Surprise! How Do Science Fiction Stories Challenge Predictive AI Readers?

Rafael B. Frigori[1], **Ariadne A. Costa**[2]

[1]Universidade Tecnológica Federal do Paraná – Toledo, PR – Brazil

frigori@utfpr.edu.br

[2]Grupo de Redes Complexas Aplicadas de Jataí (GRAJ)
Instituto de Ciências Exatas e Tecnológicas – Universidade Federal de Jataí (UFJ)
Jataí, GO – Brazil

ariadne.costa@ufj.edu.br

Abstract. *Science fiction repeatedly violates expectation while preserving enough structure for strange worlds and technologies to become intelligible. We convert ten public-domain stories into token-level surprisal trajectories produced by a frozen GPT-2 predictive reader and compare each Human original with a conservative GPT-5.5 AI-polished descendant and an independently AI-written same-topic control. All three conditions receive identical whitespace canonicalization before tokenization. Under this common preprocessing, Human and AI-polished texts become nearly indistinguishable: mean surprisal differs by only 0.013 nats, their 99th-percentile surprisals do not differ detectably, and their multifractal scaling functions almost coincide. By contrast, AI-written controls retain much lower mean surprisal but a heavier low-probability tail, markedly narrower singularity spectra, strong compression relative to iterative amplitude-adjusted Fourier-transform surrogate nulls, and distinct entropy–complexity geometry. The results show that conservative AI polishing can preserve both local predictability and multiscale surprisal organization, whereas independent same-topic generation changes the organization of the predictive signal itself.*

1. Introduction

Science fiction is a literature of expectancy under stress. Readers are asked to make sense of a moon built from bricks, an experiment that alters the behavior of a town, a remote future society, or an alien object whose meaning is initially obscure. Such devices are surprising without being random: a successful narrative gradually supplies enough regularity for an initially impossible event to become interpretable. Science fiction is therefore a natural domain for asking whether surprise is only a sequence of isolated local events or whether it also develops a longer-range organization through the course of a story.

Our starting point is the psycholinguistic notion of *expectancy*. The N400 is an event-related electrical potential measured at the scalp during electroencephalography; it was first associated with semantically incongruous sentence completions [Kutas and Hillyard 1980] and is now understood more broadly as a marker sensitive to

contextual expectancy and semantic integration [Kutas and Federmeier 2011]. In simple terms, words that are difficult to anticipate or integrate tend to elicit a stronger N400 response. A language model offers a computationally reproducible, although much narrower, analogue of that idea: it assigns a conditional probability to each possible continuation of a text. *Surprisal* is the negative logarithm of the probability assigned to the continuation that actually occurs, so a probable continuation has low surprisal and an improbable one has high surprisal. Surprisal predicts human reading difficulty [Hale 2001, Levy 2008, Smith and Levy 2013] and has been related to N400 amplitude [Frank et al. 2013, Michaelov et al. 2024]. Different computational predictors can nevertheless explain different electrophysiological components [Salicchi and Hsu 2025]. We therefore use GPT-2 only as a frozen *predictive AI reader*: its surprisal is one model-specific coordinate of expectancy violation, not a measurement of subjective suspense, literary quality, semantic understanding, or neural activity.

Once a story is represented by the ordered surprisal values produced while the predictive reader advances through the text, surprise becomes a time-like signal rather than a bag of isolated words. This permits a second question: do fluctuations in predictability have structure across several narrative scales? Detrended fluctuation analysis (DFA) measures scale-dependent persistence in nonstationary signals [Peng et al. 1994]; multifractal detrended fluctuation analysis (MFDFA) extends the idea by asking whether small and large fluctuations obey the same scaling law or require a spectrum of scaling exponents [Kantelhardt et al. 2002, Ihlen 2012]. Related multiscale concepts appear in critical and near-critical complex systems [Bak et al. 1987, Beggs and Plenz 2003, Hesse and Gross 2014, Plenz et al. 2021]. We use that literature only as methodological motivation: a text is not a cortex, and observing scale-dependent surprisal does not establish physical criticality. Surrogate time series provide explicit null models for deciding which apparent multifractal features survive destruction of temporal order while preserving simpler statistical properties [Schreiber and Schmitz 2000].

A complementary description places a normalized signal distribution in entropy–complexity space. We combine generalized Tsallis entropy [Tsallis 1988] with the López-Ruiz–Mancini–Calbet (LMC) entropy–disequilibrium idea [López-Ruiz et al. 1995, Rosso et al. 2007] and adapt the power-spectral-entropy logic of normalizing power before measuring complexity [Mao et al. 2018] to DFA scales. These descriptors complement rather than replace local surprisal. Related literary work tracks emotional or narrative arcs [Jockers 2013, Reagan et al. 2016, Boyd et al. 2020]; human–AI studies examine co-writing [Clark et al. 2018, Lee et al. 2022] and constrained rewriting [Li et al. 2018, Zhang and Lapata 2017]; GLTR and DetectGPT address generated-text attribution [Gehrmann et al. 2019, Mitchell et al. 2023]. Our question is instead how conservative editing versus independent generation changes the same predictive signal across scales.

We analyze ten matched triplets: a public-domain Human source, a GPT-5.5 AI-polished descendant, and an independent same-topic GPT-5.5 AI-written control. All variants are frozen before measurement and subjected to identical surface preprocessing. With ten stories and one predictive reader, the aim is not universal AI detection but a controlled comparison of local expectancy and multiscale narrative organization.

Table 1. Human originals. S1–S10 are the public story identifiers used in the figures.

ID	Title	Author	Year	Words
S1	The Unparalleled Adventure of One Hans Pfaall	Edgar Allan Poe	1835	21,488
S2	The Brick Moon	Edward Everett Hale	1869	24,720
S3	Doctor Ox's Experiment	Jules Verne	1872	19,395
S4	Master Zacharius	Jules Verne	1854	12,409
S5	A Dream of Armageddon	H. G. Wells	1901	10,187
S6	A Story of the Days to Come	H. G. Wells	1899	32,192
S7	The Time Machine	H. G. Wells	1895	32,747
S8	The Scarlet Plague	Jack London	1912	20,398
S9	The Moon Metal	Garrett P. Serviss	1900	24,590
S10	The Red One	Jack London	1918	10,120

2. Methods

2.1. Corpus, AI transformations, and common preprocessing

The corpus contains ten long-form, public-domain English science-fiction stories (Table 1), curated before analysis as a convenience rather than a probability sample. Human originals came from Project Gutenberg; Gutenberg header/footer material was removed and anthology sources were restricted to the selected story body. The public display labels S1–S10 are contiguous identifiers used throughout the paper; their Project Gutenberg eBook identifiers are, respectively, 2147, 1633, 28657, 28657, 1743, 27365, 35, 21970, 8199, and 788. The same fixed story colors identify S1–S10 in every story-level figure.

Before the analytical stage, GPT-5.5 was run once under a fixed prompt-and-seed protocol to create two variants per source; the resulting texts were frozen and the generator was never invoked during measurement. The POLISHED instruction was deliberately conservative, intended to reduce historical linguistic artifacts and improve readability without reconstructing the narrative: *"Rewrite the human original in cleaner contemporary prose while preserving the narrative order and most lexical content."* The independent SAME_TOPIC control used the complete template *"Write an independent modern science-fiction mesotext on the same broad conceptual theme as '[TITLE]'. Do not imitate the author's style. Use a neutral, regular, explanatory machine-like narrative style. Keep the recurring themes: [THEME TERMS]. Target approximately [N] words."* Human and AI-polished word counts differ by only one word on average; their mean lexical overlap is 0.9969 and mean 8-gram Jaccard similarity is 0.971. AI-written controls average 98.1% of source length but have mean lexical overlap 0.072 with their Human counterparts.

Whitespace is an experimental variable for a subword language model because line breaks and paragraph boundaries enter the tokenizer's context. We therefore apply the *same* canonicalization to all three final conditions immediately before GPT-2 tokenization: leading/trailing whitespace is stripped and every contiguous whitespace sequence (regular-expression `\s+`) is replaced by one ASCII space. GPT-2 receives the resulting plain text without lemmatization, stop-word removal, semantic annotation, or manual sentence-level rewriting; no lexical or punctuation content is changed by canonicalization. This prevents historical hard wrapping or generated paragraph layout from acting as an unintended predictor.

Table 2. Most locally altered aligned Human–AI-polished window (8/80 changed words); the displayed excerpt is shortened and changed pronouns are emphasized.

Condition	Excerpt
Human	"No, **thou** wouldst not kill **thy** father, and **thou** wilt accept this man for **thy** husband [...] Gerande, since I gave **thee** life, give life to **thy** father!"
AI-polished	"No, **you** wouldst not kill **your** father, and **you** wilt accept this man for **your** husband [...] Gerande, since I gave **you** life, give life to **your** father!"
AI-written	"Because technical mastery becomes metaphysical anxiety when every mechanism starts to fail, the committee treated each anomaly as evidence and each witness as an instrument."

Table 2 displays the maximum-alteration Human–AI-polished window found by deterministic alignment of all ten pairs: *Master Zacharius*, with 8 of 80 Human words involved in an insertion, deletion, or substitution. It is illustrative only; the AI-written row is same-topic, not aligned.

2.2. Predictive reader and local descriptors

GPT-2 is an autoregressive language model: from the tokens already observed it assigns probabilities to possible next tokens [Radford et al. 2019]. Let w_t be the GPT-2 subword token at position t and $w_{<t}$ the preceding tokens available as context. Token surprisal is

$$s_t = -\ln p_{\mathrm{GPT2}}(w_t \mid w_{<t}), \tag{1}$$

where p_{GPT2} is the conditional probability assigned to the token that actually occurs. The logarithm is natural, so s_t is measured in nats: unlikely tokens have large s_t, expected tokens small s_t. The ordered values $s_1, \ldots, s_T$, for T scored tokens, form the surprisal trajectory.

The reader is implemented in Python/PyTorch with Hugging Face Transformers, using `AutoTokenizer` and `AutoModelForCausalLM` with `model_name="gpt2"`, i.e. the standard GPT-2 small checkpoint (approximately 124M parameters), and tokenization uses `add_special_tokens=False`. Long texts are evaluated in 1024-token windows with stride 512: later windows reuse earlier tokens as left context but contribute only their newest 512 targets, so overlap does not double-count observations. The first token has no causal left context; it is excluded from GLTR ranking and assigned the within-stream median surprisal as a bookkeeping placeholder, an $O(1/T)$ contribution. All analyses remain at GPT-2 subword-token resolution. The same logits give Giant Language Model Test Room (GLTR) rank diagnostics [Gehrmann et al. 2019]; `topk=1000` provides ranks 1–10, 11–100, 101–1000, and the complementary tail above 1000. Mean and 99th-percentile (p99) surprisal summarize typical and extreme local unpredictability.

2.3. DFA, MFDFA, and surrogate nulls

Detrended fluctuation analysis (DFA) measures scale-dependent persistence in a nonstationary signal [Peng et al. 1994]. We center and integrate surprisal,

$$Y(k) = \sum_{t=1}^{k} (s_t - \bar{s}), \qquad \bar{s} = T^{-1} \sum_{t=1}^{T} s_t, \tag{2}$$

where $Y(k)$ is the cumulative profile after token k and $\bar{s}$ is the story mean. At each box length ℓ, a straight trend is removed inside each segment and the root-mean-square residual gives $F(\ell)$. A relation $F(\ell) \propto \ell^{\alpha_{\mathrm{DFA}}}$ defines the DFA exponent α_{DFA}; values above or below the uncorrelated reference near $1/2$ indicate persistent or anti-persistent increments. DFA uses 22 logarithmic scales from 8 to $\min(T/4, 2048)$ tokens, retaining scales with at least four segments.

Multifractal detrended fluctuation analysis (MFDFA) repeats this construction after weighting segment fluctuations by a moment order q [Kantelhardt et al. 2002, Ihlen 2012]. Positive q emphasizes large fluctuations and negative q emphasizes small ones. The generalized fluctuation function obeys $F_q(\ell) \propto \ell^{h(q)}$, defining the generalized Hurst exponent $h(q)$. We use 251 moment orders over $q \in [-15, 20]$ and 80 logarithmic scales from 10 to approximately $T/4$. From $h(q)$ we compute the mass exponent $\tau(q) = qh(q) - 1$ and its Legendre representation

$$\alpha(q) = \frac{d\tau}{dq}, \qquad f(\alpha) = q\alpha - \tau(q), \tag{3}$$

where α is the local scaling-strength coordinate and $f(\alpha)$ is the associated singularity spectrum. Its width $\Delta\alpha = \alpha_{\max} - \alpha_{\min}$ measures the range of scaling strengths; α_{peak} is the most represented strength and unequal left/right widths quantify asymmetry. Numerically, quadratic Savitzky–Golay smoothing uses a scale-wise window of at most 7 points, then a $\sqrt{\ell}$-weighted log–log fit for $h(q)$; the $h(q)$ curve is smoothed with a window of at most 21 points (polynomial order 2, `mirror` boundary mode), and $\alpha(q)$ is obtained by cubic-spline differentiation of $\tau(q)$.

Each text is also compared with 99 shuffled and 99 iterative amplitude-adjusted Fourier-transform (IAAFT) surrogates [Schreiber and Schmitz 2000], generated with seed 20260518. Shuffling preserves the observed value distribution but destroys order. IAAFT additionally preserves the amplitude distribution and closely matches the linear Fourier power spectrum while randomizing phase organization; 60 iterations are used per surrogate. For any scalar descriptor X, its surrogate-normalized deviation is $z_X = (X_{\mathrm{obs}} - \langle X_{\mathrm{IAAFT}} \rangle)/\sigma_{X,\mathrm{IAAFT}}$. We further define

$$q_{\mathrm{eff}} = 1 + \left[\Delta\alpha_{\mathrm{obs}} - \langle \Delta\alpha_{\mathrm{IAAFT}} \rangle\right], \tag{4}$$

where the bracketed term is the observed width minus the mean IAAFT width. Thus $q_{\mathrm{eff}} = 1$ means no width excess relative to this null. It is an empirical coordinate of this analysis, not the MFDFA moment order q and not a universal thermodynamic entropic index.

2.4. PSDFA, entropy–complexity geometry, and thermodynamic embedding

Power-spectral entropy normalizes Fourier power into probabilities. Our *Power Spectral DFA* (PSDFA) applies the same idea to DFA scales: if $F(\ell)$ is the detrended fluctuation amplitude at scale ℓ, then

$$P_\ell = \frac{F(\ell)^2}{\sum_u F(u)^2}, \tag{5}$$

so the nonnegative P_ℓ values sum to one. We evaluate 20 logarithmic scales from 10 to approximately $T/4$. PSDFA is therefore a probability distribution over detrended

fluctuation scales, not a Fourier spectrum; $F(\ell)^2$ plays the role that spectral power plays in power-spectral entropy [Mao et al. 2018].

For $P = \{P_\ell\}$ we compute Tsallis entropy $S_q = (1 - \sum_\ell P_\ell^q)/(q - 1)$, normalize it by its uniform-distribution maximum to obtain H_q, and multiply by normalized Jensen–Tsallis disequilibrium D_q to obtain $C_q = H_q D_q$. This is the LMC entropy–disequilibrium idea expressed with generalized entropy [Tsallis 1988, López-Ruiz et al. 1995, Rosso et al. 2007]. Here q is an entropy index, not the MFDFA moment order. An 801-point linear grid over $0 \le q \le 15$, with the Shannon-limit point $q = 1$ inserted explicitly, traces the (H_q, C_q) curve. Its *sickle index* is $(A_{\text{hull}} - A_{\text{curve}})/A_{\text{hull}}$, where A_{hull} is the convex-hull area and A_{curve} the area enclosed by the ordered trajectory; larger values denote a more strongly re-entrant curve.

The thermodynamic visualization combines q_{eff} with IAAFT-normalized spectral asymmetry z_{asym}, where $A = (\alpha_{\max} - \alpha_{\text{peak}})/(\alpha_{\text{peak}} - \alpha_{\min})$ is standardized against the 99 IAAFT surrogates. In leave-one-story-out calibration, the held-out triplet never defines its own reference state: the other nine stories determine class centers and shared dispersions. Equal-prior Student-t likelihoods with 98 degrees of freedom map the two coordinates to three nonnegative barycentric weights. These are descriptive embedding weights, not universal authorship probabilities.

2.5. Story-level inference and validation

The independent experimental unit is the *story*, not the token. Scalar contrasts use one paired difference per story and all $2^{10} = 1024$ exact sign reversals. We report the paired standardized effect $d_z = \bar{d}/s_d$, where $\bar{d}$ is the mean of the ten paired differences and s_d their sample standard deviation. For the complete MFDFA function $h(q)$, the same sign-flip scheme is applied to max-T, the largest absolute standardized difference over the tested q grid, and L_2, the root sum of squared mean-function differences. We examine the central sector $q \in [-5, 5]$, the extreme-negative sector $q \in [-15, -9]$, and the full range. Because non-significance does not establish equality, δ_{95} denotes the smallest symmetric band containing a 95% simultaneous paired-bootstrap band over the central Human–AI-polished mean difference; it is a sensitivity bound, not a formal equivalence margin.

Classification is secondary. We use scikit-learn `LeaveOneGroupOut`, withholding all three variants of one story together. The four reported feature blocks contain no missing values; `StandardScaler` and `LogisticRegression` are fitted inside each training fold with $C = 1.0$, `max_iter=5000`, and `class_weight=balanced`. Balanced accuracy is mean recall across classes. The blocks are GPT-2/GLTR descriptors, no-length stylometry, global DFA/MFDFA/q-LMC descriptors, and their no-length combination. Numerical post-processing uses NumPy, pandas, SciPy, and scikit-learn; Matplotlib is used for visualization. Some descriptors were computed with proprietary research software whose source is not publicly redistributed, but the scientific definitions and parameters used here are stated above.

Table 3. Key averages after identical whitespace canonicalization. GLTR tail is the fraction ranked beyond 1000 under GPT-2.

Condition	Mean s_t	p99 s_t	GLTR tail	$\Delta\alpha$	q-LMC sickle	q_{eff}
Human	3.533	11.71	0.0378	0.469	0.271	0.994
AI-polished	3.546	11.73	0.0381	0.469	0.267	0.993
AI-written	1.631	12.47	0.0711	0.243	0.140	0.736

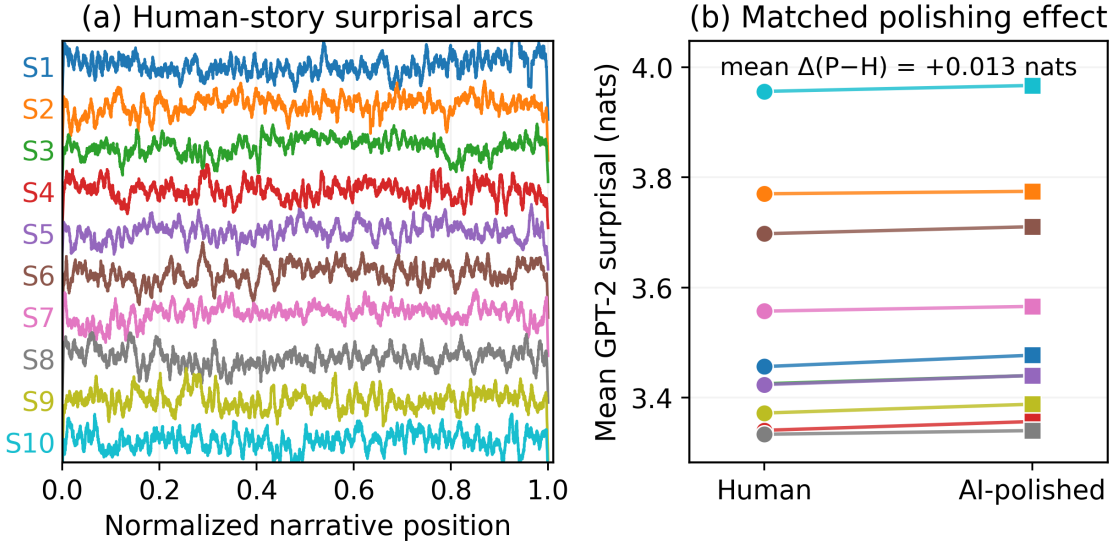


Figure 1. Narrative surprise. (a) Smoothed Human arcs (S1–S10). (b) Matched Human → AI-polished mean surprisal. Story colors are fixed across panels; circles and squares mark condition.

3. Results

3.1. After common preprocessing, conservative polishing barely changes local surprisal

Table 3 summarizes the central descriptors. Human and AI-polished texts have almost identical local statistics: mean surprisal is 3.533 versus 3.546 nats, p99 is 11.71 versus 11.73 nats, and the GLTR tail is 0.0378 versus 0.0381. The AI-written controls are different: their mean surprisal is much lower (1.631 nats) but their p99 and GLTR tail are higher, showing a highly predictable bulk punctuated by a heavier low-probability tail.

The small mean-surprisal displacement is systematic but tiny: AI-polished minus Human is $+0.0128$ nats in all ten stories (exact $p = .00195$, $d_z = 2.53$), only about 0.36% of the Human mean. The p99 difference is $+0.0177$ nats and is not detected by the exact paired test ($p = .154$). The GLTR-tail increase is only 3.53×10^{-4} in absolute fraction ($p = .00586$). By contrast, Human minus AI-written mean surprisal is 1.903 nats in all ten pairs ($p = .00195$). Figure 1 makes the scale of the Human–AI-polished change visible: the matched lines are nearly horizontal.

3.2. Human and AI-polished multiscale organization is nearly coincident

Figure 2 compares singularity-spectrum summaries and the complete mean MFDFA scaling functions. Human and AI-polished curves are nearly coincident over the full displayed

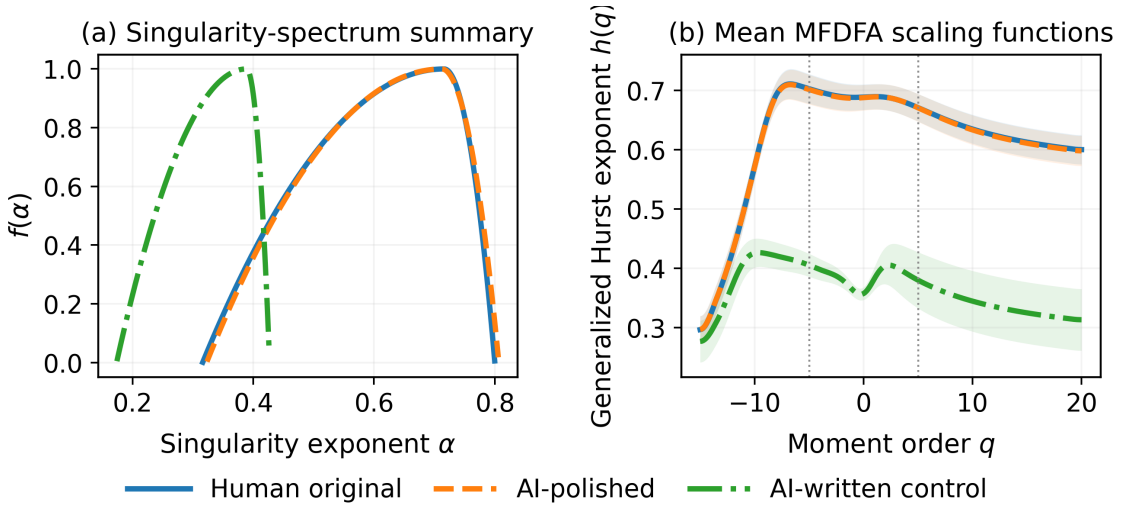


Figure 2. Multifractal organization after common preprocessing. (a) Condition-level singularity-spectrum summaries. (b) Mean MFDFA $h(q)$; bands are standard errors and dotted lines delimit $q \in [-5, 5]$. Human and AI-polished overlap closely; AI-written controls are displaced.

range, whereas the AI-written curve is displaced. Their mean spectrum widths are 0.46878 and 0.46883 ($p = .982$); AI-written controls contract to 0.24333.

The functional tests give no Human–AI-polished displacement in the central sector ($p_{\max T} = p_{L_2} = .686$), in the extreme-negative sector ($p_{\max T} = .748$, $p_{L_2} = .754$), or over the full range ($p_{\max T} = .898$, $p_{L_2} = .654$). In the central sector, the root-mean-square Human–AI-polished mean-function distance is only 0.00105. The AI-polished–AI-written distance is 0.30205, approximately 287 times larger, with both global functional tests giving $p = .00195$. The 95% simultaneous bootstrap band for the central Human–AI-polished difference fits inside $|\Delta h(q)| \leq \delta_{95} = 0.00542$. This bound quantifies how close the curves are without converting a non-significant test into a formal equivalence claim.

3.3. Surrogate and entropy–complexity analyses isolate the AI-written regime

IAAFT surrogates sharpen this distinction. Mean observed spectrum widths are 0.469, 0.469, and 0.243 for Human, AI-polished, and AI-written texts. Human and AI-polished remain close to their own amplitude- and spectrum-matched IAAFT nulls: mean width z-scores are -0.178 and -0.209 ($p = .678$). AI-written controls are strongly compressed relative to their null expectation, with mean $z_{\Delta\alpha} = -4.794$. The AI-polished–AI-written contrast is positive in all ten stories ($p = .00195$, $d_z = 3.65$).

The same result appears in q_{eff}: Human and AI-polished means are 0.9940 and 0.9935 ($p = .848$), whereas AI-written controls average 0.7360; every AI-polished value exceeds its AI-written counterpart ($p = .00195$, $d_z = 3.31$). PSDFA q-LMC geometry agrees. Mean sickle indices are 0.2706, 0.2674, and 0.1399; Human–AI-polished is not detected ($p = .219$), while AI-polished exceeds AI-written in 9/10 pairs ($p = .00391$).

3.4. A two-coordinate embedding and classifiers recover the same asymmetry

The thermodynamic embedding combines q_{eff} and z_{asym} using leave-one-story-out reference states (Figure 4). The Human and AI-polished centroids are almost coincident at approximately $(0.538, 0.003)$ and $(0.532, 0.003)$ in barycentric Cartesian coordinates, a

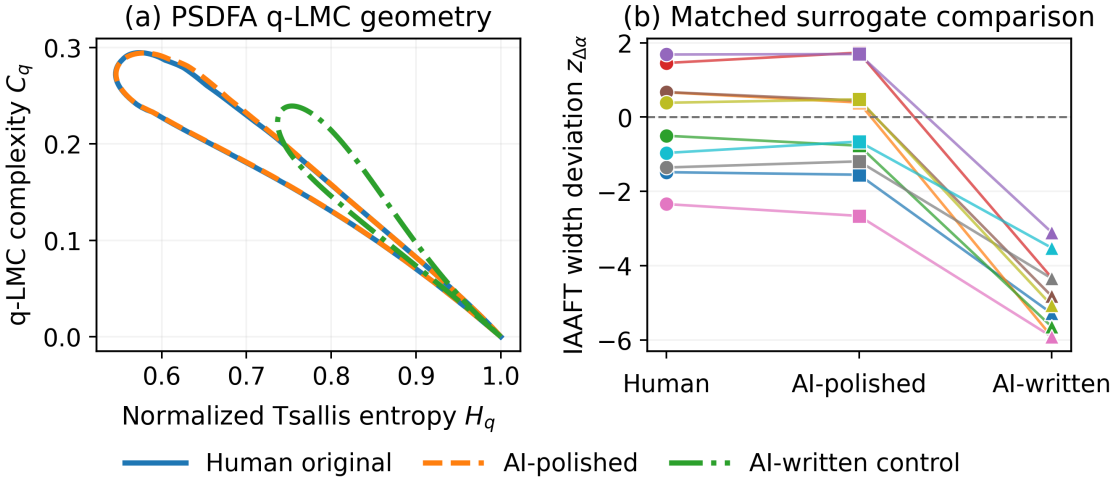


Figure 3. Complementary global structure. (a) PSDFA q-LMC trajectories. (b) Story-level MFDFA-width deviations from the IAAFT null. Condition styles match Figure 2; story colors and marker shapes follow Figure 1.

Table 4. Balanced accuracy under leave-one-story-out validation. Explicit length features are excluded from stylometry and the combined block.

Feature block	3-way	Human/Polished	Human/AI-written
GPT-2/GLTR	0.70	0.65	1.00
Stylometry, no length	0.77	0.70	1.00
Global DFA/MFDFA/q-LMC	0.70	0.50	1.00
Combined, no length	0.70	0.60	1.00

separation of only 0.0063. The AI-written centroid lies near $(0.499, 0.842)$, about 0.84 units away. Panel (b) therefore magnifies the base region so the small Human–AI-polished displacements remain visible. The embedding is descriptive; its principal result is geometric overlap of the source and conservative polish versus strong displacement of independent generation.

Classification is a secondary consistency check rather than the scientific endpoint. Table 4 shows that, after identical preprocessing, Human–AI-polished discrimination is weak: balanced accuracy ranges from 0.50 to 0.70. All four transparent blocks separate Human from the present AI-written control under leave-one-story-out validation, but this perfect separation is conditional on the deliberately constrained SAME_TOPIC prompt, which explicitly requested a neutral, regular, explanatory machine-like style; it must not be read as general AI-text detection performance. The three-way task is limited mainly by Human/AI-polished confusion. With only ten story groups, these values are corpus diagnostics.

4. Discussion

The experiment reveals a sharper separation than a simple "AI makes text predictable" account. After all three conditions are presented to GPT-2 with the same whitespace representation, conservative AI polishing preserves the source trajectory to a remarkable degree. Its mean-surprisal change is only about 0.36%, its p99 is statistically indistinguishable from the Human source, its spectrum width and q_{eff} are essentially unchanged, and its

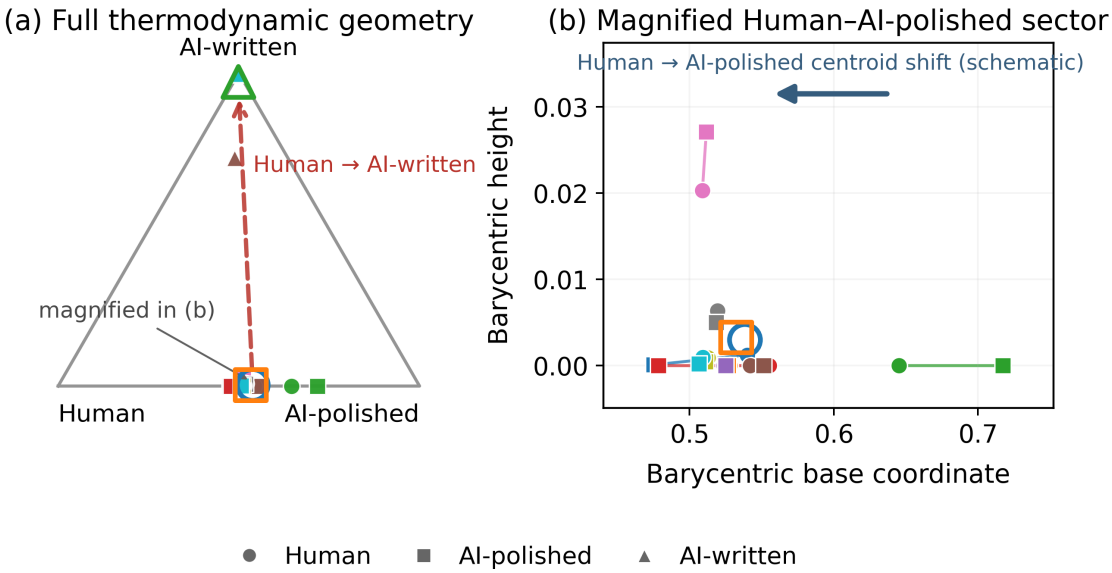


Figure 4. Thermodynamic multifractal embedding of the 30 texts. Fixed story colors and condition markers follow earlier figures; open symbols are true centroids. (a) Full geometry. (b) Magnified Human–AI-polished sector; the top Human→AI-polished arrow is schematic.

MFDFA function differs from the Human mean by only about 10^{-3} in root-mean-square distance over the central range. The same conclusion holds in the negative-q sector that preferentially weights small fluctuations.

Independent same-topic generation is different in kind rather than merely degree. Its lower mean surprisal coexists with larger p99 and GLTR tail, so low average surprise is not uniform predictability. Its singularity spectrum is about half as wide as the Human/AI-polished spectra, its IAAFT-normalized width is strongly negative, and PSDFA q-LMC plus the thermodynamic embedding place it in a separate regime. Because IAAFT preserves the amplitude distribution and approximately the linear power spectrum, the residual width compression points to temporal organization beyond those constraints.

The common whitespace step is methodologically important: a subword model treats boundaries as context, so historical wrapping or generated paragraph layout can masquerade as predictability when conditions are preprocessed differently. Removing this nuisance variable before token-level measurement makes formatting normalization part of the experimental specification rather than an incidental implementation detail.

Limitations remain: only ten English science-fiction sources and one GPT-2 reader are studied. The AI-written control is one frozen GPT-5.5 prompt-and-seed protocol, not contemporary machine writing in general; temperature, top-p, and model snapshot were not varied as robustness factors. GPT-2 surprisal is neither subjective suspense nor neural activity, and the embedding is not a universal authorship detector. Reader-response experiments and additional predictive models are needed to test generalization.

5. Conclusion

Science-fiction narratives can be studied as evolving predictability signals rather than detector scores. Under identical preprocessing, conservative GPT-5.5 polishing leaves

Human local surprisal and multiscale geometry nearly unchanged; independent same-topic generation instead displaces MFDFA, IAAFT-normalized width, PSDFA q-LMC, and the thermodynamic embedding. The robust distinction is source-preserving editing versus independent generation, a quantitative starting point for studying machine intervention in narrative expectancy across scales.

References

Bak, P., Tang, C., and Wiesenfeld, K. (1987). Self-organized criticality: An explanation of the 1/f noise. *Physical Review Letters*, 59(4):381–384.

Beggs, J. M. and Plenz, D. (2003). Neuronal avalanches in neocortical circuits. *Journal of Neuroscience*, 23(35):11167–11177.

Boyd, R. L., Blackburn, K. G., and Pennebaker, J. W. (2020). The narrative arc: Revealing core narrative structures through text analysis. *Science Advances*, 6(32):eaba2196.

Clark, E., Ross, A. S., Tan, C., Ji, Y., and Smith, N. A. (2018). Creative writing with a machine in the loop: Case studies on slogans and stories. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, pages 329–340.

Frank, S. L., Otten, L. J., Galli, G., and Vigliocco, G. (2013). Word surprisal predicts N400 amplitude during reading. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 878–883.

Gehrmann, S., Strobelt, H., and Rush, A. M. (2019). GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*. doi:10.3115/1073336.1073357.

Hesse, J. and Gross, T. (2014). Self-organized criticality as a fundamental property of neural systems. *Frontiers in Systems Neuroscience*, 8:166.

Ihlen, E. A. F. (2012). Introduction to multifractal detrended fluctuation analysis in Matlab. *Frontiers in Physiology*, 3:141.

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

Kantelhardt, J. W., Zschiegner, S. A., Koscielny-Bunde, E., Havlin, S., Bunde, A., and Stanley, H. E. (2002). Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A*, 316(1–4):87–114.

Kutas, M. and Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, 62:621–647.

Kutas, M. and Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205.

Lee, M., Liang, P., and Yang, Q. (2022). CoAuthor: Designing a human–AI collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, Article 388, pages 1–19.

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

Li, J., Jia, R., He, H., and Liang, P. (2018). Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *Proceedings of NAACL-HLT, Volume 1*, pages 1865–1874.

López-Ruiz, R., Mancini, H. L., and Calbet, X. (1995). A statistical measure of complexity. *Physics Letters A*, 209(5–6):321–326.

Mao, X., Shang, P., Wang, J., and Ma, Y. (2018). Characterizing time series by extended complexity–entropy curves based on Tsallis, Rényi, and power spectral entropy. *Chaos*, 28(11):113106. doi:10.1063/1.5038758.

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., and Coulson, S. (2024). Strong prediction: Language model surprisal explains multiple N400 effects. *Neurobiology of Language*, 5(1):107–135.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., and Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR*, pages 24950–24962.

Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., and Goldberger, A. L. (1994). Mosaic organization of DNA nucleotides. *Physical Review E*, 49(2):1685–1689.

Plenz, D., Ribeiro, T. L., Miller, S. R., Kells, P. A., Vakili, A., and Capek, E. L. (2021). Self-organized criticality in the brain. *Frontiers in Physics*, 9:639389.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. Technical report, OpenAI.

Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., and Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):31.

Rosso, O. A., Larrondo, H. A., Martin, M. T., Plastino, A., and Fuentes, M. A. (2007). Distinguishing noise from chaos. *Physical Review Letters*, 99(15):154102.

Salicchi, L. and Hsu, Y.-Y. (2025). Not every metric is equal: Cognitive models for predicting N400 and P600 components during reading comprehension. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3648–3654.

Schreiber, T. and Schmitz, A. (2000). Surrogate time series. *Physica D*, 142(3–4):346–382.

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

Tsallis, C. (1988). Possible generalization of Boltzmann–Gibbs statistics. *Journal of Statistical Physics*, 52(1–2):479–487.

Zhang, X. and Lapata, M. (2017). Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594.