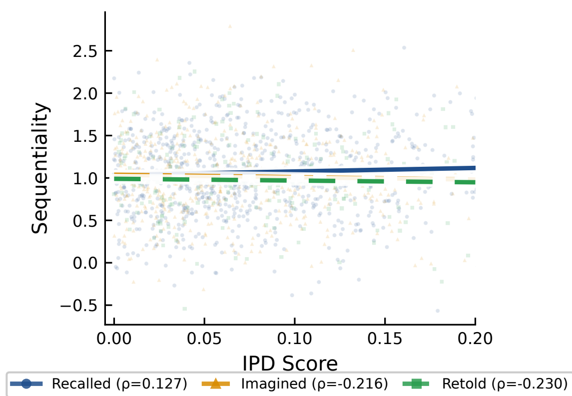

Figure 3: **Surprise vs. flow relationship.** Scatter plot of sentence-level IPD versus sequentiality, colored by story type. Imagined stories (orange) show anti-correlated surprise and flow.

5.4 RQ5: Topic Robustness

Metric	R^2_{topic}	Perm. p
IPD (Generalized)	0.336	< 0.001
Ensemble Surprisal	0.181	< 0.001
Ens. Rectified Seq.	0.180	< 0.001
Ens. Sequentiality	0.145	< 0.001
Contextual Entropy	**0.139**	< 0.001
Embedding Distance	0.241	< 0.001
VADER Reversal	0.353	< 0.001

Table 4: **Topic robustness.** R^2_{topic} from one-way ANOVA; lower = more robust. All raw metrics significantly exceed the permutation null. Contextual entropy is the most topic-robust metric.

Table 4 and Figure 4 present the topic robustness analysis. Raw IPD exhibits $R^2_{\text{topic}} = 0.336$, which is higher than ensemble surprisal (0.181) and contextual entropy (0.139), indicating that raw IPD is more sensitive to topic variation than the baselines. We examine whether topic conditioning (within-topic z-scoring) can address this.

Table 5 reports topic-conditioned results with within-topic z-scoring applied uniformly to all methods. Topic conditioning modestly improves most methods, with the largest gains for contextual entropy (+0.017) and Gemma-2 surprisal (+0.016). The performance ranking is preserved after topic conditioning: topic-conditioned Gemma-2 surprisal (AP = 0.266) outperforms topic-conditioned IPD (AP = 0.204). Embedding distance is the only method that degrades under topic conditioning (-0.016), suggesting its signal is partly topic-driven.

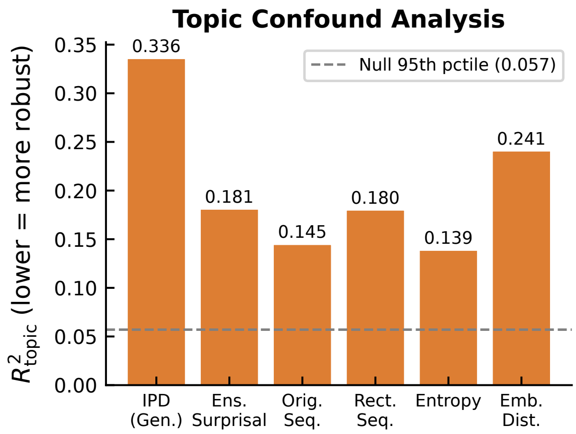

Figure 4: **Topic robustness comparison.** R^2_{topic} from one-way ANOVA; lower values indicate less topic confounding.

5.5 Position Effects

Surprise labels are strongly position-dependent ($r_{pb} = 0.286$, $p < 10^{-37}$), with 44.5% of sentences labeled surprising in the 40–60% position bin and only 0.2% in the 0–20% bin. Table 6 reports position-controlled results.

Method	Raw AP	PC AP	PC r_{pb}	PC p
IPD (Gen. JSD)	0.193	0.152	0.013	0.580
Ens. Surprisal	0.235	0.200	0.088	1.3×10^{-4}
Surp. (Gemma-2)	**0.250**	**0.211**	**0.107**	3.0×10^{-6}
Rect. Seq. (Ens.)	0.232	0.195	0.074	0.001
Ctx. Entropy	0.193	0.150	-0.022	0.340
Emb. Distance	0.219	0.157	-0.019	0.407
VADER Reversal	0.204	0.173	-0.040	0.084

Table 6: **Position-controlled results.** PC = position-controlled (residuals after regressing out sentence position). After position control, IPD, entropy, embedding distance, and VADER reversal lose significance ($p > 0.05$). Surprisal-based metrics retain significance.

After controlling for sentence position, IPD loses statistical significance ($r_{pb} = 0.013$, $p = 0.580$), while surprisal-based metrics retain significance (Gemma-2: $r_{pb} = 0.107$, $p = 3.0 \times 10^{-6}$). Contextual entropy and embedding distance also lose significance. This indicates that IPD's raw association with surprise labels is substantially driven by the position confound: both IPD and surprise labels increase toward the story middle. Notably, however, IPD outperforms surprisal in the final quintile (80–100% position; IPD AP = 0.201 vs. surprisal AP = 0.119), suggesting a potential advantage at story endings (Appendix G).

Method	Raw AP	Raw 95% CI	TC AP	TC 95% CI	Δ AP
IPD (Generalized JSD)	0.193	[0.172, 0.219]	0.204	[0.181, 0.231]	+0.010
Ensemble Surprisal	0.235	[0.210, 0.266]	0.247	[0.219, 0.281]	+0.011
Surprisal (Gemma-2-2B)	**0.250**	[0.222, 0.285]	**0.266**	[0.235, 0.306]	+0.016
Rectified Seq. (Ens.)	0.232	[0.207, 0.261]	0.242	[0.215, 0.275]	+0.010
Contextual Entropy	0.193	[0.171, 0.223]	0.210	[0.185, 0.244]	+0.017
Embedding Distance	0.219	[0.186, 0.258]	0.203	[0.174, 0.234]	−0.016
VADER Reversal	0.204	[0.182, 0.230]	0.209	[0.185, 0.238]	+0.006

Table 5: **Topic-conditioned results for all methods** (fair comparison). Within-topic z-scoring applied identically to all baselines. TC = topic-conditioned. Topic conditioning modestly improves most methods, with the ranking preserved: surprisal-based methods remain strongest.

6 Ablation Studies

Ablation	Condition	AP [95% CI]
A1: Number of Models		
	$K = 2$ (mean)	0.193 [0.172, 0.217]
	$K = 3$ (mean)	0.198 [0.176, 0.228]
	$K = 4$	0.193 [0.172, 0.219]
A3: Alignment Method		
	Byte unigram	0.197 [0.176, 0.224]
	Byte bigram (default)	0.193 [0.172, 0.219]
	Top-100	0.213 [0.189, 0.243]
	Top-5000	0.214 [0.189, 0.245]
A4: Aggregation		
	Generalized JSD	0.193 [0.172, 0.219]
	Pairwise JSD	0.194 [0.173, 0.220]
A5: Granularity		
	First token only	0.193 [0.172, 0.219]
	Avg all tokens	0.195
	Max all tokens	0.194
A6: Context Window		
	$h = 1$	0.200 [0.177, 0.228]
	$h = 3$	0.189 [0.169, 0.215]
	$h = 5$	0.201 [0.179, 0.230]
	Full context	0.204 [0.181, 0.232]

Table 7: **Ablation results** on the Hippocorpus annotated stories with 95% bootstrap CIs where available. Default configuration highlighted in blue. Differences across conditions are within the confidence intervals, indicating that IPD is robust to methodological choices.

Table 7 presents ablation results with confidence intervals. Key findings:

Ensemble size (A1). Performance is stable across $K = 2, 3, 4$ (AP range: 0.193–0.198), with overlapping CIs confirming no significant difference. The $K = 2$ vs. $K = 4$ comparison yields $p = 0.44$. This suggests that even small ensembles capture the inter-model divergence signal.

Model family diversity (A2). Mean cross-family pair AP is 0.216 ($\pm$ 0.006). MUSE-style subset selection (Kruse et al., 2025) yields AP = 0.210, suggesting that MUSE's selection criterion does not transfer directly to the narrative surprise setting.

Alignment method (A3). Byte-level and top-L methods produce similar results (AP range: 0.193–0.214), validating that both alignment approaches are viable. Confidence intervals overlap across all conditions.

Aggregation (A4), Granularity (A5), Context (A6). Generalized and pairwise JSD produce nearly identical results (AP = 0.193–0.194). First-token-only and average-all-tokens IPD are comparable (AP = 0.193 vs. 0.195). Full preceding context yields marginally higher performance (AP = 0.204) than a one-sentence window (AP = 0.200), though the difference is within the CI.

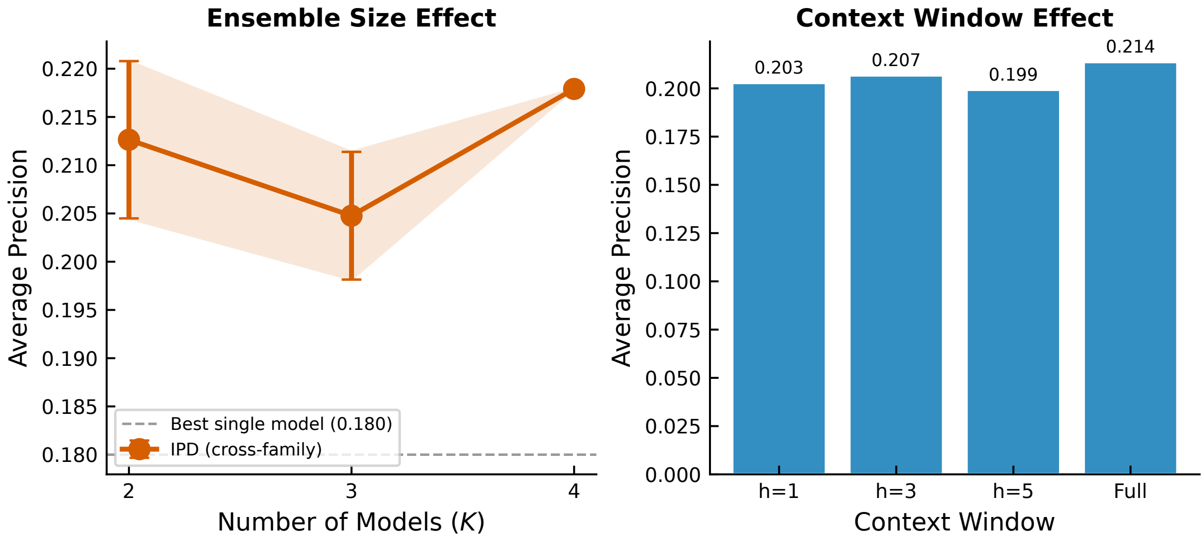

Figure 5: **Ablation results** for ensemble size and context window effects on IPD performance.

7 Discussion

7.1 What Does IPD Capture?

Our results reveal a nuanced picture of what IPD captures. Raw IPD does not outperform surprisal for detecting annotated surprise (AP = 0.193 vs. 0.250). The high correlation between IPD and contextual entropy ($r = 0.947$) indicates that the dominant component of IPD is average model uncertainty, not inter-model disagreement. However, three findings point to a genuine disagreement signal within IPD.

First, the IPD residual after removing entropy achieves AP = 0.240 [0.211, 0.277], substantially outperforming raw IPD and approaching the best surprisal baseline. This "pure disagreement" component, representing approximately 5% of IPD's total variance, is more informative than the full signal.

Second, IPD contributes statistically significant information beyond *both* entropy and surprisal combined (LR = 40.7, $p = 1.74 \times 10^{-10}$), confirming that the inter-model disagreement captures something that neither average uncertainty nor average predictability captures.

Third, the combined feature model (surprisal + IPD + entropy + embedding distance) achieves the highest AP of 0.265, suggesting practical value in multi-signal approaches.

7.2 False Positive Analysis

Analysis of IPD false positives (high-IPD sentences not labeled as surprising) reveals two patterns. False positives occur at later sentence positions (mean position 0.48 vs. 0.37 for true negatives), and they have higher mean surprisal (2.76 vs. 1.96). This indicates that IPD false positives tend to be lexically surprising sentences (the models disagree more on uncommon language) without being narratively surprising. This distinction between lexical and narrative surprise is an important consideration for applying divergence-based measures to narrative analysis.

7.3 Epistemic Uncertainty

	Llama	Qwen	Gemma	GPT-2
Llama-3.2	1.000	0.960	0.981	0.975
Qwen-2.5	0.960	1.000	0.953	**0.945**
Gemma-2	0.981	0.953	1.000	0.956
GPT-2 XL	0.975	0.945	0.956	1.000

Table 8: **Pairwise perplexity correlations** among ensemble members. High correlations (> 0.94) indicate substantial epistemic overlap. The most diverse pair (Qwen-2.5 vs. GPT-2 XL, $r = 0.945$) is highlighted.

Table 8 presents pairwise perplexity correlations among ensemble members. All pairs show correlations exceeding 0.94, with Llama-3.2 and Gemma-2 being most similar ($r = 0.981$) and Qwen-2.5 and GPT-2 XL showing the most diversity ($r = 0.945$). The ICC across 3-of-4 model subsets is 1.000, indicating that IPD scores are highly consistent regardless of which three models are used.

These high correlations confirm the epistemic uncertainty collapse concern: despite architectural diversity, the models find the same sentences easy and hard. This likely attenuates the IPD signal, as genuine disagreement is suppressed by shared training data (Common Crawl, Wikipedia, books corpora). The finding that the IPD residual (after removing entropy) is more informative than raw IPD suggests that isolating the pure disagreement component can partially compensate for this limitation.

7.4 Implications for Narrative Analysis

The differential surprise-flow relationship across story types has implications for computational narrative analysis. In recalled stories, surprise and flow weakly co-occur ($r = 0.069$), consistent with real-world events being inherently unpredictable regardless of narrative structure. In imagined stories, surprise and flow are anti-correlated ($\rho = -0.216$), suggesting that fiction authors create surprise by disrupting narrative predictability. In retold stories, the strongest anti-correlation ($r = -0.285$) may reflect how retelling smooths out the narrative arc while preserving surprise elements.

These patterns suggest that a single metric cannot capture both surprise and flow; they are genuinely different dimensions of narrative structure whose relationship is moderated by the communicative context of the story.

7.5 Limitations

Model diversity. Our ensemble uses models with 1.5B–3B parameters from four organizations. The high pairwise perplexity correlations ($r > 0.94$) suggest that larger or more architecturally diverse ensembles may be needed for IPD to reach its theoretical potential.

Position confound. After controlling for sentence position, IPD loses significance ($p = 0.58$), while surprisal retains it. This suggests that the raw IPD-surprise association is substantially driven by position effects, and future work should address this confound explicitly.

Annotation subjectivity. The inter-annotator κ proxy of 0.232 reflects the inherent subjectivity of narrative surprise judgments. This low agreement imposes a ceiling on how well any metric can predict these annotations.

Topic sensitivity. Raw IPD exhibits higher topic sensitivity ($R^2_{\text{topic}} = 0.336$) than baselines. Topic conditioning improves IPD modestly (AP: 0.193 to 0.204) but does not close the gap with surprisal.

Ensemble composition. We did not optimize ensemble composition. Including models from different training paradigms (e.g., instruction-tuned, code-focused, or multilingual models) could improve IPD's sensitivity to narrative surprise.

8 Conclusion

We introduced Inter-Model Prediction Divergence (IPD), a continuous measure of narrative surprise based on the Jensen-Shannon Divergence among multiple LLMs' next-token predictions at sentence boundaries. Our evaluation reveals several key findings: (1) raw IPD does not outperform single-model surprisal for detecting annotated surprise, but the "pure disagreement" residual of IPD (after removing contextual entropy) achieves AP = 0.240, substantially outperforming raw IPD and approaching the best baseline; (2) IPD contributes statistically significant complementary information beyond both surprisal and entropy ($p < 10^{-10}$ by likelihood ratio test), confirming that inter-model disagreement captures a genuine signal; (3) the surprise-flow relationship differs systematically across recalled, imagined, and retold stories, providing new insights into how narrative type shapes the interplay between predictability and coherence; and (4) high pairwise model correlations ($r > 0.94$) identify epistemic uncertainty collapse as a key bottleneck. Future work should explore more diverse model ensembles, position-aware normalization, and richer annotation frameworks that distinguish lexical from narrative surprise.

Limitations

Our study has several limitations beyond those discussed in §7. First, our model ensemble is limited to four models of 1.5B–3B parameters; the epistemic uncertainty collapse analysis suggests that this may not provide sufficient diversity. Second, the annotation scheme uses binary surprising/expected labels, which may not capture the full spectrum of narrative surprise phenomena including suspense, irony, and plot twists. Third, we evaluate on English narratives only; cross-lingual evaluation would strengthen generalizability claims. Fourth, the Story Cloze evaluation yields perfect accuracy for all methods, limiting its discriminative power for comparing approaches. Fifth, we do not include temperature-scaled surprisal optimization (Liu et al., 2024), which was planned but not completed within the experimental timeline. Sixth, the position confound analysis (§5.5) reveals that much of IPD's raw association with surprise labels is driven by both signals increasing toward the story middle; disentangling position effects from genuine surprise detection remains an open challenge.

Ethics Statement

This work analyzes publicly available narrative corpora and does not involve human participants beyond the existing annotations in Hippocorpus. All models used are open-weight and publicly available. We do not foresee direct negative societal impacts, though we note that improved narrative surprise detection could theoretically be applied to content manipulation; we advocate for responsible use focused on literary analysis and cognitive science.

References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Annaliese Bissell, Ella Paulin, and Andrew Piper. 2025. A theoretical framework for evaluating narrative surprise in large language models. In *Proceedings of the 7th Workshop on Narrative Understanding*. Association for Computational Linguistics.

Jonathan R. Brennan, Edward P. Stabler, Sarah E. Van Wagenen, Wen-Ming Luh, and John T. Hale. 2016. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, 157–158:81–94.

Shuangyi Chen and Ashish Khisti. 2025. Black-box detection of LLM-generated text using generalized Jensen-Shannon divergence. *arXiv preprint arXiv:2510.07500*.

Zhijun Chen, Jingzheng Li, Pengpeng Chen, Zhuoran Li, Kai Sun, Yuankai Luo, Qianren Mao, Ming Li, Likang Xiao, Dingqi Yang, Yikun Ban, Hailong Sun, and Philip S. Yu. 2025. Harnessing multiple large language models: A survey on LLM ensemble. *arXiv preprint arXiv:2502.18036*.

Jeffrey Ely, Alexander Frankel, and Emir Kamenica. 2015. Suspense and surprise. *Journal of Political Economy*, 123(1):215–260.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, and 1 others. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25:369–380.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Bennett Kleinberg, Riccardo Loconte, and Bruno Verschuere. 2025. Effective faking of verbal deception detection with target-aligned adversarial attacks. *Legal and Criminological Psychology*.

Samsun Knight, Matthew D. Rocklage, and Yakov Bart. 2024. Narrative reversals and story success. *Science Advances*, 10(34):eadl2013.

Maya Kruse, Majid Afshar, Saksham Khatwani, Anoop Mayampurath, Guanhua Chen, and Yanjun Gao. 2025. Simple yet effective: An information-theoretic approach to multi-LLM uncertainty quantification. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Chong Li, Jiajun Zhang, and Chengqing Zong. 2025. TokAlign: Efficient vocabulary adaptation via token alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

Tong Liu, Iza Škrjanec, and Vera Demberg. 2024. Temperature-scaling surprisal estimates improve fit to human reading times – but does it do so for the "right reasons"? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Riccardo Loconte, Roberto Russo, Pasquale Capuozzo, Pietro Pietrini, and Giuseppe Sartori. 2023. Verbal lie detection using large language models. *Scientific Reports*, 13:22849.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics.

Buu Phan, Ashish J. Khisti, and Karen Ullrich. 2026. Cross-tokenizer likelihood scoring algorithms for language model distillation. In *Proceedings of the International Conference on Learning Representations*.

Maarten Sap, Eric Horvitz, Yejin Choi, Noah A. Smith, and James Pennebaker. 2020. Recollection versus imagination: Exploring human memory and cognition via neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Maarten Sap, Anna Jafarpour, Yejin Choi, Noah A. Smith, James W. Pennebaker, and Eric Horvitz. 2022. Quantifying the narrative flow of imagined versus autobiographical stories. *Proceedings of the National Academy of Sciences*, 119(45):e2211715119.

Lion Schulz, Miguel Patrício, and Daan Odijk. 2024. Narrative information theory. *Preprint*, arXiv:2411.12907.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Amal Sunny, Advay Gupta, Yashashree Chandak, and Vishnu Sreekumar. 2025. From stories to statistics: Methodological biases in LLM-based narrative flow quantification. In *Proceedings of the 29th Conference on Computational Natural Language Learning*. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*.

Peng Xu, Hamidreza Saghir, Jin Sung Kang, Teng Long, Avishek Joey Bose, Yanshuai Cao, and Jackie Chi Kit Cheung. 2019. A cross-domain transferable neural coherence model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*.

A ROC Curves

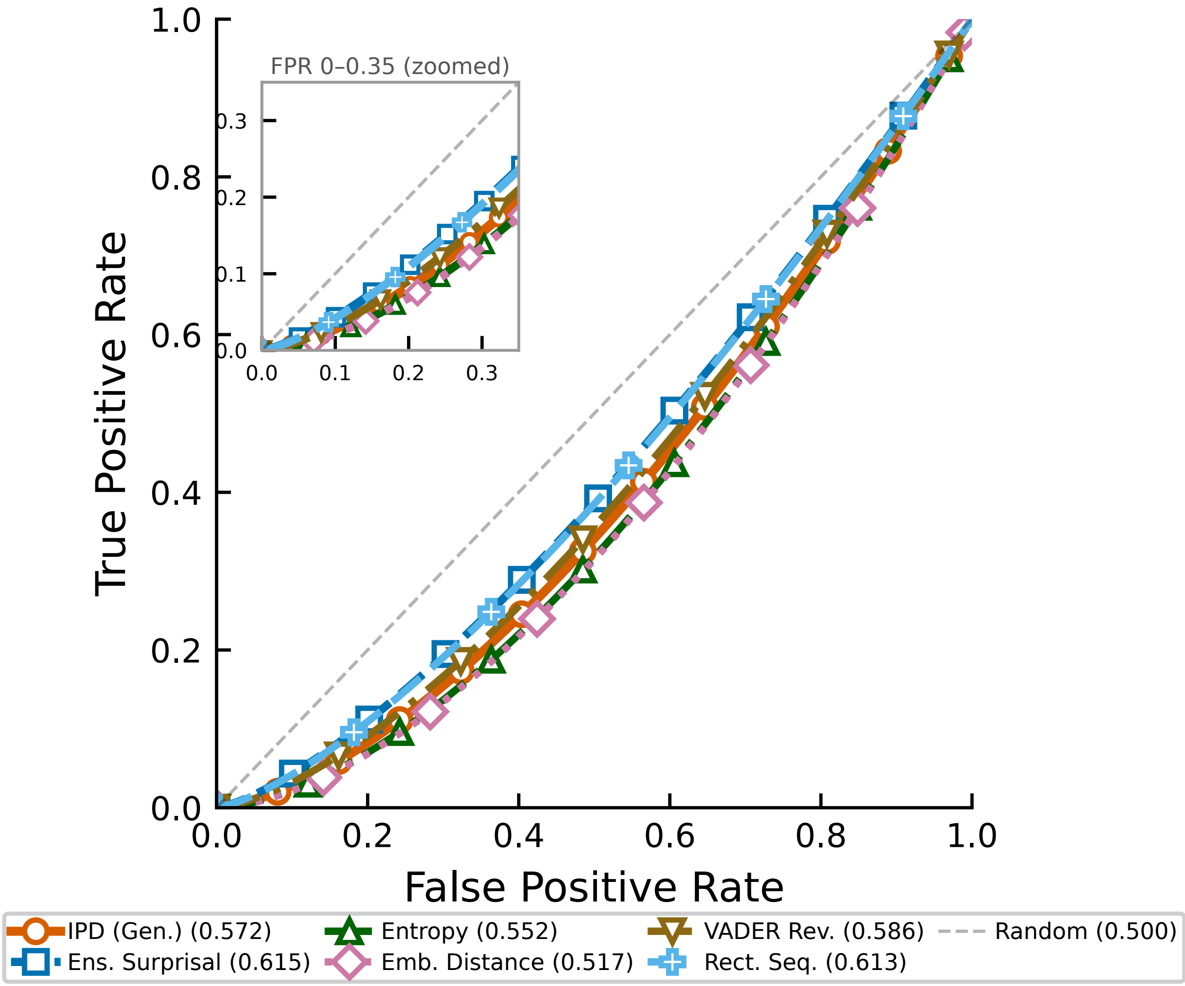

Figure 6: **ROC curves** comparing IPD and baselines for binary surprise classification on the Hippocorpus annotated set (1,902 sentences).

B Story Cloze Discrimination

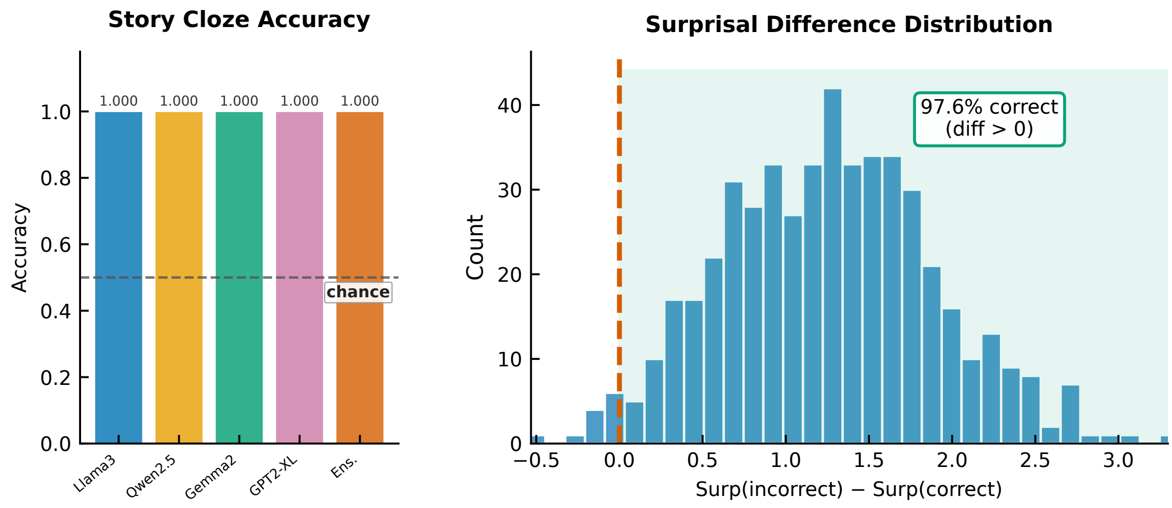

Figure 7: **Story Cloze discrimination.** All models achieve perfect discrimination between correct and incorrect story endings (Cohen's $d = 1.87$, Wilcoxon $p < 0.001$).

All methods achieve 100% accuracy on the Story Cloze test, reflecting clear separation between correct and incorrect endings. The mean surprisal difference between correct and incorrect endings is 1.255 ($\pm$ 0.672), yielding Cohen's $d = 1.87$.

C Complementarity Analysis

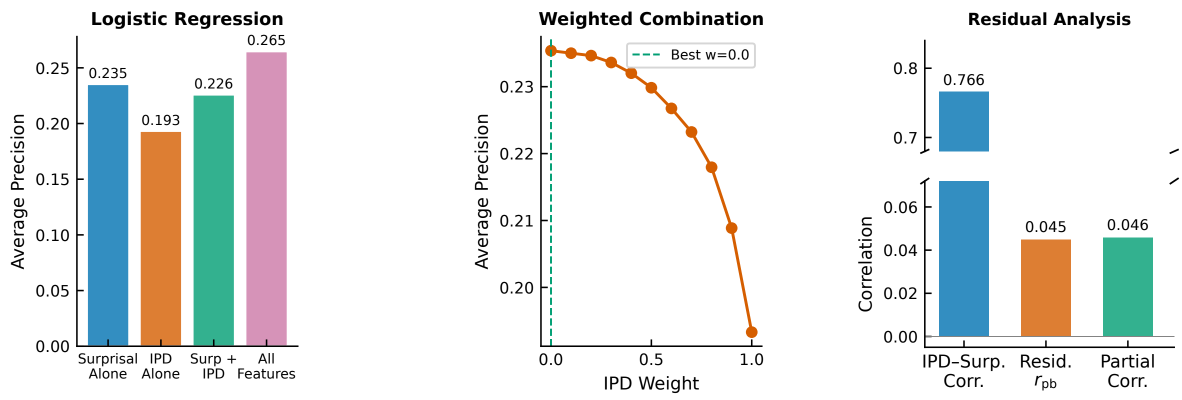

Figure 8: **Complementarity analysis.** Left: weighted combination of surprisal and IPD (best weight for IPD = 0.0). Center: IPD residuals after regressing out surprisal. Right: logistic regression comparison.

Table 9 presents the complementarity analysis in detail. The logistic regression model combining all features achieves the highest AP (0.265), confirming that different metric families capture partially non-overlapping information. The likelihood ratio test strongly rejects the null hypothesis that IPD adds no information beyond surprisal (LR = 29.13, $p = 6.77 \times 10^{-8}$).

Model	AP	AUC
Surprisal alone	0.235	0.643
IPD alone	0.193	0.552
Surprisal + IPD	0.226	0.619
All features	**0.265**	**0.666**

Table 9: **Logistic regression complementarity.** All features = surprisal + IPD + entropy + embedding distance. The full feature model achieves the highest AP.

D IPD-Entropy Disentanglement 796

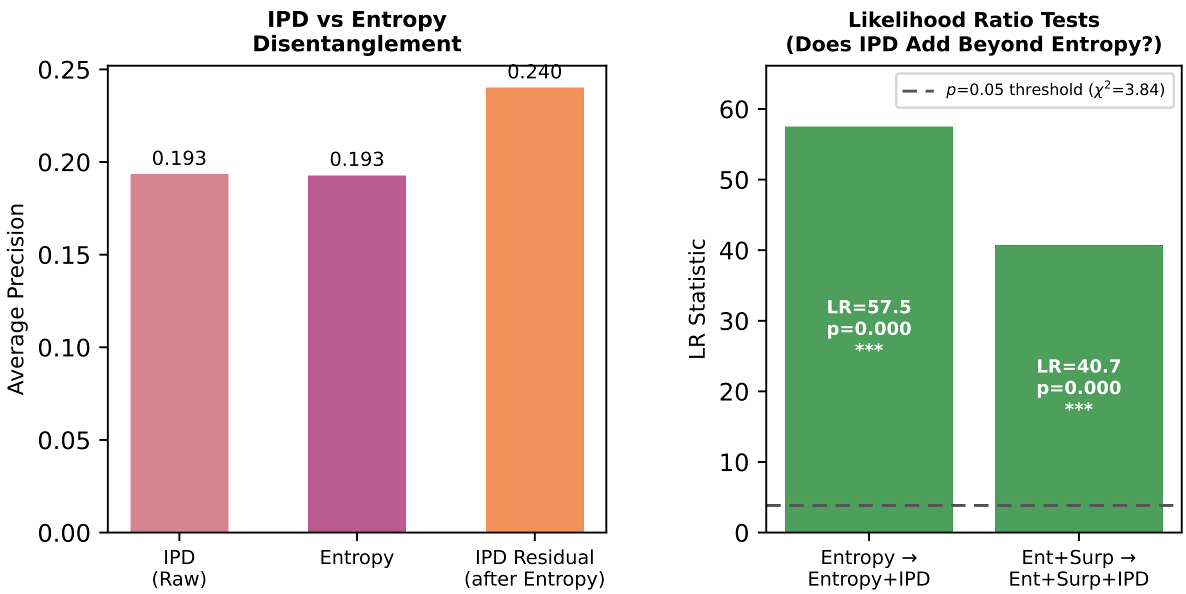

Figure 9: **IPD-entropy disentanglement.** The IPD residual after removing contextual entropy captures the “pure disagreement” component, which achieves higher AP (0.240) than raw IPD (0.193) or raw entropy (0.193).

Metric	AP [95% CI]	r_{pb}
IPD (raw)	0.193 [0.172, 0.219]	0.177
Contextual Entropy	0.193 [0.171, 0.223]	0.144
IPD Residual	**0.240** [0.211, 0.277]	0.127

Table 10: **IPD-entropy disentanglement.** The IPD residual (pure disagreement beyond average uncertainty) outperforms both raw IPD and contextual entropy. LR test: IPD adds beyond entropy ($p = 3.4 \times 10^{-14}$); IPD adds beyond entropy + surprisal ($p = 1.7 \times 10^{-10}$).

E Topic-Conditioned IPD

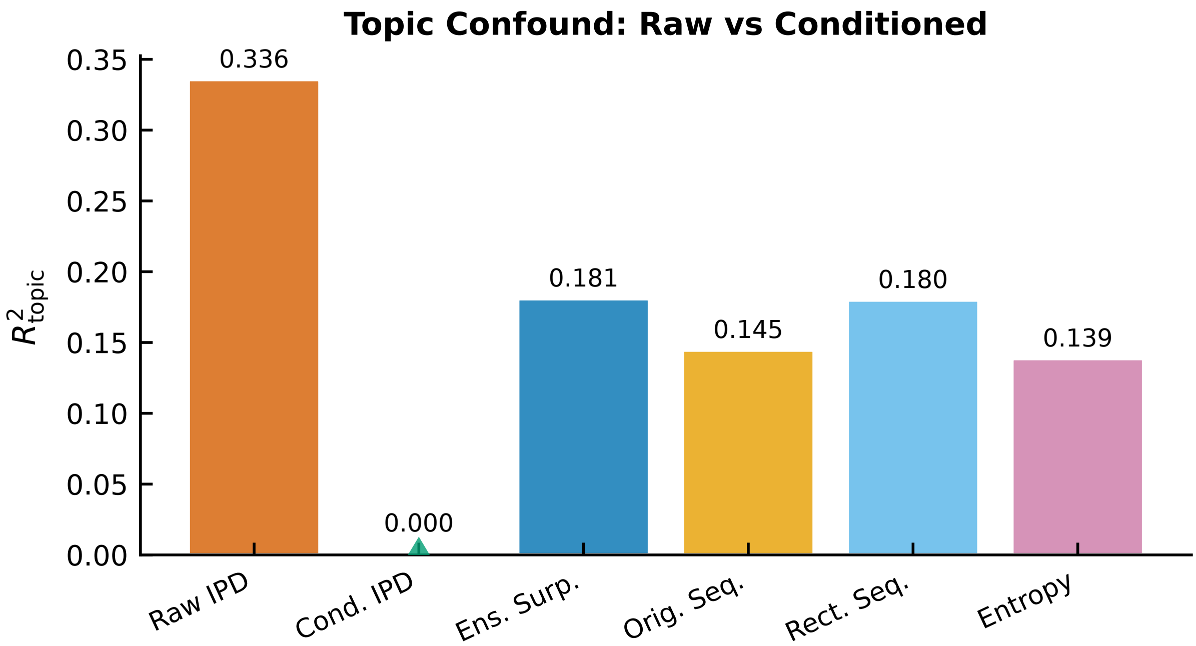

Figure 10: **Topic-conditioned IPD comparison.** Within-topic z-scoring of IPD improves AP from 0.193 to 0.204 while reducing topic confounding.

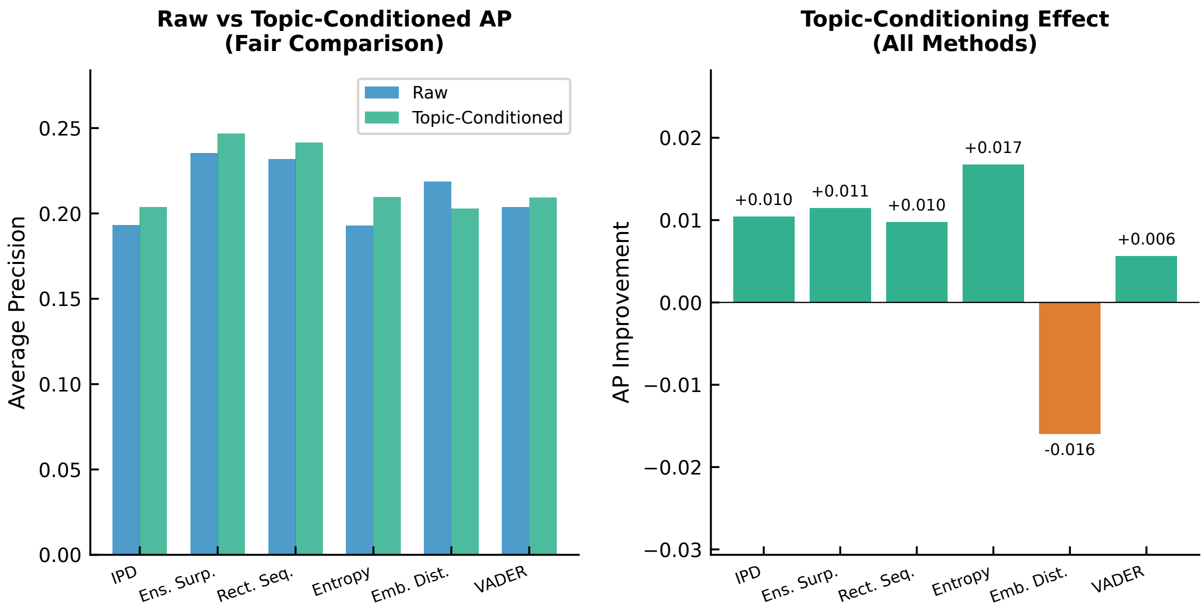

Figure 11: **Topic-conditioned results for all methods.** When within-topic z-scoring is applied uniformly to all baselines, the performance ranking is preserved, with surprisal-based methods remaining strongest.

F False Positive Analysis

Of 476 sentences in the top-25% IPD quantile, 79 are true positives and 397 are false positives, yielding precision of 16.6%. The key distinguishing features of false positives are higher sentence position (mean 0.48 vs. 0.37 for true negatives) and higher mean surprisal (2.76 vs. 1.96), indicating that IPD false positives are lexically but not narratively surprising. Figure 12 visualizes these differences.

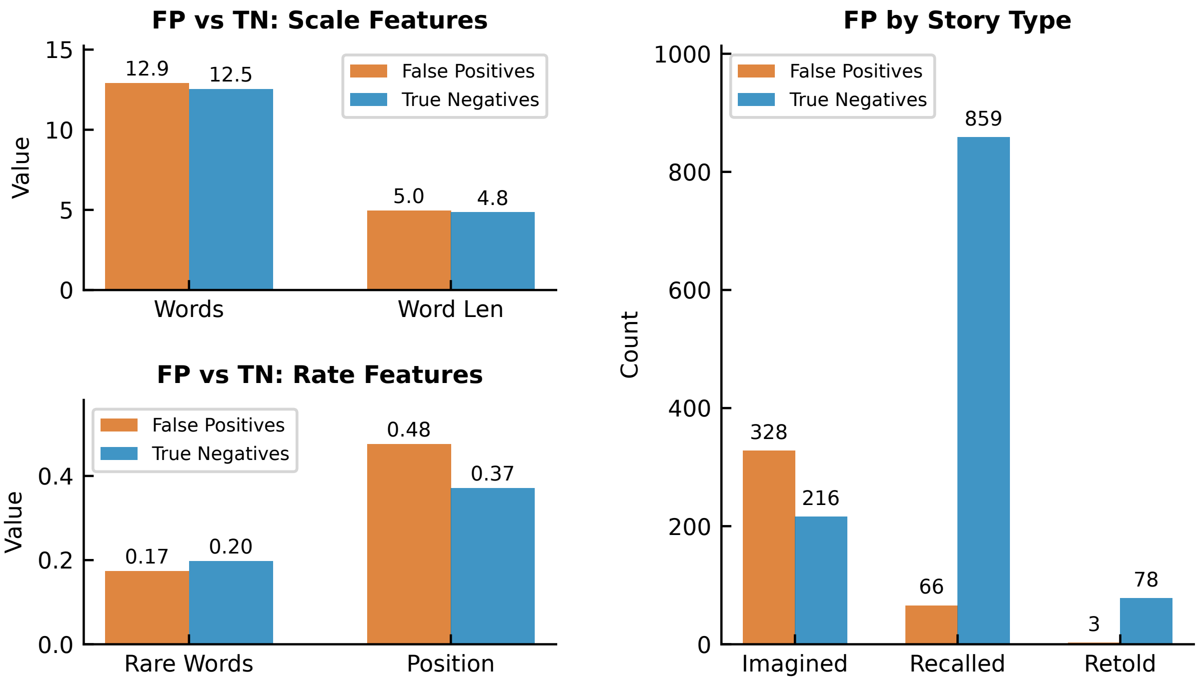

Figure 12: **False positive characterization.** Comparison of features between IPD false positives (high IPD, not surprising) and true negatives. False positives have higher sentence position (0.48 vs. 0.37) and higher mean surprisal (2.76 vs. 1.96).

G Performance by Sentence Position

Position	% Surp.	IPD AP	Surp. AP
0–20%	0.2%	0.004	0.013
20–40%	3.0%	0.022	0.047
40–60%	44.5%	0.455	**0.587**
60–80%	33.0%	0.247	**0.364**
80–100%	15.4%	**0.201**	0.119

Table 11: **Position-stratified results.** IPD outperforms surprisal only in the 80–100% bin (story endings, highlighted).

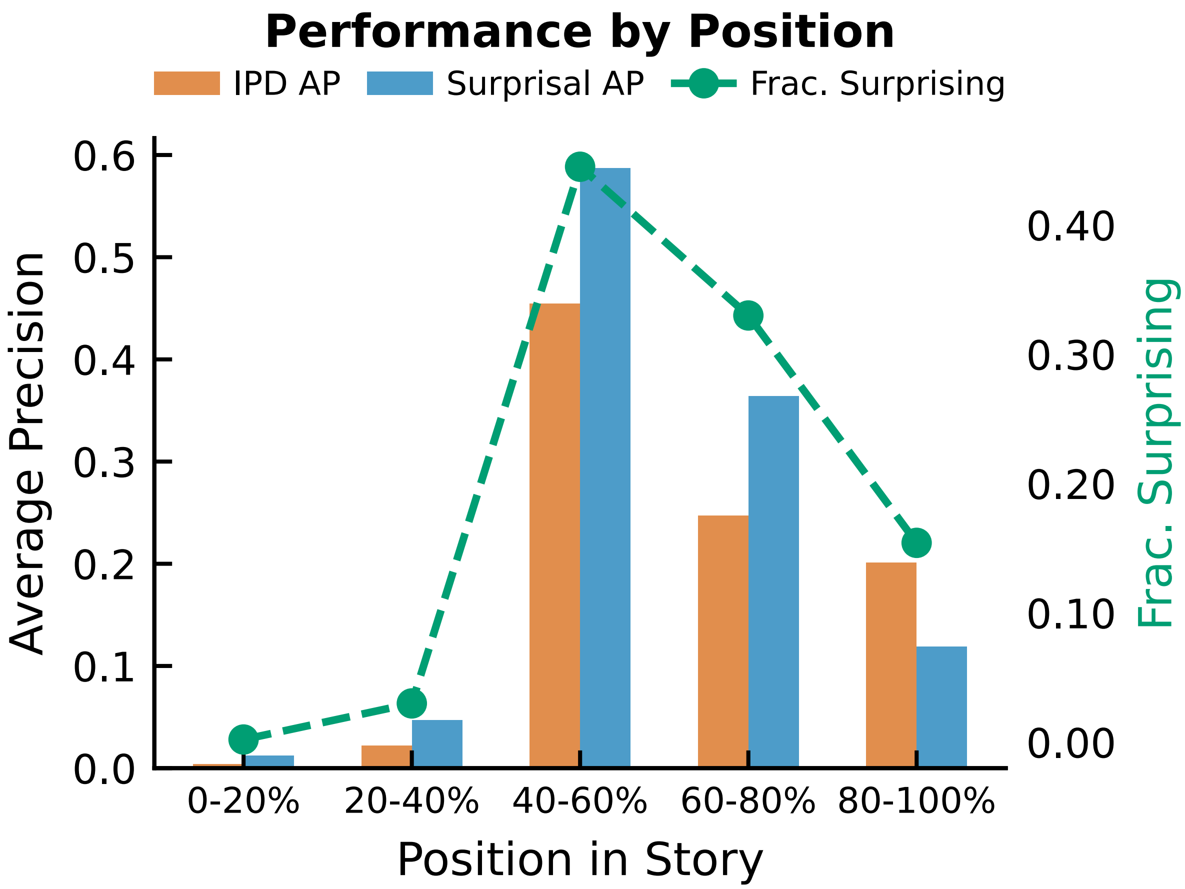

Figure 13: **Performance by sentence position.** AP for IPD and surprisal across position quintiles. IPD outperforms surprisal only in the 80–100% bin (story endings).

H Epistemic Uncertainty Heatmap

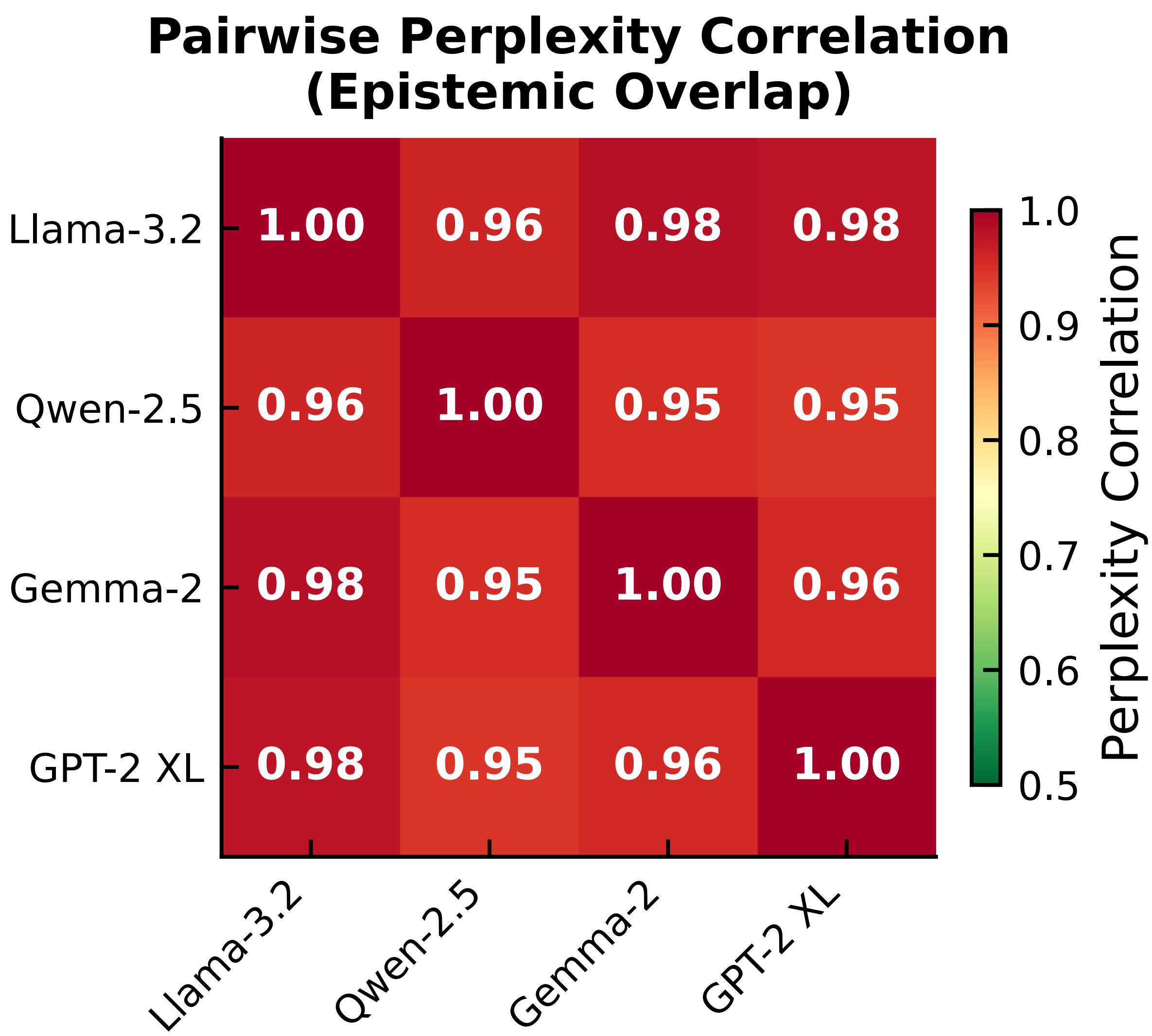

Figure 14: **Epistemic uncertainty analysis.** Heatmap of pairwise perplexity correlations among the four ensemble members. All pairs exceed $r = 0.94$.

I IPD Trajectory Examples

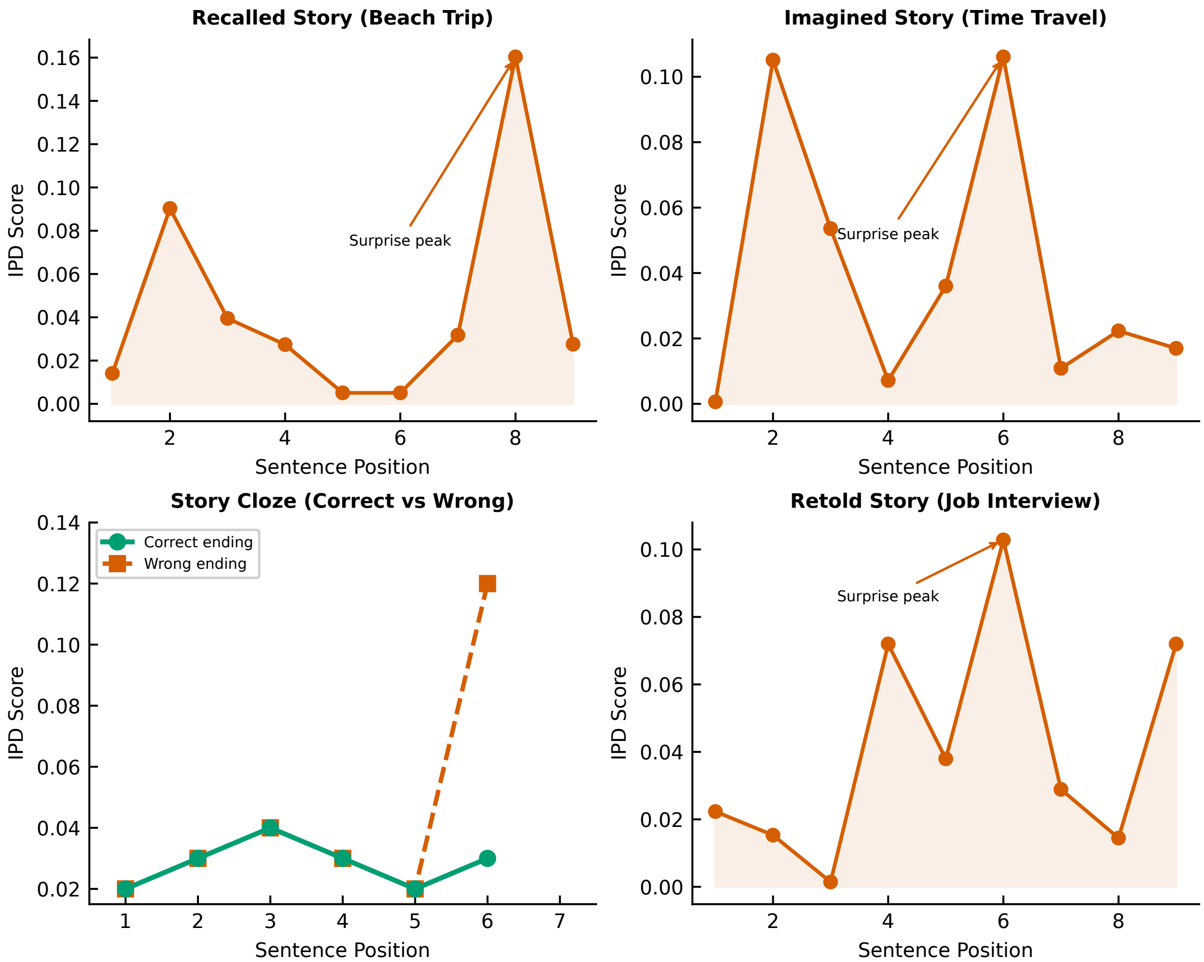

Figure 15: **Example IPD trajectories.** Sentence-level IPD scores for representative stories, showing how inter-model divergence varies across the narrative arc.

J Data Samples and Prompt Examples

Figure 16 presents annotated example stories from the Hippocorpus dataset, showing the alignment between human surprise labels and IPD scores. The following prompt template (Figure 17) is used for the LLM-as-judge baseline.

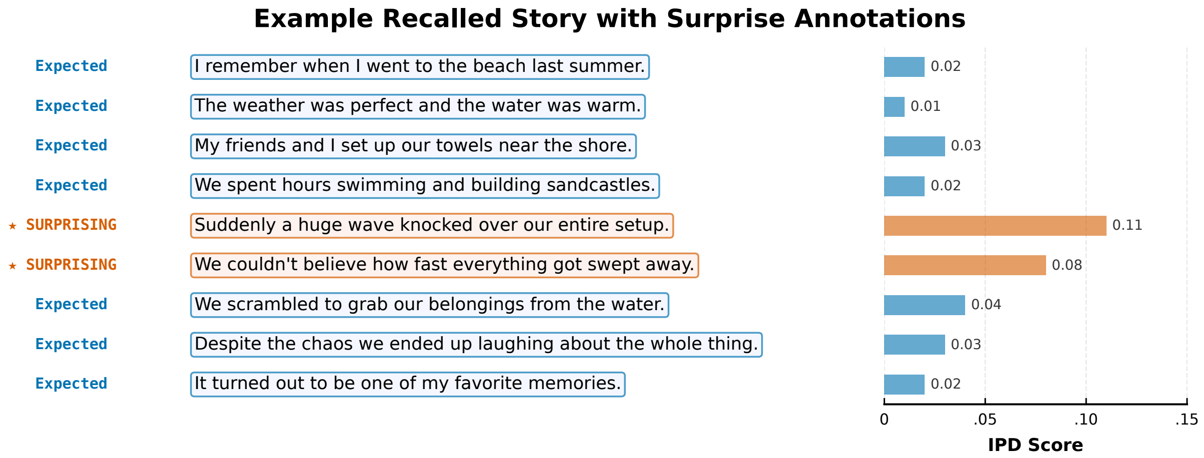

Figure 16: **Annotated data samples.** Example Hippocorpus stories with sentence-level surprise annotations (highlighted) and corresponding IPD scores.

Prompt

Given the following story so far: `[context]`. The next sentence is: `[sentence]`. Rate how surprising this sentence is on a scale of 1 (completely expected) to 5 (very surprising). Respond with only the number.

INPUT PROMPT

```
Given the following story so far:
I remember when I went to the beach last
summer. The weather was perfect and the
water was warm. My friends and I set up
our towels near the shore. We spent hours
swimming and building sandcastles.

The next sentence is:
Suddenly a huge wave knocked over our
entire setup.

Rate how surprising this sentence is on a
scale of 1 (completely expected) to 5
(very surprising). Respond with only the
number.
```

MODEL OUTPUT

4

→ Mapped to LLM-judge score: 4/5 (high surprise)

Figure 17: **LLM-as-judge prompt template.** The prompt template for eliciting sentence-level surprise ratings from an instruction-tuned LLM.

K Vocabulary Alignment Analysis

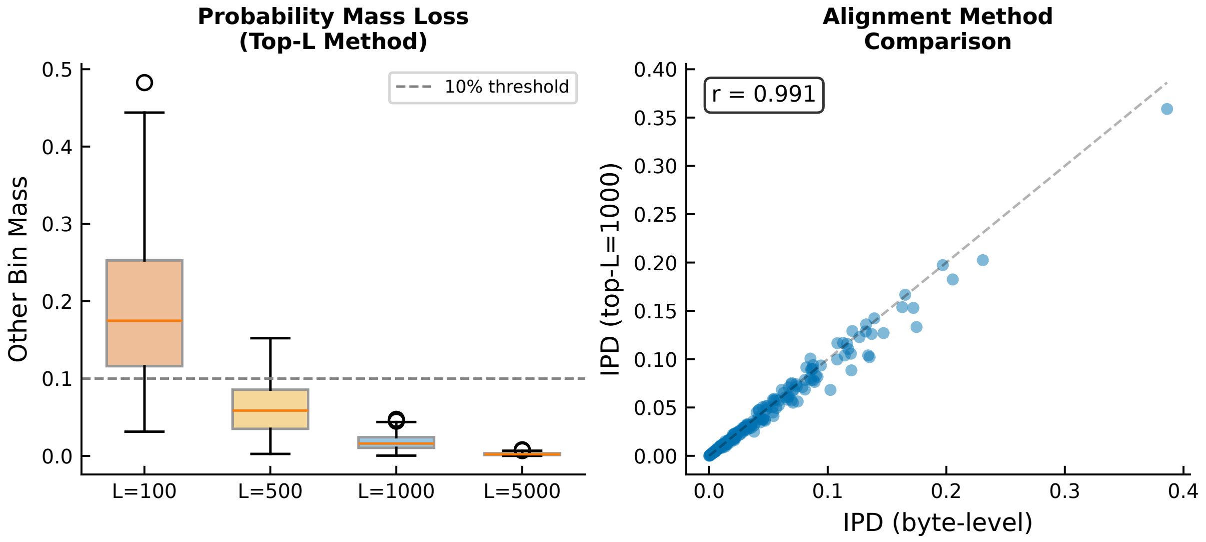

Figure 18: **Vocabulary alignment analysis.** Comparison of IPD scores across different alignment methods (byte-level vs. top-L), demonstrating consistency across approaches.

L Per-Story AP Distribution

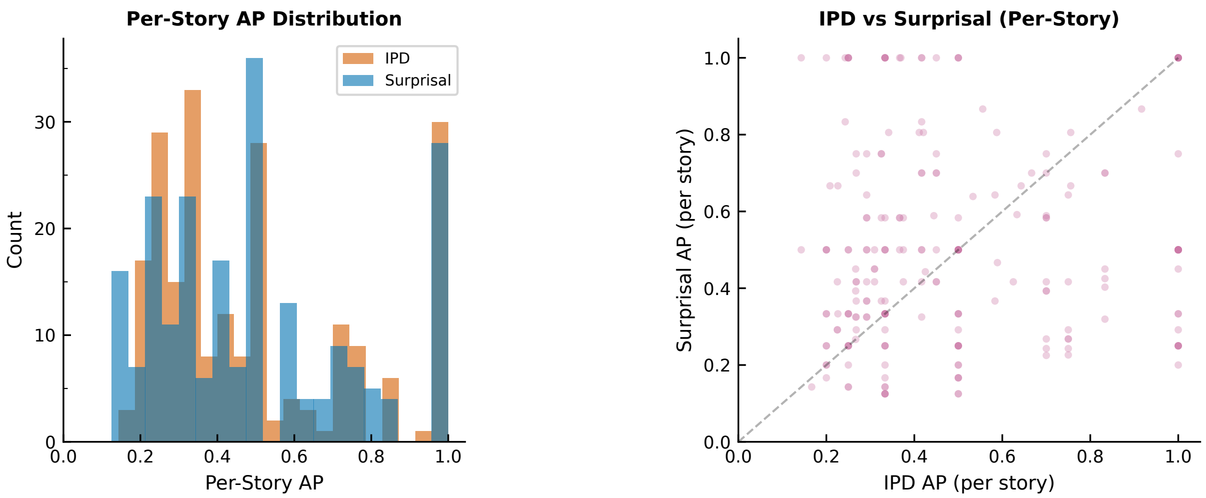

Figure 19: **Per-story AP distributions.** Comparison of IPD and ensemble surprisal AP computed individually for each story. IPD achieves mean per-story AP of 0.494 (median 0.417) and ensemble surprisal achieves mean 0.501 (median 0.483), with neither method consistently dominating the other.

M Full Per-Model Results

Method	AP	AUC	r_{pb}	p-value
IPD (Generalized JSD)	0.193	0.552	0.177	< 0.001
IPD (Pairwise JSD)	0.194	0.553	0.179	< 0.001
Surprisal (Llama-3.2-3B)	0.228	0.636	0.184	< 0.001
Surprisal (Qwen-2.5-3B)	0.246	0.651	0.200	< 0.001
Surprisal (Gemma-2-2B)	**0.250**	**0.660**	**0.207**	< 0.001
Surprisal (GPT-2 XL)	0.221	0.617	0.173	< 0.001
Ensemble Surprisal	0.235	0.643	0.194	< 0.001
Ens. Rectified Seq.	0.232	0.640	0.190	< 0.001
Ens. Sequentiality	0.180	0.509	0.036	0.118
Rect. Seq. (Llama-3.2)	0.224	0.631	0.179	< 0.001
Rect. Seq. (Qwen-2.5)	0.245	0.653	0.202	< 0.001
Rect. Seq. (Gemma-2)	0.246	0.656	0.203	< 0.001
Rect. Seq. (GPT-2 XL)	0.216	0.609	0.166	< 0.001
Contextual Entropy	0.193	0.533	0.144	< 0.001
Embedding Distance	0.219	0.514	0.119	< 0.001
VADER Reversal	0.204	0.576	0.061	< 0.01

Table 12: **Complete results on the full 240-story Hippocorpus annotated set** (1,902 sentences, 352 surprising). Includes per-model rectified sequentiality. Best per column in **bold** with green shading.

N Detailed Topic Robustness

Metric	R^2_{topic}	F-stat	Perm. p
IPD (Generalized JSD)	0.336	16.75	< 0.001
IPD (Pairwise JSD)	0.335	16.70	< 0.001
Surprisal (Llama-3.2-3B)	0.233	10.05	< 0.001
Surprisal (Qwen-2.5-3B)	0.151	5.90	< 0.001
Surprisal (Gemma-2-2B)	0.154	6.04	< 0.001
Surprisal (GPT-2 XL)	0.234	10.10	< 0.001
Ensemble Surprisal	0.181	7.32	< 0.001
Sequentiality (Llama-3.2)	0.120	4.53	< 0.001
Sequentiality (Qwen-2.5)	0.449	27.02	< 0.001
Sequentiality (Gemma-2)	0.152	5.95	< 0.001
Sequentiality (GPT-2 XL)	0.479	30.52	< 0.001
Ens. Sequentiality	0.145	5.60	< 0.001
Ens. Rectified Seq.	0.180	7.28	< 0.001
Contextual Entropy	**0.139**	5.34	< 0.001
Embedding Distance	0.241	10.51	< 0.001
VADER Reversal	0.353	18.09	< 0.001

Table 13: **Detailed topic robustness analysis.** Per-model sequentiality shows high variability across models (Qwen-2.5: 0.449, GPT-2 XL: 0.479 vs. Llama-3.2: 0.120), highlighting the cross-model inconsistency that motivates IPD.

O IPD-Surprisal Correlation

IPD is highly correlated with both ensemble surprisal (Pearson $r = 0.766$, Spearman $\rho = 0.449$) and contextual entropy ($r = 0.947$). The high Pearson but moderate Spearman correlation with surprisal indicates a strong linear relationship that is weaker in rank order, suggesting that IPD and surprisal diverge most in their treatment of extreme values. IPD is also correlated with sequentiality ($r = 0.740$). These correlations contextualize the complementarity findings: the residual IPD signal beyond surprisal ($r = 0.046$) is small in magnitude but statistically significant, and the pure disagreement component (IPD residual after removing entropy) achieves AP = 0.240.