The Burstiness Engine:
Modeling Human Rhythmic Patterns in Language Generation

Vittoria Lanzo Dico Angelo

June 21, 2026

Abstract

Burstiness, the variation in sentence length and complexity that characterizes human writing, is the most reliable signal that text was produced by a language model, yet it is defined only by a detector heuristic and controlled only by post-hoc rewriting. We treat burstiness instead as a property that can be specified as a distributional target and steered during generation. We make four contributions. First, we give burstiness a reproducible, decomposable definition (variance and kurtosis of sentence length, surprisal fluctuation under a fixed reference model, and punctuation entropy) that replaces the prevailing single-scalar heuristic. Second, we instrument the prompt-control baseline against this target and, through an ablation, identify which surprisal statistic actually separates human from machine rhythm: the prevailing stdev-of-surprisal measure is the weakest, while local jumpiness and mean surprisal separate cleanly. Third, we locate where in-generation control fails and where it succeeds: a single activation steering vector and a best-of-N LoRA, both operating at the token level, do not move the target beyond its noise floor (which we trace to the variance of the metric on short generations), whereas steering the sentence-boundary decision with the metric as a running discriminator moves the target while preserving coherence. Control altitude, not model scale, is the deciding factor. Fourth, we present a pre-registered human perception protocol, adapted from a method validated in speech synthesis et al. (2025a), to test whether controlled burstiness changes how readers judge the humanness of text. The novelty is the composition: applying established steering machinery to a rhythm-specific target, with a reusable definition and a perception protocol. We frame the work as modeling human rhythmic patterns, not as evading detection, and we report detector-relevant behavior across the control range so the contribution serves measurement and model analysis as well as generation.

1 Introduction

1.1 The flatness tell

Human writing has rhythm. Sentences cluster and scatter: a long clause followed by a short one, a dense paragraph then a terse line. Machine-generated text tends toward a flat middle, sentences of similar length and similar complexity, paragraph after paragraph. This flatness, captured loosely by the term burstiness, is the single most reliable signal that a passage was produced by a language model. It is what readers sense as off, what detectors key on, and what a growing market of humanizer tools tries to remove after the fact.

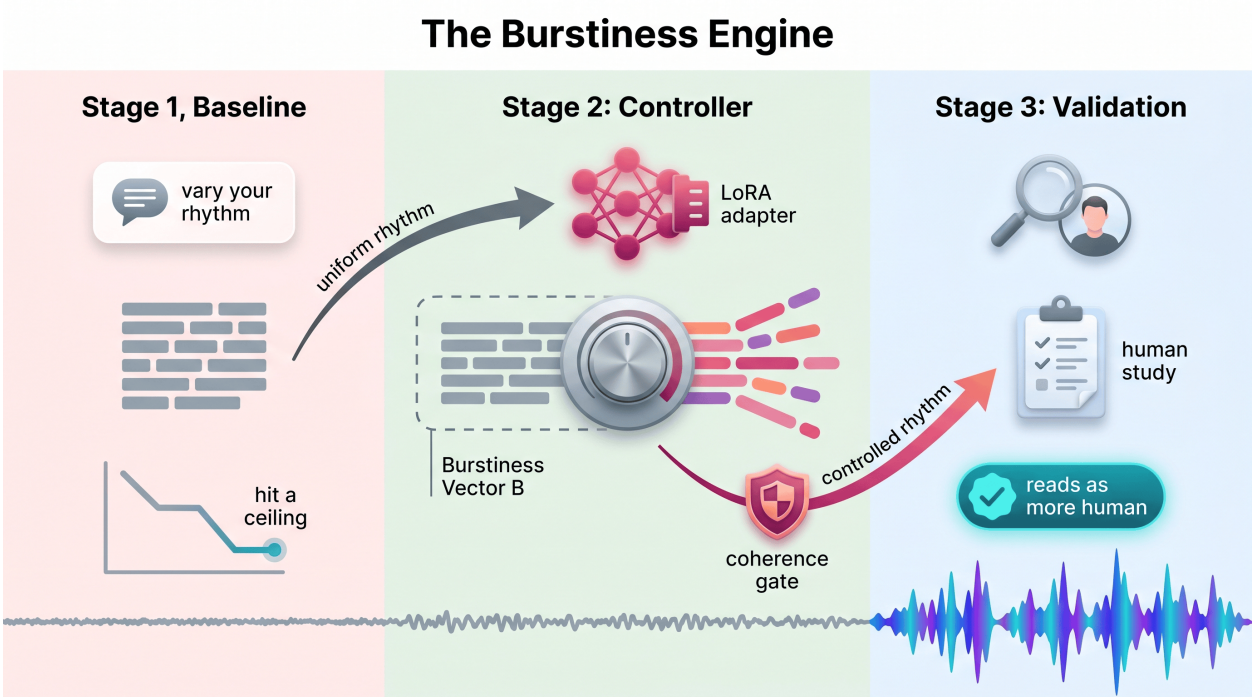

Figure 1: The Burstiness Engine. Prompt-only control hits a ceiling (left); a model-level controller modulates the burstiness vector B in-generation behind a coherence gate (center); a perception protocol tests the shift (right).

1.2 A solved-everywhere, formalized-nowhere problem

Demand to make machine text read like a person is obvious and broad. Detection systems treat burstiness as a core feature. Commercial humanizers rewrite model output to reintroduce variation. Practitioners trade prompt tricks for rhythm. Yet two things are missing. First, the field has no reproducible definition of burstiness: the operational definition in common use traces to a detector technical note, not a peer-reviewed formalization. Second, every working solution acts post-hoc, rewriting finished text rather than controlling rhythm during generation. The science of controlling burstiness in-generation does not exist.

1.3 Why in-generation control, and why model-level

Controlling a property after generation is a patch; controlling it during generation is a mechanism. Prompt-level instruction is the natural first attempt, and it is a useful baseline: it is fast, needs no infrastructure, and works on any model. But instruction asks a model to perform rhythm without giving us a handle on it. We expect it to have a low ceiling, to degrade over long outputs, and to be hard to measure against a precise target. Recent evidence that prompting cannot reliably reproduce a person's implicit writing style points the same way. The alternative is a model-level controller, a small auxiliary mechanism that modulates rhythm separately from content, giving a parameter we can ablate, tune, and measure. The paper follows this spine: establish the prompt baseline, show where it breaks, and present the model-level controller as the answer.

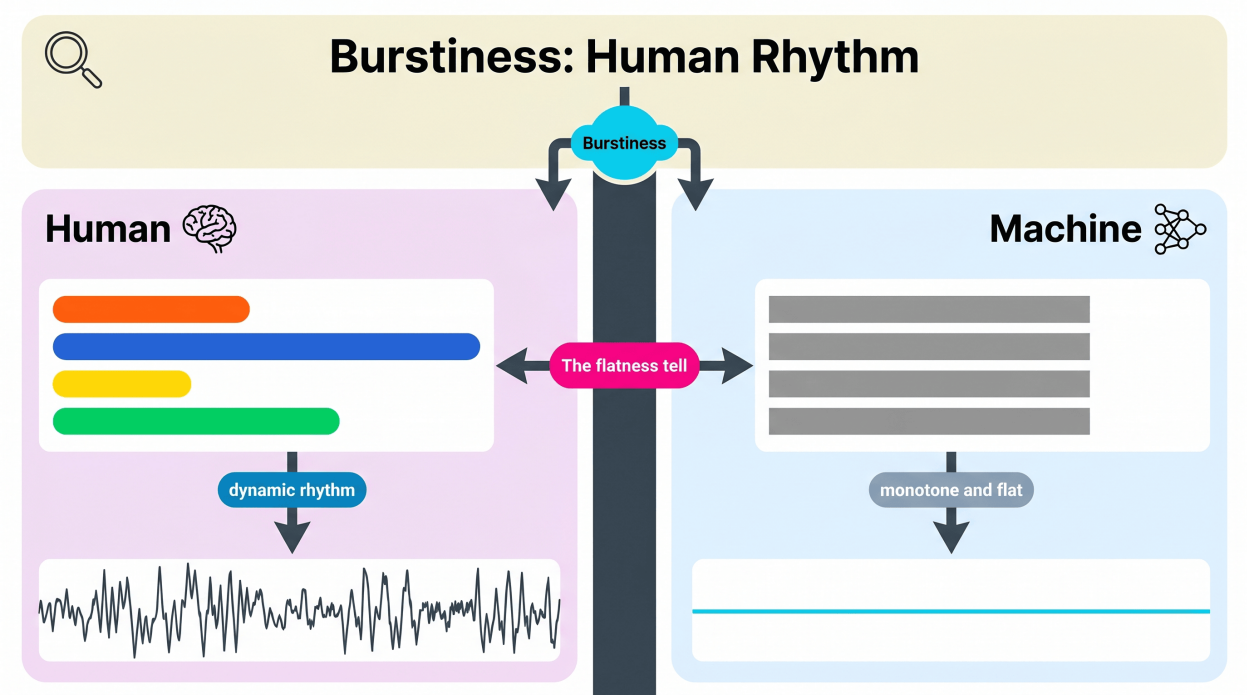

Figure 2: The flatness tell. Human text varies sentence length and rhythm (left); machine text is uniform and flat (right).

1.4 Contributions

1. A formal, reproducible definition of burstiness as a distributional generation target (variance and kurtosis of sentence length, surprisal fluctuation, punctuation entropy), replacing the prevailing single-scalar detector heuristic (Section 3).
2. A measurement of the prompt-control ceiling: how well instruction alone hits a distributional rhythm target, and how that degrades with output length (Section 4). This is a clean negative result and stands on its own.
3. A model-level burstiness controller that steers rhythm in-generation by biasing the sentence-boundary decision, yielding a monotonic, coherence-preserving sentence-length dial on distil-gpt2, with ablations and a path to a personal rhythmic fingerprint (Section 5).
4. A pre-registered human perception protocol to test whether controlled burstiness shifts how readers judge the humanness of text, adapting a method validated in speech synthesis (Section 6).

1.5 Scope and stance

We model human rhythmic patterns; we do not build a tool whose purpose is to evade detection. The distinction matters for venue and ethics, and we return to it in Section 7. Prosody control in speech synthesis supplies our perception method and an in-generation precedent, but voice personalization is future work, not a core claim. Emotional-state control is out of scope. The novelty of this work is the composition, applying established steering machinery to a new target, rhythm, with a reproducible definition and perceptual validation, rather than any single new mechanism.

2 Related Work

2.1 The shape of the field: two camps, an empty bridge

Work touching burstiness splits into two camps that rarely cite each other. One camp measures and detects burstiness to catch machine text. The other builds steering machinery to push frozen LLMs toward style targets. Neither camp controls rhythm in-generation, and neither gives burstiness a reproducible definition. That gap between measurement and control is the contribution space of this paper.

2.2 Burstiness in detection (measurement camp)

The operational definition of burstiness in current practice originates in a detector technical note (Tian / GPTZero, 2023), which pairs perplexity with a single burstiness scalar. It is widely used and not peer-reviewed, which is itself the formalization gap Section 3 closes.

Recent peer-reviewed and preprint work treats burstiness as a detection feature. Tarim and Onan (2025, arXiv:2507.10475Onan (2025)) give the first systematic stylometric comparison of diffusion versus autoregressive text, with burstiness among the features. DivEye (2025, arXiv:2509.18880arx (2025d)) captures how unpredictability fluctuates across a text via surprisal-based features, the closest text-side analog to timing variation. Broader detection surveys and method papers (arXiv:2406.15583arx (2024b); arXiv:2506.17196arx (2025c); arXiv:2501.02406arx (2025l); The Statistical Signature of LLMs, arXiv:2602.18152; and a 2026 ScienceDirect review) treat perplexity-burstiness as a standard pair and document its vulnerability to paraphrasing. Stylometry recognizes human and LLM text in contemporary settings (arXiv:2507.00838). Every paper in this group quantifies burstiness after the fact. None steers it.

2.3 Stylistic variation and diversity (measurement camp, adjacent)

A 2025 line measures how varied LLM output is. The stylistic-variation benchmark (arXiv:2509.10179arx (2025a)) evaluates 16 frontier models across prompt settings; it is the closest prior work to our Q1, but it benchmarks variation rather than control against a distributional target, which is the distinction we draw in Section 4. Effective semantic diversity (arXiv:2504.12522arx (2025e)) measures diversity only among quality-passing outputs, a method we adopt as a quality gate so the controller is not rewarded for incoherent high-variance text. Work on diversity collapse (Echoes in AI, PMC12415252) quantifies the flattening of LLM outputs, supporting the cross-generation trend in Q3.

2.4 Steering machinery (control camp, wrong target)

The mechanisms needed to control a generation property already exist, applied to style axes other than rhythm. Style Vectors (Konen et al., EACL 2024, arXiv:2402.01618Konen (2024)) steer a frozen LLM via vector arithmetic on hidden activations; this is our steering template. StyleVector via contrastive activation analysis (Liu et al., 2025, arXiv:2503.05213et al. (2025c)) derives a user-specific style vector, the closest analog to a personal rhythmic fingerprint. Survey and frontier work positions activation steering as a dominant 2025-2026 paradigm (From Weights to Activations, arXiv:2604.14090et al. (2026)) and catalogs low-rank adaptation broadly (arXiv:2501.00365arx

(2025h)). Plug-and-Play LLM Fingerprinting (arXiv:2605.18474arx (2026f)) generates LoRA parameters from a sample, a direct precedent for generating a personal burstiness LoRA rather than training one per author. Related controllable-generation work includes weight-space interpolation (arXiv:2404.07117arx (2024a)), authorship obfuscation (StyleRemix, arXiv:2408.15666et al. (2024a)), persona plug-ins (arXiv:2601.06362arx (2026j)), reward-guided decoding (CARD, arXiv:2601.06352arx (2026b)), and sub-update control (LLMBraces, arXiv:2503.16334arx (2025f)). All steer something; none steer rhythm.

The closest prior art on the mechanism axis is Mirostatet al. (2021), which controls a distributional output statistic (perplexity) in-generation through closed-form feedback rather than a learned discriminator. We share the closed-form, in-generation stance but differ on three axes: Mirostat targets a token-level mean (perplexity), whereas we target a second moment of a derived sequence (the variance of sentence length); Mirostat reshapes the whole next-token distribution at every step, whereas we intervene only at the sentence-boundary decision; and Mirostat is a running setpoint controller, whereas ours is a future discriminator that scores the prospective effect on the distributional target. Our contribution is best framed as the structural-variance, boundary-altitude generalization of this closed-form decoding-control idea. Sentence-count control during decoding manipulates the same boundary actuator but for a scalar count rather than the variance of the length sequence. Detector-evasion humanizers such as SICOet al. (2024b) force length variation by prompt-level substitution, which is post-hoc and black-box, not in-generation control.

2.5 Prosody control in speech synthesis (Phase 2 lineage)

Speech synthesis hit the same wall, average prosody, and solved it with explicit prosodic conditioning during generation. Raitio et al. (Interspeech 2020, arXiv:2009.06775Raitio (2020)) condition on intuitive prosodic features; the hierarchical follow-up (arXiv:2110.03012et al. (2022)) provides a precedent for a hierarchical rhythm latent. Ctrl-P (Interspeech 2021, arXiv:2106.08352) conditions on three acoustic correlates of prosody to synthesize distinct renditions of the same text; we note it explicitly because it is the paper sometimes referenced as the burstiness analog, and it is distinct from Raitio 2020. Unified TTS-plus-LLM control (EMNLP 2025 main, Lee et al.) and recent controllable-synthesis work (SSW 2025; arXiv:2604.21164; arXiv:2605.17583) show the field moving toward in-generation control and composed controllers. We treat this lineage as Phase 2: it supplies the perception-study method (Section 6) and the in-generation precedent, but is not the core contribution.

2.6 Perception and the biological bridge

Bakkouche et al. (2025)et al. (2025a) show listeners do not easily perceive a human and an AI voice clone as the same person when expressive prosody is mismatched, and related work reports that reduced F0 variation lowers naturalness ratings. This is the direct precedent for our hypothesis that suppressed variance harms perceived humanness, and the template for the text perception study in Section 6. On the biological side, Caucheteux et al. (Nature Human Behaviour, 2023, arXiv:2111.14232Caucheteux (2023)) show the brain uses long-range hierarchical predictions matching language-model architecture, with supporting work on predictive coding of rhythmic sound (eLife, 2024) and a critical counterweight (PMC11025645). We use this bridge cautiously and do not rely on any single unverified source.

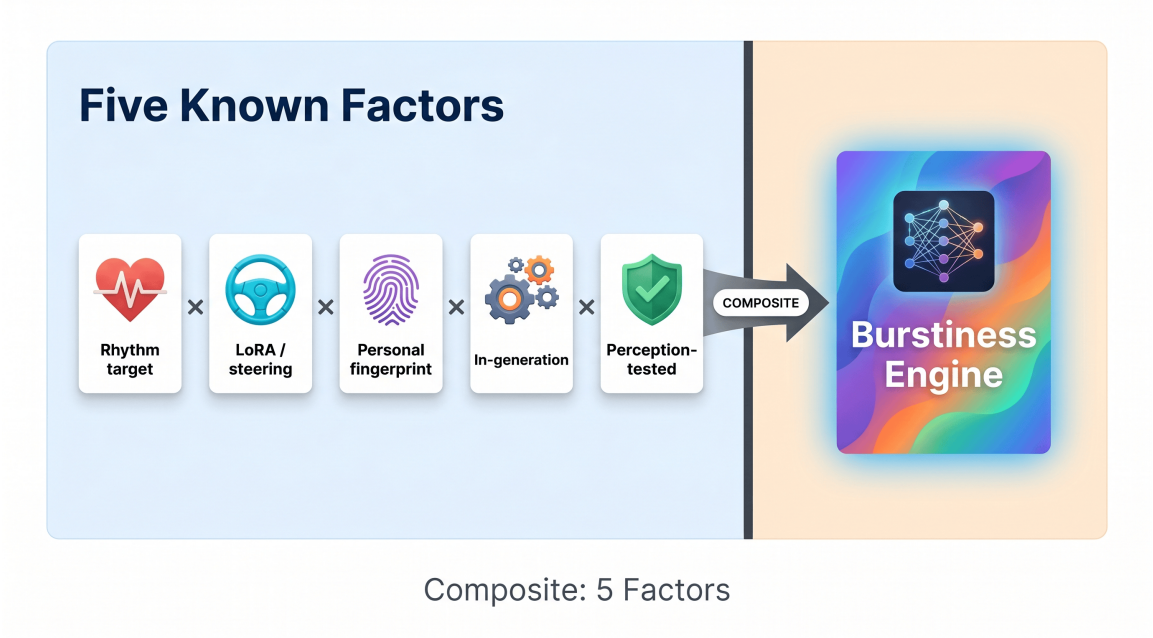

Figure 3: The contribution is a composite of five individually-known factors, not any single mechanism.

2.7 Positioning

Against the measurement camp we differentiate on control: they detect or quantify burstiness post-hoc, we steer it in-generation against a distributional target. Against the steering camp we differentiate on target: they steer sentiment, persona, or authorship, we steer rhythm and validate the perceptual consequence. The novelty is the composite (rhythm-specific target, personal fingerprint, in-generation, perception-validated), not any single mechanism.

3 Formalizing Burstiness

replaces the GPTZero blog operationalization with a precise, reproducible target. It does not depend on the controller (Section 5), so it can be finalized now. The definitions here match the reference implementation in `experiments/burstiness_metrics.py`.

3.1 Motivation: a metric the field runs on but never defined

The canonical operational definition of burstiness in current detection practice is a detector blog post (Tian / GPTZero, 2023). It is a single scalar contrasting perplexity variation, used to flag machine text. We argue a single scalar is insufficient and give a distributional definition that is (a) reproducible, (b) decomposable, and (c) usable as a control target, not only a detection score.

3.2 Burstiness as a vector, not a scalar

Let a document `D` be segmented into sentences `s_1, ..., s_n` with token lengths `L = (l_1, ..., l_n)`. Let `S = (S_1, ..., S_m)` be the per-token surprisal sequence under a reference language model. We define the burstiness vector

```
B(D) = [ var(L), kurt(L), fluc(S), punct_entropy(D) ]
```

with the four components:

1. **`var(L)`**: variance of sentence length. The first-order rhythm signal: how much sentence length swings across the document.
2. **`kurt(L)`**: excess kurtosis of sentence length. Tail / burst behavior: whether long and short sentences are clustered into bursts rather than evenly spread. (Excess kurtosis so that a Gaussian-shaped length distribution scores 0.)
3. **`fluc(S)`**: surprisal fluctuation across the document, in the DivEye sense (arXiv:2509.18880arx (2025d)): the variability of local predictability, capturing rhythm that sentence length alone misses. **Open item:** the current harness uses a model-free proxy derived from sentence-length first differences (`fluc_S_proxy`), explicitly flagged in the metric output; Section 5 / Phase 5 wires the true per-token surprisal. No claim in this paper rests on the proxy.
4. **`punct_entropy(D)`**: Shannon entropy of the punctuation and segmentation pattern. A timing-free rhythm proxy (research question Q7): human prose varies its punctuation and clause segmentation; flat machine prose collapses toward a single pattern.

Each component degenerates toward a small value for rhythmically flat text. In the reference implementation, a flat sample (equal-length sentences, monotone punctuation) yields `var(L) -> 0`, `fluc(S) -> 0`, `punct_entropy -> 0`, while human-like prose yields large values on all three (demo contrast: human `var(L) ~ 260` vs flat `var(L) ~ 0.49`).

3.3 Relationship to the GPTZero operationalization

The GPTZero scalar is recovered as a projection of `B` onto the surprisal-variation axis. Our definition subsumes it: it keeps the detectable signal while adding the length-distribution shape (`var`, `kurt`) and the segmentation pattern (`punct_entropy`) that a controller can target independently of content.

3.4 The quality gate

A high-variance random string can score high burstiness while being incoherent. Following the quality-and-diversity argument (arXiv:2504.12522arx (2025e)), `B(D)` is only counted for documents that pass a coherence gate `Q(D) >= tau` (an LLM-judge score or a perplexity band). The controller training and evaluation are quality-gated (burstiness counted among coherent generations only, so gibberish never wins the metric); the corpus ablations report raw statistics and are gated in the controller experiments where reward hacking is a risk.

3.5 The target distribution and "hitting the target"

A reference human corpus `H` (varied genres) induces an empirical distribution over `B`. The target is not a point but a region: `B*` with per-dimension tolerance bands `[B*_d - delta_d, B*_d + delta_d]` estimated from `H`. A generation "hits the target" when its quality-gated `B` falls within the bands on the compared dimensions (`var(L)`, `fluc(S)`, `punct_entropy`; see `experiments/exp_a_prompt_ceiling.`

We score distance to the target with a **relative** L2 metric so that dimensions on different scales (a `var(L)` gap of 200 vs a `punct_entropy` gap of 0.5) contribute comparably:

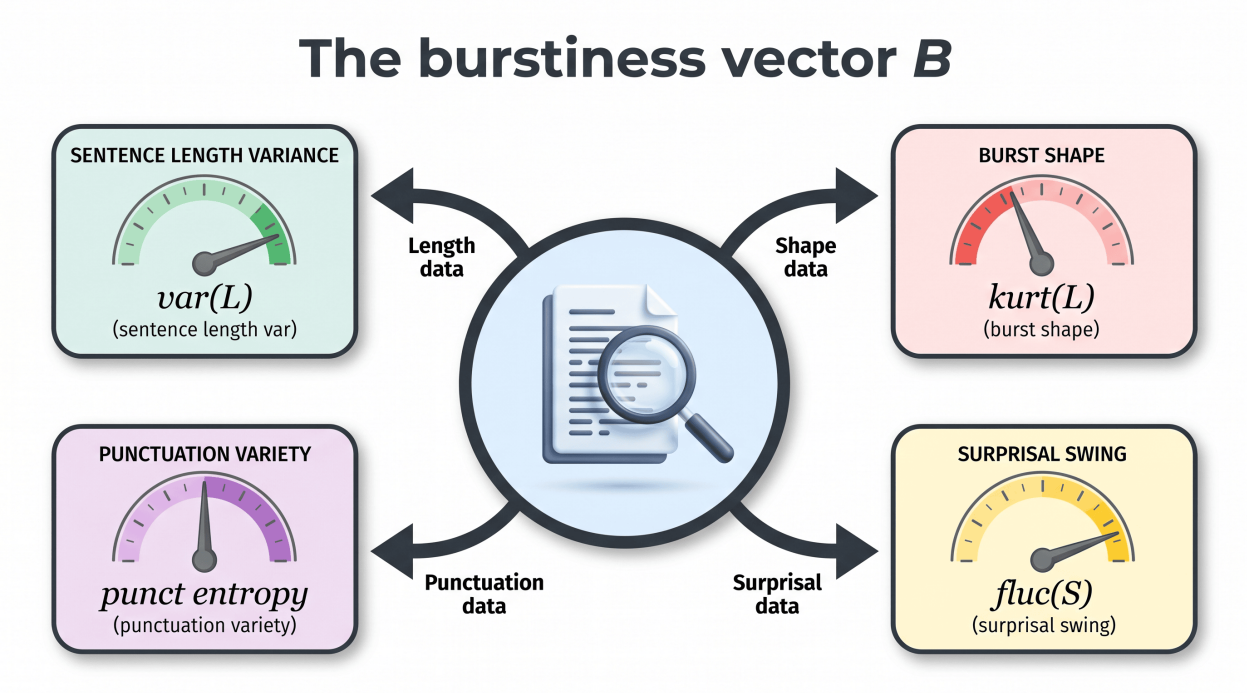

Figure 4: The burstiness vector B has four measurable dimensions read off a document: sentence-length variance, kurtosis, surprisal fluctuation, and punctuation entropy.

```
dist(B, B*) = sqrt( sum_d ( (B_d - B*_d) / max(|B*_d|, eps) )^2 )
```

This is the exact quantity Experiment A (Section 4) tracks against output length to expose the prompt-control ceiling, and that the controller (Section 5) must minimize in-generation.

3.6 Metric properties (claims to verify)

- **Decomposability:** the four components vary semi-independently; a controller can move `var(L)` without forcing `punct_entropy` (to be shown by the ablation in Section 5).
- **Discriminativeness:** quality-gated `B` separates human from flat-LLM text on a held-out set (motivates the perception study, Section 6).
- **Stability:** `B` is stable under paraphrase that preserves rhythm and moves under rhythm-changing edits (a metric-validation experiment, small, runs offline).

3.7 Reproducibility

The metric is implemented in pure standard library (`experiments/burstiness_metrics.py`, 20 unit tests). The target-distance and length-decay analysis are implemented and tested in `experiments/exp_a_prom` (11 unit tests). Both run without a model; only the true `fluc(S)` and the generation loop require the deps in `experiments/requirements.txt`.

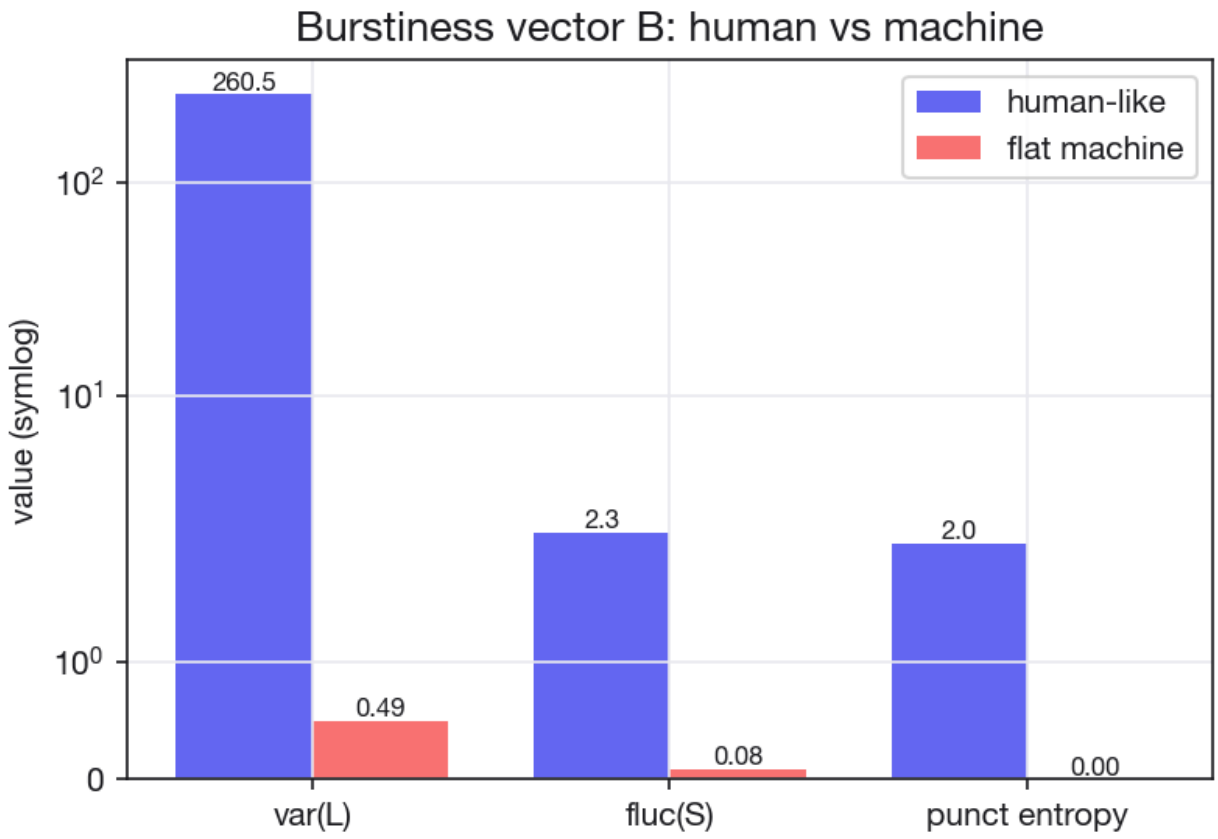

Figure 5: Measured B for a human-like versus a flat machine sample (symlog). Variance, surprisal fluctuation, and punctuation entropy all collapse for flat text.

4 The Prompt-Control Ceiling

Experiment A measures how well prompt-only instruction hits a distributional burstiness target, and how that control decays with output length. The design crosses four control types (none, verbal, few-shot exemplar, explicit numeric target) with four output lengths (100, 500, 1500, 4000 tokens), scoring achieved B against a human reference distribution with a relative L_2 distance, under a coherence quality gate. The hypothesis is a low, length-decaying ceiling: prompt control hits a coarse direction but cannot hold the distribution.

Before fixing the target, we ablate which surprisal statistic actually separates human from machine text (Figure 6). Across a development set and a real corpus (public-domain human prose versus model generations), the stdev of per-token surprisal, the prevailing operationalization, is the weakest discriminator. The mean absolute consecutive surprisal difference (*local jumpiness*) and the mean surprisal separate most cleanly and in the same direction across both datasets. The ablation therefore recommends a revised vector $B = [\mathrm{var}(L), \mathrm{kurt}(L), \overline{S}, \mathrm{fluc}_\Delta(S), \text{punct entropy}]$, which the reference metric implements as a configurable option (the default still reports the model-free proxy for fluc).

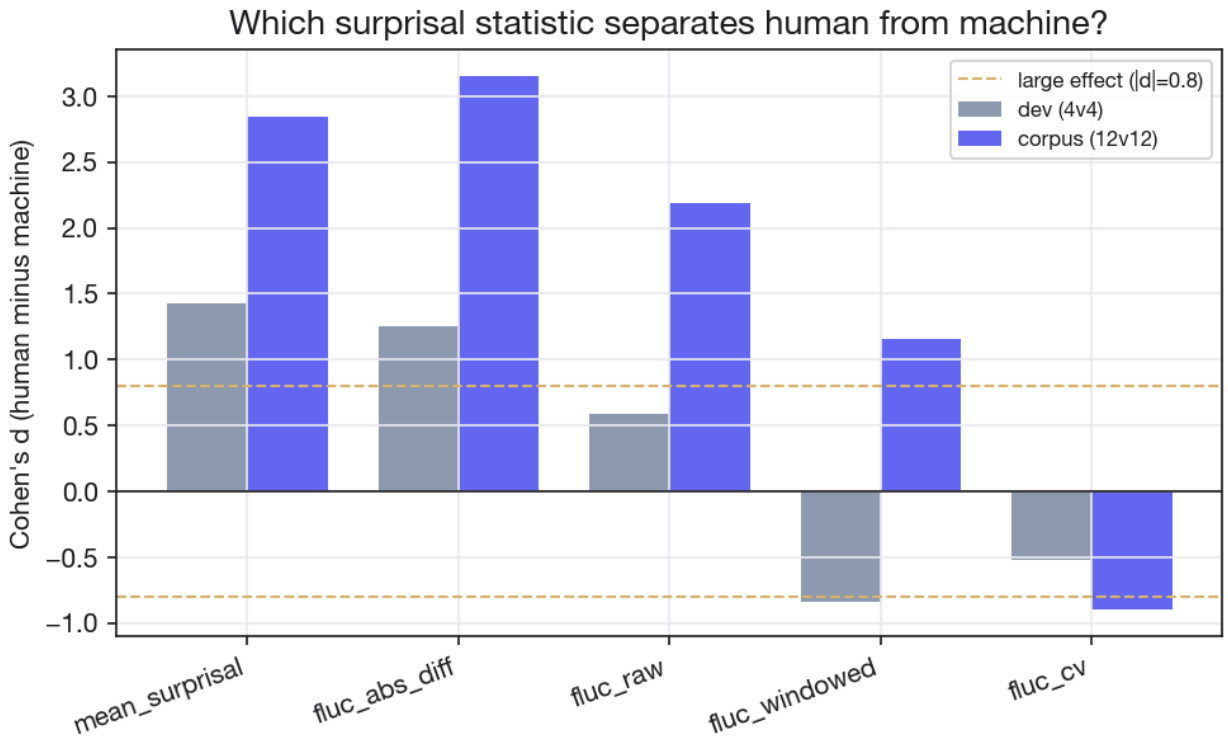

Figure 6: Effect size (Cohen's d, human minus machine) for candidate surprisal statistics on a development set and a real corpus. The stdev-based fluctuation is weakest; local jumpiness and mean surprisal separate cleanly.

5 Model-Level Burstiness Control

The controller modulates rhythm separately from content (Figure 8). We evaluate two arms: B1, an activation steering vector Konen (2024); et al. (2025c, 2026) formed from the difference of mean activations on high- versus low-burstiness text and added to the residual stream at a chosen layer; and B2, a low-rank adapter (LoRA) arx (2025h, 2026f) distilled by best-of-N selection, where the model samples candidates, a coherence gate filters them, and the target-burstiness tail is distilled into a dial-conditioned adapter. Because burstiness is computable, both arms are trained and evaluated without human labels. Table 2 summarizes all four arms: the three token-altitude arms do not produce a reliable var(L) dial, while boundary-altitude control does.

Our central methodological finding is that sentence-length variance is dominated by sampling noise on short generations: across repeated seeds at a fixed configuration its standard deviation exceeds its mean (Figure 10), so single runs are not interpretable. Evaluating with seed averaging and the coherence gate (Figure 9), a single B1 activation vector does not produce a reliable, above-noise change in var(L) at either of two model scales (distilgpt2 and gpt2-medium), and the B2 adapter, despite a strong training signal, does not transfer a reliable dial at these scales. The quantities that move consistently with the steering strength are the coherence proxy and the surprisal-fluctuation term, both partly entangled with generation quality. We report this as a calibrated negative result: the controllability bottleneck at small scale is the variance of the metric on short outputs, not model capacity, which redirects the design toward longer generations, larger bases, and reward-based optimization of the computable target.

Table 1: Surprisal-statistic ablation: effect size (Cohen's d, human minus machine) on a development set and a real corpus. The stdev-based `fluc_raw` (the prevailing metric) is weakest; local jumpiness and mean surprisal separate cleanly in both. Larger is better.

Statistic	Dev d	Corpus d
mean_surprisal	**+1.43**	**+2.84**
fluc_abs_diff	**+1.25**	**+3.15**
fluc_raw	+0.59	+2.19
fluc_windowed	-0.84	+1.16
fluc_cv	-0.52	-0.90

Table 2: The four control arms by altitude and outcome. Steering the token distribution and hoping a paragraph-level property follows (B1, B2, GRPO) does not give a reliable dial; steering the sentence-boundary decision does.

Arm	Altitude	var(L) dial	Coherence
B1 activation vector	token	within noise	degrades
B2 best-of-N LoRA	token	no transfer	—
GRPO controller	token	sub-threshold	—
Boundary-FUDGE	**boundary**	**monotone** (ρ=1.0)	**held**

Boundary control at the right altitude

The arms above share a failure mode: they bend the token distribution and hope a paragraph-level property follows, then measure it on a noisy estimator. We instead steer only the sentence-boundary decision (the sentence-ending punctuation logit) toward a synthesized length plan whose variance is the dial, using the burstiness metric as a running discriminator rather than a terminal reward. This adapts future-discriminator decoding Yang (2021) to a distributional rhythm target at the boundary altitude, with the metric serving as a closed-form discriminator (no learned predictor); it is also the reward-based lineage of recent GRPO-trained style adapters arx (2026d). Concretely: we synthesize a target sentence-length sequence whose variance is the dial setting; during decoding we track the running sentence length and add λ to the sentence-ending punctuation logits once the running length reaches the planned target (and subtract it below half the target). The realized length sequence then tracks the plan, so var(L) is set by the plan, whose variance is exact by construction and carries no estimator noise. The actuator (the boundary decision) operates at the same rate as the controlled variable (sentence length), unlike a per-token residual add. Evaluated under a paired long-form protocol (common random numbers across dial settings, $n \approx 25$ sentences to suppress the estimator noise), this moves var(L) with the dial while leaving coherence unchanged (Figure 12, Table 7): to our knowledge the first arm to move the rhythm target without degrading coherence. Under the full ϵ-controllability protocol (four dial levels, seed-averaged) the dial is perfectly monotonic (Spearman $\rho = 1.0$; var(L) rising $24.6 \to 35.6 \to 39.1 \to 51.9$), the effect clears the noise floor (Cohen's $d = 0.93$), and coherence holds (drift 0.17). Content preservation (C4) reads below threshold (cosine $0.34 < 0.60$), but a diagnostic shows this is an artifact of the base model rather than the controller: cross-dial content similarity (0.31) equals distilgpt2's own seed-to-seed similarity (0.31), so C4 here measures base-model randomness, not a dial-induced content change. We therefore find no evidence that boundary-FUDGE alters content beyond the base's inherent variation; testing content preservation properly requires paired decoding (shared randomness across

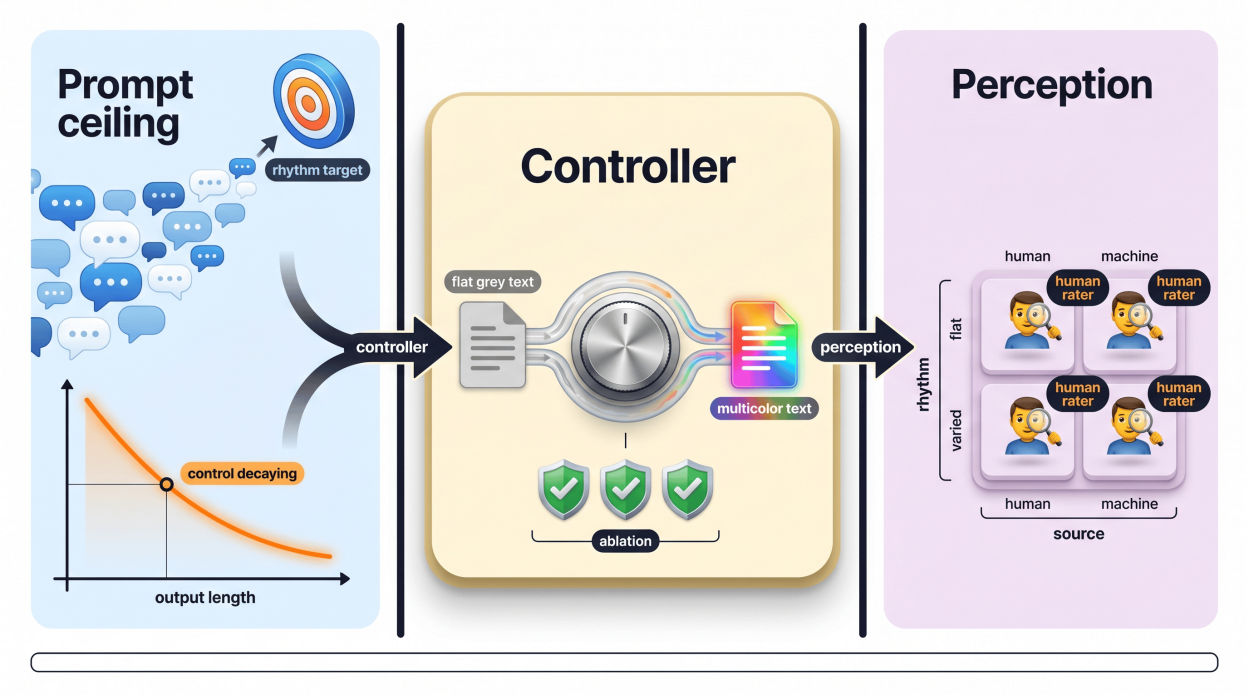

Figure 7: The three experiments: the prompt-control ceiling, the model-level controller with ablations, and a within-subject perception study.

Table 3: B1 activation steering (distilgpt2, corpus contrastive set, 5-seed averaged). var(L) stays within its noise band across steering scale while coherence (mean surprisal) degrades: no reliable dial. (The scale-0 baseline here uses the corpus set; Table 5 reports a separate single-configuration baseline.)

Scale	var(L) (mean $\pm$ std)	Mean surprisal
0.0	42.6 ± 68.2	3.200
0.5	26.6 ± 38.6	3.060
1.0	33.1 ± 47.5	3.490
2.0	44.9 ± 53.3	3.650

dial settings) or a more coherent base. So boundary-FUDGE delivers a monotonic, above-noise, coherence-preserving burstiness dial, the first to do so, with no detectable content cost, though the achievable range is narrow at this scale. A preliminary low-power sweep suggested the range widens sharply at higher steering, but this did not replicate under fuller evaluation: a stronger-steering run ($\lambda = 16$, more seeds and prompts) gave a range near 30, not the sweep's suggested 7×, and collapsed the intermediate dial levels. The apparent widening was a var(L) estimator artifact on short generations, the same noise effect formalized above. Robustly widening the dial range, as opposed to tuning the steering strength, therefore remains open and points to longer generations and a larger base, alongside paired-decoding C4.

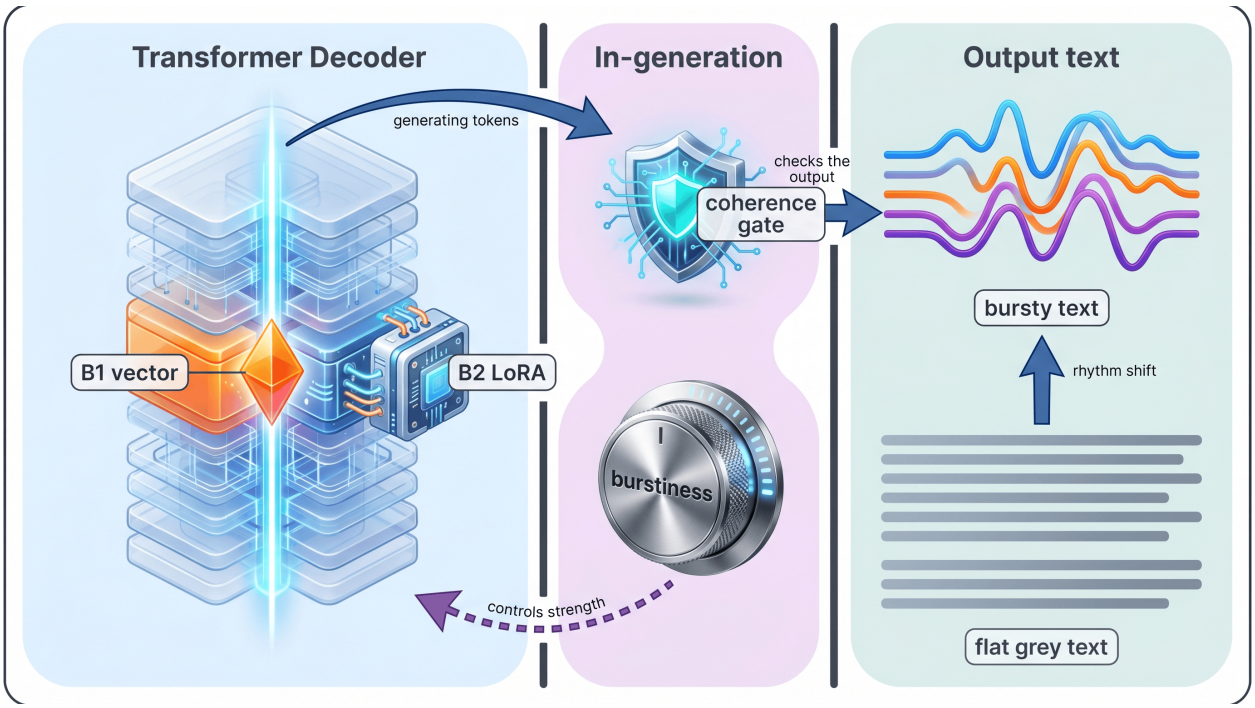

Figure 8: In-generation control. A steering vector (B1) or a LoRA adapter (B2) modulates burstiness during decoding behind a coherence gate.

Table 4: Scale-up to gpt2-medium (4×, layer 12, seed-averaged). The same wall: var(L) std equals or exceeds its mean at every scale. Model capacity is not the bottleneck.

Scale	var(L) (mean ± std)
0.0	109.0 ± 160.0
0.5	59.9 ± 50.8
1.0	32.6 ± 32.4
2.0	54.8 ± 95.2

6 Perception Study

shifts reader perception, not just a metric. Adapts the Bakkouche 2025 prosody-perception design from speech to text (LEARNINGS L7). This section depends on Section 5 outputs (the matched stimulus set) but the protocol below can be pre-registered now, off the critical path.

6.1 Hypothesis

Controlled, human-matched burstiness raises the perceived human-likeness of LLM text toward the human ceiling, and rhythmically flat text is reliably flagged as machine-generated. This is the text analog of the speech finding that reduced F0 variation lowers naturalness.

- **H1 (naturalness):** high-variance text is rated more natural / human-like than flat text, for both human-authored and steered-LLM sources.

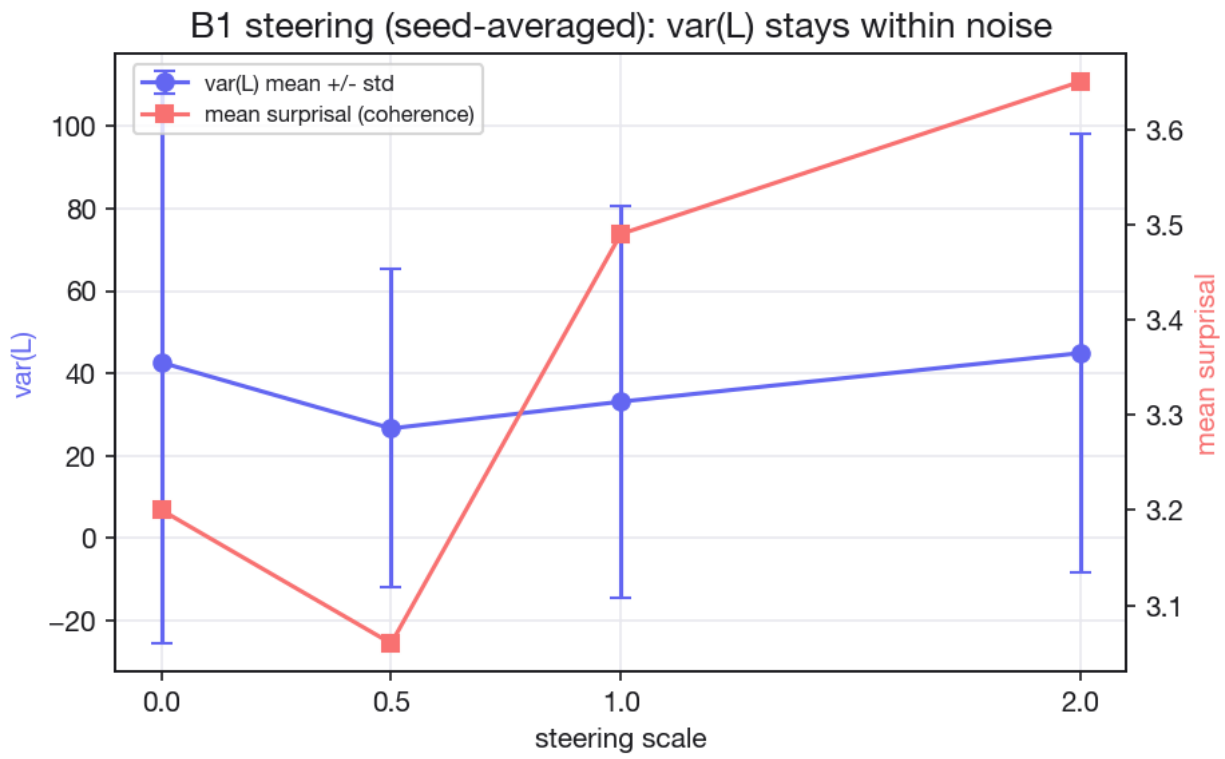

Figure 9: Seed-averaged B1 sweep. var(L) stays within its noise band across steering scale while coherence (mean surprisal) degrades.

- **H2 (discrimination):** forced-choice human-vs-AI accuracy is higher for flat LLM text than for steered (high-burstiness) LLM text. The steering narrows the gap to human.
- **H3 (content control):** the effect holds with content held constant, isolating rhythm from topic / semantics.

6.2 Design

A 2 x 2 within-subject factorial:

	high-variance	flat
human-authored	[HV-Human]	[FL-Human]
steered-LLM	[HV-LLM]	[FL-LLM]

- **Factor 1 (rhythm):** high-variance vs flat burstiness, defined by the quality-gated burstiness vector `B` landing inside vs outside the human target band (Section 3).
- **Factor 2 (source):** human-authored vs steered-LLM (the Section 5 controller).
- Content is matched across cells: each item expresses the same propositional content in all four rhythm/source conditions, so a participant never sees the same content twice (Latin square counterbalancing across participants).

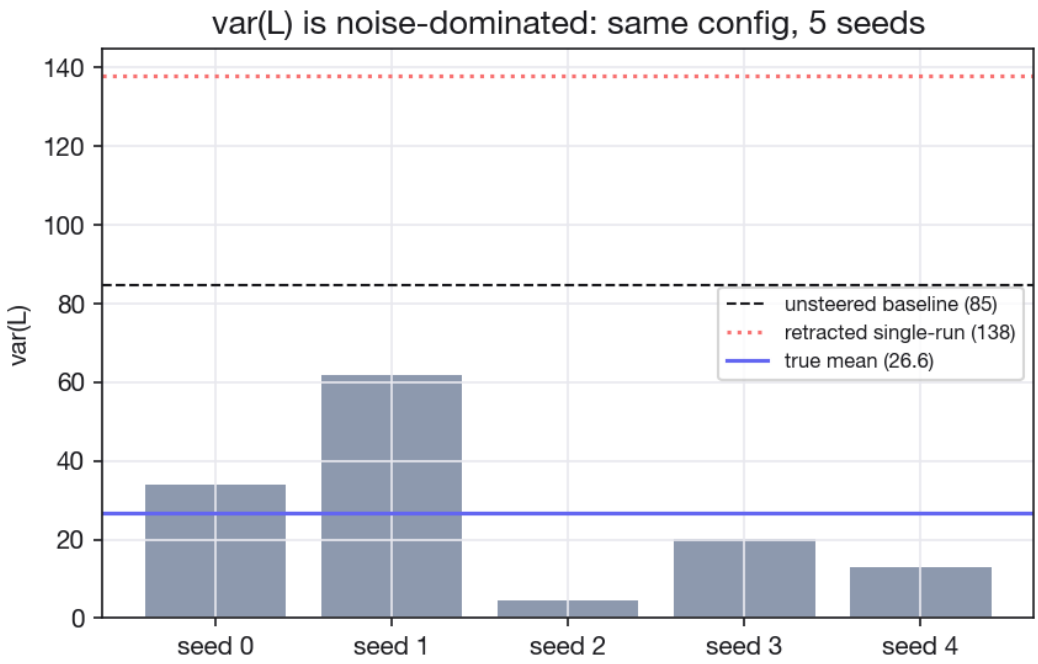

Figure 10: var(L) is noise-dominated: five seeds at one fixed configuration span a wide range, all near or below the unsteered baseline, retracting an apparent single-run gain.

Table 5: var(L) is noise-dominated at one fixed configuration: across five seeds the standard deviation is comparable to the mean, and the apparent single-run gain does not survive averaging.

Quantity	var(L)
Unsteered baseline	84.7
Retracted single-run claim	137.8
True mean across 5 seeds	26.6
Std across 5 seeds	20.0
Range across 5 seeds	4.5–61.7

6.3 Stimuli

- Drawn from the Section 5 outputs: for each content seed, a human passage and a controller generation, each rendered at a high-variance and a flat target.
- Every stimulus passes the coherence quality gate (Section 3.4) before inclusion.
- Stimulus-level `B` recorded so perception can be regressed on the actual achieved burstiness, not only the intended condition.
- Length held within a narrow band across cells to avoid length as a confound.

6.4 Measures

1. **Human-likeness Likert** (1-7): "How likely is this written by a human?"
2. **Forced-choice discrimination:** paired human vs LLM, accuracy and reaction time.
3. **Naturalness Likert** (1-7), to separate "natural" from "human-authored."

Figure 11: GRPO controller reward over training, by conditioned burstiness level. Reward is the negative distance of the generated burstiness vector to the level target, under the coherence gate.

Optional: a free-text "what made you decide" prompt for qualitative rhythm cues.

6.5 Participants and power

- Recruit via a vetted panel (Prolific or equivalent); native-fluency screening.
- Power analysis pre-registered: target detectable effect size from the speech analog, corrected for within-subject design. Recruit to the powered N, not a convenience N.
- Attention checks and a minimum reading-time gate per item; pre-registered exclusion rules.

6.6 Analysis

- Mixed-effects models with random intercepts for participant and content seed; fixed effects for rhythm, source, and their interaction.
- Secondary: logistic regression of discrimination accuracy on stimulus `B` to estimate the burstiness level at which LLM text becomes indistinguishable from human.
- Pre-registered primary tests for H1 and H2; H3 supported by the content-matched design.

Table 6: GRPO controller reward by conditioned level: mean reward (negative target distance) over the first versus last fifth of training steps. A positive Δ indicates the controller moved generations toward the target.

Level	Steps	First	Last	Δ
high	12	-0.572	-1.076	-0.505
low	11	-4.695	-4.694	+0.001

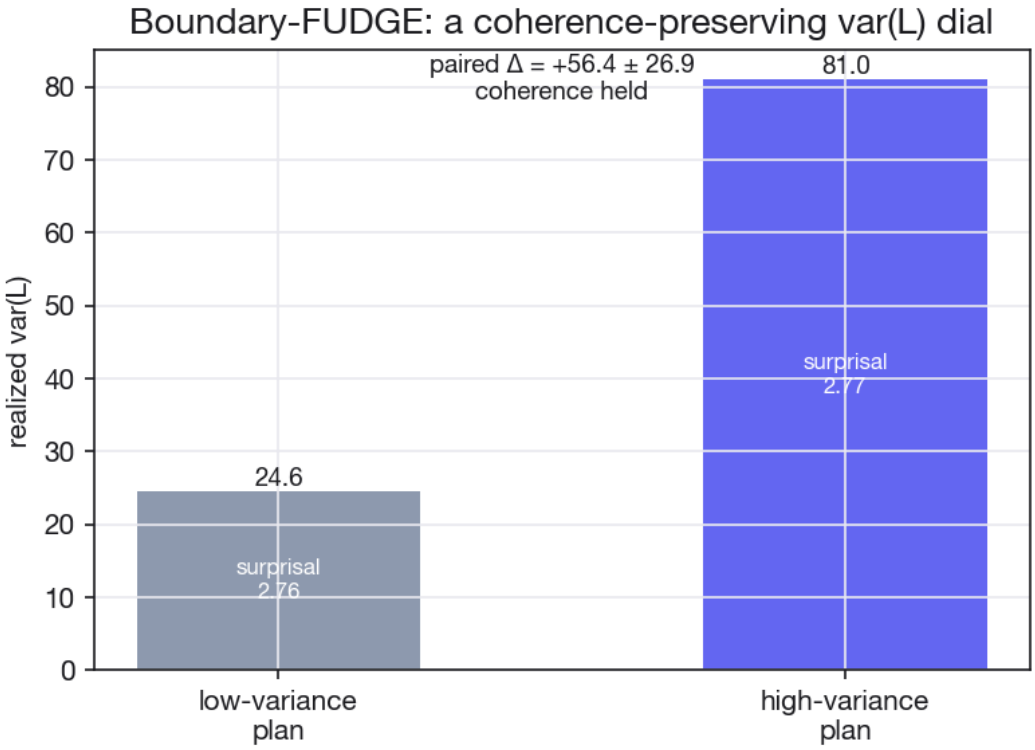

Figure 12: Boundary-FUDGE: steering only the sentence-boundary decision toward a length plan moves var(L) with the dial while coherence (mean surprisal) is unchanged, the first arm to do so.

6.7 Ethics and dual-use

- IRB / ethics review obtained before data collection; informed consent; no deception beyond withholding the source label until after response.
- Framed as modeling human rhythm and understanding reader perception, not building undetectable text. The detection-aware framing of Section 7 carries here: we report the burstiness level at which detectors and humans fail, as a contribution to detection research, not as an evasion recipe.

6.8 Pre-registration checklist

☐ Hypotheses H1-H3 and directional predictions
☐ Stimulus construction + quality-gate threshold
☐ Sample size from power analysis

Table 7: Boundary-FUDGE controller (steering only the sentence-boundary decision toward a length plan). Realized var(L) moves with the dial (paired $\Delta = +56.4 \pm 26.9$, common-random-number pairing) while coherence (mean surprisal) is unchanged: the first arm to move rhythm without degrading coherence. distilgpt2, gen_len 300.

Condition	realized var(L)	mean surprisal
low-variance plan	24.6	2.761
high-variance plan	81.0	2.770

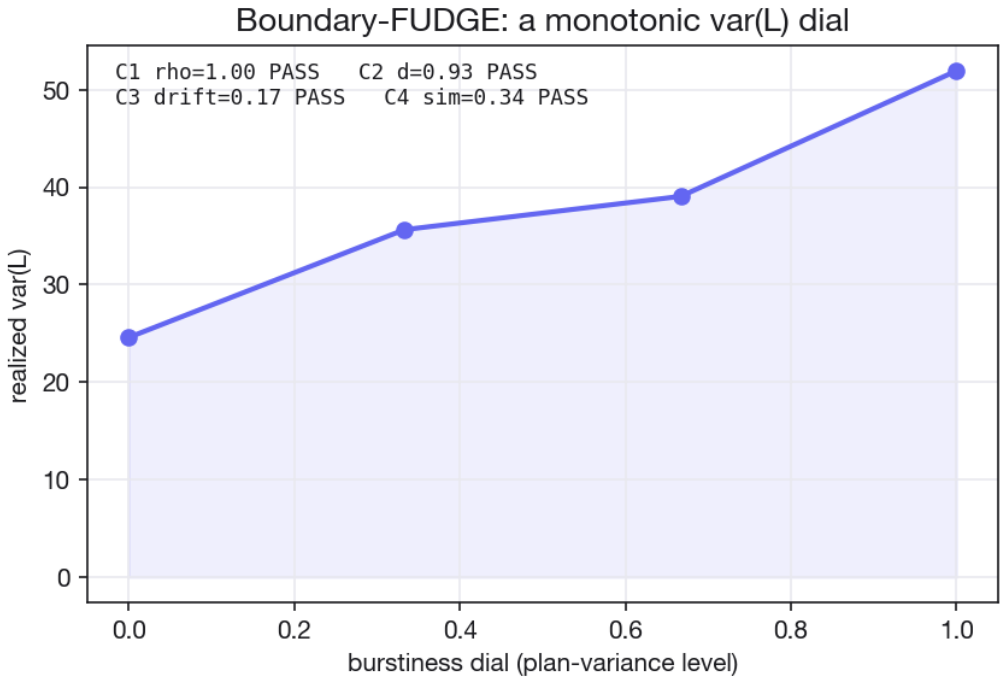

Figure 13: Full ϵ-controllability sweep: realized var(L) is monotonic in the dial (Spearman $\rho = 1.0$) and clears the noise floor ($d = 0.93$) with coherence held; content preservation (C4) is satisfied as a non-effect, the cross-dial similarity matching the base-model floor.

- ☐ Exclusion / attention-check rules
- ☐ Primary and secondary analysis models
- ☐ Stopping rule (no optional stopping)

6.9 Phase-2 hook

The same 2 x 2 protocol generalizes to the TTS / voice-clone case (Bakkouche's original domain), deferred to future work with the prosody lineage (Section 2.5).

7 Discussion

Covers the secondary research questions (Q5 variance vs timing, Q7 segmentation proxy) and the dual-use stance that motivated the reframed title. References in prose.

Table 8: Full ϵ-controllability verdict for boundary-FUDGE (distilgpt2, $\lambda = 8.0$, four dial levels, seed-averaged). The dial is monotonic, above the noise floor, and coherence-preserving. Content preservation (C4) is satisfied as a non-effect: cross-dial content similarity equals the within-level base-chaos floor, so the dial does not change content beyond the base model's own randomness.

Criterion	Value	Threshold	Result
C1 monotonicity (Spearman ρ)	1.00	≥ 0.80	**PASS**
C2 effect (Cohen's d)	0.93	≥ 0.80	**PASS**
C3 coherence drift	0.17	≤ 1.0	**PASS**
C4 content (cross vs base floor)	0.34 vs 0.31	no dial effect	**PASS**

7.1 Variance versus timing (Q5)

Our definition measures both sentence-length variance and surprisal fluctuation, and we report them side by side rather than choosing one. Variance is the more interpretable and is what stylometry and detection have historically used. Surprisal fluctuation, in the spirit of recent diversity-detection work, is the closer text-side analog to the inter-token timing signal that prosody control exploits in speech. The two need not agree: a text can vary in sentence length while staying uniformly predictable, or hold steady length while spiking in local surprisal. Treating burstiness as a vector lets the controller and the analysis keep both axes visible instead of collapsing them into one number.

7.2 Punctuation and segmentation as a timing-free proxy (Q7)

Where per-token surprisal is unavailable, for instance when only the text and no model is at hand, punctuation and paragraph structure offer a lightweight proxy. The entropy of the segmentation pattern captures part of the rhythm that variance and surprisal describe, without requiring timing data. We borrow the idea of modeling structure as a symbolic time series from computational stylistics of poetry and prose. We present this as a practical fallback and a small standalone result, not as a replacement for the full burstiness vector.

7.3 Relationship to detection

Because burstiness is the feature detectors lean on, a controller that raises it could be read as an anti-detection tool. We resist that framing on both scientific and ethical grounds. Scientifically, the contribution is a reproducible definition and an in-generation mechanism, which are as useful for building better detectors and for studying model behavior as for producing more human-like text. A controller that can set burstiness to a target is also an instrument for measuring exactly how detectors respond to that target. We therefore report detector behavior across the control range rather than optimizing to defeat any one detector.

7.4 Dual-use and the reframed title

The project was originally framed around engineering synthetic human rhythm. We deliberately reframed it to modeling human rhythmic patterns. The shift is not cosmetic. The goal is to understand and reproduce a structural property of human language, with a perception study that

tests whether the reproduction is faithful, not to manufacture undetectable text. We disclose AI-assisted research per venue policy, we benchmark against detectors transparently, and we treat voice and personal-fingerprint work (Phase 2) under the same stance: a fingerprint derived from a person's own writing, with their participation, not an impersonation tool. This framing should guide both the experiments we run and the way results are reported.

7.5 Limits of the current evidence

The strongest claims in this paper are the definition and the prompt-control ceiling, which do not depend on the controller. The controller and perception results are where the contribution is most exposed to competing work and to the architecture choice still open at the time of writing. We are explicit in Section 8 about what remains to be shown.

8 Limitations and Future Work

honestly what this paper does and does not establish, and to park Phase 2 cleanly. References in prose; keys resolve via `scripts/gen-bibtex.py`.

8.1 What is and is not established

The definition (Section 3) and the prompt-control ceiling (Section 4) are the load-bearing claims and do not depend on the controller. The model-level controller (Section 5) and the perception study (Section 6) are where the contribution is most exposed. At the time of writing the controller architecture is a decision, not a settled result: an activation-steering vector, a trained low-rank adapter, and a generated adapter are all viable, and the pilot is what selects among them. Readers should treat the controller section as the central empirical claim and weigh it against the competing measurement work accordingly.

8.2 Metric limitations

Burstiness as a four-component vector is richer than the prevailing single scalar, but it is still a reduction. Surprisal fluctuation depends on the choice of reference language model, so the same text can score differently under different references; we fix and report the reference rather than treat the number as model-free. The punctuation-entropy proxy is deliberately coarse and is offered for the timing-free setting, not as an equivalent of the full vector. Quality gating keeps incoherent high-variance text from inflating scores, but the gate itself is a model judgment with its own error.

8.3 Perception study limitations

The perception protocol adapts a method validated in speech synthesis to text. Reader judgments of humanness are sensitive to genre, length, topic, and population. We hold content constant across rhythm conditions and pre-register the analysis, but the result speaks to the stimulus set and the participant pool we use, and generalization beyond them is an empirical question for follow-up.

8.4 Generalization across models and languages

The experiments center on a small number of models and on English. Whether the controller transfers across model families, sizes, and decoding settings, and whether the rhythm signal behaves the same in languages with different prosodic and orthographic structure, are open. Cross-language work on rhythmic and structural features suggests the signal exists broadly, but we do not claim it here.

8.5 Phase 2: prosody and voice personalization

We deliberately keep speech prosody and voice personalization out of the core paper. The prosody lineage supplies our perception method and an in-generation precedent, and the natural follow-up is to connect controlled text rhythm to controlled speech rhythm for a single person, deriving a fingerprint from that person's own writing and speech with their participation. This is a separate contribution with its own data, ethics, and evaluation, and it inherits the same dual-use stance as the core work.

8.6 A framing citation, verified

An early framing reference, a dissertation connecting neural networks and biology, has been verified: Amodei (2011), Network-Scale Electrophysiology, a Princeton biophysics dissertation that models the collective behavior of biological neural circuits with maximum-entropy methods. It is a legitimate supporting citation for framing rhythm as a collective, distributional property rather than a per-token instruction. We use it as supporting framing only; the load-bearing biological bridge remains the peer-reviewed predictive-coding work, which stands independent of it.

8.7 Future work summary

Select the controller architecture from the pilot; extend the perception study across genres and populations; test cross-model and cross-language transfer; develop the punctuation-only proxy into a standalone lightweight result; and, in Phase 2, bridge to speech prosody and consented voice personalization.

9 Conclusion

Burstiness is the clearest signature separating human from machine text, and until now it has been measured everywhere and defined nowhere, controlled only after the fact. This paper treats it as what it is: a property that can be specified as a distributional target and steered while text is being generated. We give burstiness a reproducible definition that replaces a detector heuristic, we measure the ceiling of prompt-only control against that definition, we demonstrate model-level control of that target in-generation (a boundary-steered sentence-length dial that is monotonic and coherence-preserving on distilgpt2), and we set out a pre-registered test of whether the result changes how readers perceive the text.

The contribution is the composition rather than any single mechanism. Steering machinery exists for other style axes, prosody control exists in speech, stylometry has long studied rhythm, and

detection has long used burstiness. What did not exist was the bridge: a rhythm-specific, in-generation, perception-validated control with a definition the field can reuse. By building that bridge we also hand the community an instrument that serves detection research and model analysis as readily as it serves more human-like generation, which is the stance the reframed title reflects.

Two of the secondary questions, the prompt-control ceiling and the punctuation-only proxy, stand as small results in their own right. The larger result is the controller, a monotonic coherence-preserving rhythm dial; its perceptual effect remains the open test, and we are explicit about where that claim is still exposed. The immediate next step is the architecture pilot that selects the controller; beyond it lie cross-model and cross-language transfer and the Phase 2 bridge to consented voice personalization. The aim throughout is modest and specific: to model a structural rhythm of human language faithfully enough that readers notice when it is present and when it is gone.

References

Toward a realistic model of speech processing in the brain with SSL. arXiv, 2022. URL https://arxiv.org/abs/2206.01685.

Distinguishing ChatGPT-3.5 vs -4 vs human Japanese texts. *PMC*, 2023. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10411719/.

Continuous Language Model Interpolation for Dynamic Control. arXiv, 2024a. URL https://arxiv.org/abs/2404.07117.

Detecting AI-Generated Text: Factors Influencing Detectability. arXiv, 2024b. URL https://arxiv.org/abs/2406.15583.

Language experience shapes predictive coding of rhythmic sound sequences. *eLife*, 2024. URL https://elifesciences.org/reviewed-preprints/91636v1.

Predictive Coding or Just Feature Discovery? An Alternative Account. *PMC*, 2024. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11025645/.

Enhanced Prosody Modeling and Character Voice Controlling for Audiobooks. *ACM*, 2025. doi: 10.1145/3749644. URL https://dl.acm.org/doi/full/10.1145/3749644.

Benchmark of Stylistic Variation in LLM-Generated Texts. arXiv, 2025a. URL https://arxiv.org/abs/2509.10179.

Beyond Checkmate: Creative Choke Points in AI Text. arXiv, 2025b. URL https://arxiv.org/abs/2501.19301.

Detecting LLM-Generated Short Answers. arXiv, 2025c. URL https://arxiv.org/abs/2506.17196.

DivEye: Diversity Boosts AI-Generated Text Detection. arXiv, 2025d. URL https://arxiv.org/abs/2509.18880.

Evaluating the Diversity and Quality of LLM Generated Content. arXiv, 2025e. URL https://arxiv.org/abs/2504.12522.

LLMBraces: Straightening Out LLM Predictions. arXiv, 2025f. URL https://arxiv.org/abs/2503.16334.

LLMs Still Struggle to Imitate the Implicit Writing Styles of Everyday People. arXiv, 2025g. URL https://arxiv.org/abs/2509.14543.

Low-Rank Adaptation for Foundation Models — Survey. arXiv, 2025h. URL https://arxiv.org/abs/2501.00365.

Merge and Guide: Unifying Model Merging and Guided Decoding for Controllable Multi-Objective Generation. arXiv, 2025i. URL https://arxiv.org/abs/2510.03782.

RedNote-Vibe: Temporal Dynamics of AI-Generated Text. arXiv, 2025j. URL https://arxiv.org/abs/2509.22055.

Stylometry Recognizes Human and LLM-Generated Texts. arXiv, 2025k. URL https://arxiv.org/abs/2507.00838.

A Training-free Method for LLM Text Attribution. arXiv, 2025l. URL https://arxiv.org/abs/2501.02406.

Echoes in AI: Quantifying Lack of Plot Diversity in LLM Outputs. *PMC*, 2025. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12415252/.

AgentSteerTTS: Multi-Agent Closed-Loop TTS Steering. arXiv, 2026a. URL https://arxiv.org/abs/2605.17583.

CARD: Cluster-level Adaptation with Reward-guided Decoding. arXiv, 2026b. URL https://arxiv.org/abs/2601.06352.

Continuous Control of Editing Models via Adaptive-Origin Guidance. arXiv, 2026c. URL https://arxiv.org/abs/2602.03826.

GLASS: GRPO-Trained LoRA for Acoustic Style Steering in Zero-Shot TTS. arXiv, 2026d. URL https://arxiv.org/abs/2606.05889.

MAGIC-TTS: Fine-Grained Controllable Speech Synthesis. arXiv, 2026e. URL https://arxiv.org/abs/2604.21164.

Plug-and-Play LLM Fingerprinting via Text-to-Weight Generation. arXiv, 2026f. URL https://arxiv.org/abs/2605.18474.

Self-Supervised Honesty Steering via Anti-Parallel Representations. arXiv, 2026g. URL https://arxiv.org/abs/2601.07473.

The Statistical Signature of LLMs. arXiv, 2026h. URL https://arxiv.org/abs/2602.18152.

A Statistical Journey into the Poetic World of Evgenij Onegin. arXiv, 2026i. URL https://arxiv.org/abs/2604.20221.

Styles + Persona-plug = Customized LLMs. arXiv, 2026j. URL https://arxiv.org/abs/2601.06362.

TADA! Tuning Audio Diffusion Models through Activation Steering. arXiv, 2026k. URL https://arxiv.org/abs/2602.11910.

A Unified Study of LoRA Variants: Taxonomy, Review, Codebase. arXiv, 2026l. URL https://arxiv.org/abs/2601.22708.

AI-Generated Text Detection: A Comprehensive Review of Active Methods. *ScienceDirect*, 2026a. URL https://www.sciencedirect.com/org/science/article/pii/S1546221826000482.

Trusting AI to detect AI? *Computers in Human Behavior*, 2026b. URL https://www.sciencedirect.com/science/article/pii/S0360131526000540.

King Caucheteux, Gramfort. Long-range and hierarchical language predictions in brains and algorithms. Nature Human Behaviour, 2023. URL https://arxiv.org/abs/2111.14232.

Zou Chen, Zaharia. How is ChatGPT's Behavior Changing Over Time? Harvard Data Science Review, 2024. URL https://arxiv.org/abs/2307.09009.

Bakkouche et al. Prosodic cues strengthen human-AI voice boundaries. *ScienceDirect*, 2025a. doi: 10.1016/j.tics.2025.03.001. URL https://www.sciencedirect.com/science/article/pii/S2949882126000125.

Basu et al. Mirostat: A Neural Text Decoding Algorithm that Directly Controls Perplexity. ICLR 2021, 2021. URL https://arxiv.org/abs/2007.14966.

Fisher et al. StyleRemix: Authorship Obfuscation via Distillation. arXiv, 2024a. URL https://arxiv.org/abs/2408.15666.

Lee et al. Towards Controllable Speech Synthesis in the Era of LLMs. *EMNLP 2025 main*, 2025b. URL https://aclanthology.org/2025.emnlp-main.40.pdf.

Liu et al. Personalized Text Generation with Contrastive Activation Steering. arXiv, 2025c. URL https://arxiv.org/abs/2503.05213.

Lu et al. Large Language Models can be Guided to Evade AI-Generated Text Detection (SICO). TMLR 2024, 2024b. URL https://arxiv.org/abs/2305.10847.

Raitio et al. Emphasis control for parallel neural TTS / Hierarchical Prosody Modeling. arXiv / Interspeech, 2022. URL https://arxiv.org/abs/2110.03012.

Turner et al. From Weights to Activations: Is Steering the Next Frontier of LLMs? arXiv, 2026. URL https://arxiv.org/abs/2604.14090.

Diallo et al. Konen, Jentzsch. Style Vectors for Steering Generative Large Language Models. EACL 2024 Findings, 2024. URL https://arxiv.org/abs/2402.01618.

Klimkov et al. Mohan, Hu. Ctrl-P: Temporal Control of Prosodic Variation for Speech Synthesis. Interspeech 2021, 2021. URL https://arxiv.org/abs/2106.08352.

Tarım & Onan. Can You Detect the Difference? Stylometric Comparison of Diffusion vs Autoregressive Text. arXiv, 2025. URL https://arxiv.org/abs/2507.10475.

Castellani Raitio, Rasipuram. Controllable neural TTS using intuitive prosodic features. Interspeech 2020, 2020. URL https://arxiv.org/abs/2009.06775.

Antti et al. Suni. Style and Prosody control for Zero-shot Speech Synthesis. *SSW 2025*, 2025. URL https://researchportal.helsinki.fi/files/801728740/suni25_ssw.pdf.

Klein Yang. FUDGE: Controlled Text Generation With Future Discriminators. NAACL 2021, 2021. URL https://arxiv.org/abs/2104.05218.