SAME TOKENS, DIFFERENT READINGS: REPRESENTATIONAL AND CAUSAL TESTS OF LEXICAL-UNIT ENCODING IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Behavioral signatures, such as how a model's next-token probabilities move across a span, are routinely read as evidence about what language models represent internally. We ask what one such signature actually licenses, in a setting where the surface form can be held fixed: idiomatic and literal uses of the same expression. Across autoregressive models spanning 124M–122B, the surprisal-drop signal remains positive across model families. In the corpus-scale analysis, mean surprisal over the span does not strongly separate the two readings (Cohen's $d = 0.03$), whereas the first-to-last surprisal drop does ($d = 0.91$); residualizing frequency, collocation, span length and continuation predictability reduces this to $d = 0.14$, and matched controls retain positive effects ($d = 0.25$–0.52). We then ask whether this behavioral signal corresponds to internal structure. In three models (0.77B–14B), probes under nested expression-disjoint cross-validation over 402 expression types separate the readings at AUROC 0.981–0.991, against 0.535–0.564 for surface-form and context-free controls. Follow-up controls show that the preceding context alone reaches AUROC 0.989 in two models. Patching the final-word predictor state with a donor from the opposite reading of the same expression, matched on observed surface form and in-context token sequence, shifts final-word surprisal by +1.99 to +3.30 bits in difference-in-differences, while same-reading donors give intervals spanning zero (cross-minus-same-reading: +2.13 to +3.12 bits); the effect is near zero or negative before the expression and largest at the final-word predictor. Because random-expression donors are not reliably smaller, we report a reading-contingent causal component rather than an idiom-specific mechanism. More broadly, the study offers a template for validating and bounding behavioral signatures with representational and causal evidence.

1 INTRODUCTION

A central question in mechanistic interpretability is how neural language models compose token-level inputs into higher-level features, concepts, and associations (Elhage et al., 2021; Meng et al., 2022; Bricken et al., 2023). Multi-word expressions (MWEs) such as *kick the bucket* provide a controlled probe of this question: they are linguistically defined lexicalized units whose meanings and distributions are not always recoverable from their parts (Sag et al., 2002).

We investigate whether the continuation of a fixed expression depends on context-sensitive internal states. The behavioral readout is *surprisal drop*, the difference between first- and last-token surprisal, $\Delta_{\text{drop}} = s_{m_1} - s_{m_n}$. On MAGPIE with GPT-2-Large, the mean span surprisal is similar for idiomatic and literal uses (2.46 versus 2.43 bits; $p = 0.18$), whereas the endpoint trajectory differs. Much of this separation is shared with surface statistics: frequency, collocation, length, and continuation controls reduce the MAGPIE effect from $d = 0.916$ to 0.143 after residualization and 0.250 after matching, while matched DICE effects remain 0.392–0.524. We therefore test the internal states directly.

Operational claim. We use *latent lexical-unit encoding* for the functional hypothesis that, after a fixed-expression prefix, an internal state supports its continuation more strongly than a matched compositional use of the same surface form; *super-token* is shorthand for this hypothesis about internal

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

states, not a tokenizer unit, a single vector, or a dedicated head. We test it at three levels: behavioral separation (trajectory differences), representational availability (expression-disjoint probing), and causal contribution (matched interventions with explicit corruption controls).

Findings. Exact-expression cross-reading patches differ from same-reading patches in GPT-2-Large, Llama-3.1-8B, and Qwen2.5-14B and move idiomatic and literal recipients in opposite directions, whereas unrelated-expression donors raise surprisal in both readings. Expression-disjoint probes decode reading labels on held-out expression types, and the preceding context alone supports nearly the same AUROC. The central finding is thus context-dependent fixedness: internal support for continuing a fixed expression that depends on its context. The diagnostic is intended for comparative model analysis; classification results characterize it.

Contributions.

1. We operationalize first-to-last surprisal drop ($\Delta_{\text{drop}} = s_{m_1} - s_{m_n}$) as a behavioral signature for testing lexical-unit encoding in autoregressive language models, motivated by an information-theoretic account. The signal is trajectory shape rather than average phrase probability: idiomatic and literal uses have similar mean span surprisal ($d = 0.03$) but sharply different first-to-last drops ($d = 0.91$).
2. We provide controlled representational and causal evidence linking this behavioral signature to internal model structure across three model families, combining expression-disjoint probing (AUROC 0.981–0.991), same-expression activation patching with opposite- versus same-reading donors (cross-minus-same 2.13–3.12 bits), and held-out attention and head-output analyses. Context-only probes approach intact-probe performance, showing that much of the reading information is already available before the expression.
3. We map the scope of the signal: surface controls substantially reduce the raw behavioral effect (residualized $d = 0.143$, matched $d = 0.250$), while matched DICE effects remain 0.392–0.524 and same-expression pairwise accuracy is 0.654–0.716. Cross-corpus comparisons place the signal strongest for idioms, weaker for multi-word metaphor spans, and absent for sentence-level sarcasm; random-expression donor controls further bound the causal interpretation to a reading-contingent, context-dependent contribution rather than an idiom-specific mechanism.

2 Related Work

MWE and idiom detection. Multi-word expressions (MWEs) encompass a wide range of non-compositional constructions, including idioms, light verb constructions, and named entities (Sag et al., 2002; Baldwin & Kim, 2010). Idiom detection—distinguishing idiomatic from literal uses of the same expression—has received sustained attention, particularly through shared tasks. SemEval-2022 Task 2 (Tayyar Madabushi et al., 2022) provided a benchmark with multilingual training data and zero-shot/one-shot evaluation tracks. MAGPIE (Haagsma et al., 2020) introduced a large English idiom corpus drawn from the British National Corpus. DICE (Mi et al., 2025) provides a controlled contrastive set of idiomatic and literal contexts. Recent prompted-LLM work on DICE reports high instance-level accuracy but lower consistency across paired literal/figurative uses (Mi et al., 2025); our setting instead uses next-token probability dynamics as a prompt-free mechanistic diagnostic. Classical approaches to idiom detection use distributional statistics, including Canonical Form Score (Fazly et al., 2009), embedding-based compositionality (Salehi et al., 2015), sentence-level distributional semantics (Salton et al., 2016), and transformer probing (Garcia et al., 2021). We instead examine token-level probability dynamics.

Surprisal in NLP. Surprisal (negative log-probability) was introduced as a psycholinguistic predictor of reading difficulty by Hale (2001) and formalized by Levy (2008). Recent work has used LLM surprisal for syntactic evaluation (Hu et al., 2020) and as a probe for semantic structure. Alves et al. (2025) study regression-based surprisal slopes for MWE detection using GPT-2. First-to-last drop is a related endpoint statistic: over consecutive token positions, the least-squares slope is $-\Delta_{\text{drop}}$ for two tokens and $-\Delta_{\text{drop}}/2$ for three; for longer spans, intermediate values also affect the slope. We build on this cue: our contribution is its controlled representational and causal evaluation.

Controlled probing. Control tasks quantify the extent to which a probe can memorize type-level associations (Hewitt & Liang, 2019); decodable properties need not be used by the original model (Ravichander et al., 2021). Context-sensitive idiomaticity measures and contrastive examples likewise test sensitivity beyond lexical form (Garcia et al., 2021; Mi et al., 2025). Our surface-form and neutral-template baselines are lexical-form controls that complement control-task selectivity. We additionally compare intact prefixes with context-only and corrupted-prefix representations, and test causal contributions through matched activation replacement.

Mechanistic studies of idioms. Oh et al. (2025) use knockout and activation patching in Llama-3.2-1B to trace figurative and literal interpretations of 110 idioms, including the use of preceding disambiguating context; Gomes (2025) applies modified path patching to Gemma 2-2B and reports attention heads that recur across idioms. We ask what a behavioral trajectory cue reveals about internal states, using exact-expression donors with same-reading, unrelated-expression, and noise controls, since patching conclusions depend on the corruption method (Zhang & Nanda, 2024), and held-out head selection, under which the full head distribution is heterogeneous (Appendix B.4). Mi et al. (2026) relate model-internal idiom decomposability to human ratings, syntactic flexibility, and predictability; we instead test continuation under same-expression reading contrasts.

Figurative language. Metaphor detection has been approached through lexical resources, distributional semantics, and neural models (Shutova, 2010; Tsvetkov et al., 2014). The VUA20 corpus (Leong et al., 2020; CreativeLang, 2024) provides large-scale token-level metaphor annotations. Sarcasm detection (Joshi et al., 2017) has typically been framed as a sentence-level classification task. Our work connects these three strands by examining how surprisal dynamics vary across figurative language types, using phrase-level and sentence-level analyses to delimit where the drop diagnostic is informative.

3 METHOD

3.1 INFORMATION-THEORETIC MOTIVATION AND SPAN FEATURES

Given a sentence $x = x_1, x_2, \ldots, x_T$, the surprisal of token x_i given its left context is:

$$s_i = -\log_2 p_\theta(x_i | x_{<i}) \tag{1}$$

where p_θ is an autoregressive language model.

For a multi-word span $m = m_1, \ldots, m_n$, the mean surprisal $\bar{s}$ measures the average code length of the observed tokens. However, this aggregate metric is insensitive to *where* the uncertainty is resolved. This distinction is critical for fixed expressions.

Minimal predictive mixture. Let c be preceding context, r an observed expression prefix, and y its continuation. A motivating two-hypothesis account is

$$p(y \mid c, r) = q\, p(y \mid c, r, Z = \text{unit}) + (1-q)\, p(y \mid c, r, Z = \text{comp}), \quad q = p(Z = \text{unit} \mid c, r), \tag{2}$$

where the hypotheses denote conventionalized-unit prediction and composition. For a fixed continuation $y^\star$, write the two likelihoods as a and b. Holding them fixed, if $a > b$, then

$$\frac{\partial[-\log_2(qa + (1-q)b)]}{\partial q} = -\frac{a-b}{\ln(2)[qa + (1-q)b]} < 0. \tag{3}$$

Increasing the unit posterior then lowers continuation surprisal. A positive first-to-last drop further requires continuation surprisal to fall below span-entry surprisal, an empirical property of the spans rather than a consequence of the mixture. Frequency and collocation may affect both q and the component likelihoods, as causes, cues, or consequences of lexicalized prediction; residualization cannot separate these roles. Equations 2–3 motivate the diagnostic, and Section 4.2 tests the network directly.

Features and endpoints. Table 1 defines the reported trajectory scores. Here n counts aligned model subtokens; lexical-word endpoints are evaluated separately in Appendix B.3. The diagnostic

Table 1. Trajectory definitions. $f = s_{m_1}$, $l = s_{m_n}$, $\bar{s}$ is span-mean surprisal, and $\bar{s}_{\text{ctx}}$ is the context mean stored by the historical extractor (context delta is an auxiliary score). The auxiliary feature runner uses $\epsilon = 0.5$ bits. Historical feature-family conventions and clipping are audited in Appendix B.7.

Score	Formula	Interpretation
Drop	$f - l$	First minus last subtoken surprisal
Endpoint slope	$(f - l)/(n - 1)$	Positive average decrease
Span mean / std	$\bar{s}$ / σ_s	Level / within-span variation
Context delta	$\bar{s}_{\text{ctx}} - \bar{s}$	Context–span mean contrast
Boundary concentration	$f - \bar{s}$	Entry relative to span mean
Normalized drop	$(f - l)/\max(f, \epsilon)$	Proportional endpoint decrease
Length interaction	$(f - l)n$	Drop–length interaction

applies to multi-token spans, since a one-token span has no within-span trajectory. All newly audited surprisal outcomes use negative-log2 bits.

For $f \geq \epsilon$, normalized drop equals $1 - l/\max(f, \epsilon)$, so the audited auxiliary classifier removes the duplicate collapse-ratio feature; the trajectory-family *relative drop* $(f - l)/(f + l + 1)$ is distinct (Appendix B.7).

Testable contrast. The mechanistic test replaces the predictor state using another context of the same expression and compares cross-reading with same-reading donors. Surface-form matching fixes the expression; context-only probes, corruption controls, and a donor audit address the remaining context differences.

3.2 Threshold and Auxiliary Classifiers

For secondary validation, we classify an expression as idiomatic if $\Delta_{\text{drop}} > \tau$, with τ selected to maximize macro-F1 on a held-out development set. This thresholded rule tests whether the trajectory signal separates idiomatic and literal uses. Auxiliary supervised and encoder-plus-trajectory hybrid baselines are described in Appendix A.9.

4 Experiments

4.1 Datasets

MAGPIE (Haagsma et al., 2020) is the large-scale English idiom resource used for corpus-level analyses. We use the 21,733 raw-train examples whose target compound aligns to a contiguous sentence span, split 80/10/10 ($17{,}386/2{,}173/2{,}174$); each split is 83.9% idiomatic.

DICE (Mi et al., 2025) supplies repeated idiomatic and literal uses of the same expressions. Our analysis contains 2,066 sentences (1,033 per label) and 402 canonical expression types. Stored partitions contain 1,652/207/207 examples; the new analyses instead use canonical-expression folds. All uses of one canonical expression stay in the same fold. DICE's literal contexts were generated with GPT-4 and expert-validated, whereas its figurative contexts come from MAGPIE and AStitchInLanguageModels (Mi et al., 2025; Tayyar Madabushi et al., 2021) and are longer (28.1 vs. 15.4 words on average). Reading labels therefore coincide with collection source, and contrasts between DICE readings can also carry source and style cues. MAGPIE is corpus-derived and idiom-heavy and serves corpus-scale characterization; DICE supplies the within-expression contrasts used for context-conditioned inference.

SemEval-2022 Task 2 (Tayyar Madabushi et al., 2022) provides multilingual idiomaticity-detection data. We use it only for the supervised encoder baseline and aligned dev-set hybrid comparison.

VUA20 and sarcasm controls. For the figurative-language comparison, we use VUA20 (Leong et al., 2020; CreativeLang, 2024) to extract consecutive multi-word metaphor spans (≥ 2 metaphor-labeled tokens) and matched literal spans.

We also use the TweetEval irony subset (Barbieri et al., 2020), derived from SemEval-2018 Task 3 (Van Hee et al., 2018), as a sentence-level sarcasm/irony negative control; after alignment and profiling filters it contains 2,859 tweets (1,444 ironic/sarcastic and 1,415 non-ironic).

The sarcasm experiment tests whether surprisal drop is a generic figurative-language signal or specific to phrase-level fixedness.

Model coverage and default extractor. MAGPIE full-corpus runs use GPT-2-Large as the default surprisal extractor because it is cheap to profile at corpus scale and easy to reproduce. To test whether the effect survives in newer autoregressive families, DICE also includes Qwen2.5/Qwen3 (Yang et al., 2024; 2025) and Llama-3.1 (Grattafiori et al., 2024) checkpoints. The Qwen3.5-122B-A10B row (Qwen Team, 2026) adds parameter-scale coverage with a hybrid linear+full-attention sparse MoE architecture, which differs from the dense GPT-2 and Llama rows (Section 5.3). An exploratory localization survey (Appendix Table 32) additionally includes Mistral-7B (Jiang et al., 2023) as a non-Llama dense transformer of comparable scale.

4.2 Controlled Internal-State Evidence

Matched residual-stream interventions. We patch the post-block residual at the position immediately preceding the first subtoken of the final lexical word. The predeclared layer is $\mathrm{round}(0.75(L-1))$: zero-based layers 26/36, 23/32, and 35/48 for GPT-2-Large, Llama-3.1-8B, and Qwen2.5-14B. Fold 0 provides 318 contexts from 60 canonical types per model; fold 1 is development and folds 2–4 supply training-only means. Donors match canonical type, normalized observed surface, and the exact in-context token-ID sequence. Cross-label donors have the opposite reading; same-label donors are distinct contexts with the same reading. Caches are disabled and logits recomputed after each intervention.

Let $\delta_{e,l}^{(k)}$ be the context-averaged change from clean final-word-onset surprisal for expression e, recipient label l, and intervention k. The expression-balanced estimands are

$$D_k = \frac{1}{|E|}\sum_{e\in E}\left(\delta_{e,\mathrm{idiom}}^{(k)} - \delta_{e,\mathrm{literal}}^{(k)}\right), \qquad C = D_{\mathrm{cross}} - D_{\mathrm{same}}. \tag{4}$$

Writing $A_{k,r}$ for the type-averaged change in recipients with reading r, $C = C_I - C_L$, where $C_I = A_{\mathrm{cross},I} - A_{\mathrm{same},I}$ and $C_L = A_{\mathrm{cross},L} - A_{\mathrm{same},L}$. We use 10,000 percentile bootstrap resamples of canonical expression types (base seed 42; stored derived seeds) for 95% intervals. Table 2 reports D_{cross} and C. Same-label DiD intervals include zero in all three models. Decomposing C by recipient reading shows that the contrast is bidirectional: relative to same-reading donors, opposite-reading donors raise idiomatic-recipient surprisal ($C_I = 1.11$ [$0.35, 1.86$], 1.53 [$0.92, 2.18$], and 1.52 [$0.83, 2.26$] bits) and lower literal-recipient surprisal ($C_L = -1.02$ [$-1.87, -0.11$], -1.58 [$-2.17, -1.02$], and -1.59 [$-2.29, -0.93$] bits; Table 3). Unrelated-expression donors instead raise surprisal in both readings by 10.8–14.8 bits; the cross-minus-random contrast therefore compares a bidirectional shift with a same-sign disruption, and its intervals include zero. Cross-reading donors also come from the other collection source (corpus-derived vs. GPT-4-generated) and differ more from their recipients than same-reading donors in predictor position (1.91–1.97 tokens), sentence length (5.15–5.35 tokens), and clean onset surprisal (0.73–0.85 bits; Appendix Table 17), and part of the cross-reading effect may reflect these context differences. Appendix B.1 reports each condition, noise controls, and position sweeps. Earlier and pre-MWE matched effects are near zero or negative, and the final-predictor effect is largest. Because nominal first- and middle-token positions often coincide with or follow the predictor, we compare first-token positions strictly preceding it: the final-predictor mean-ablation effect exceeds the first-token effect by 1.06–1.25 bits in all three models (Appendix B.1).

Expression-disjoint probing and context controls. We evaluate contextual span-mean representations with five outer folds disjoint over all 402 canonical types. Layer, regularization, scaling, and preprocessing are fitted or selected inside training partitions. AUROC is 0.981 [$0.975, 0.987$], 0.991 [$0.988, 0.994$], and 0.991 [$0.987, 0.994$], versus 0.535–0.543 for surface-only features and 0.553–0.564 for neutral-template representations. Duplicate-sentence and annotation-marker audits pass. These comparisons establish context-dependent decodability for held-out types.

Table 2. Matched DICE interventions at final-word onset, in bits. Each model uses 318 contexts/60 canonical types. D_{cross} is idiom-minus-literal DiD; C subtracts same-label DiD. Intervals are 95% expression-bootstrap intervals. $D_{\text{cross}} - D_{\text{random}}$ intervals include zero.

Model	D_{cross}	$C = D_{\text{cross}} - D_{\text{same}}$	$D_{\text{cross}} - D_{\text{random}}$
GPT-2-Large	$1.99\ [0.26, 3.61]$	$2.13\ [0.59, 3.63]$	$-0.10\ [-1.42, 1.20]$
Llama-3.1-8B	$3.24\ [2.04, 4.50]$	$3.11\ [1.99, 4.24]$	$0.57\ [-0.42, 1.60]$
Qwen2.5-14B	$3.30\ [1.93, 4.76]$	$3.12\ [1.88, 4.45]$	$1.04\ [-0.28, 2.34]$

Table 3. Recipient decomposition at final-word onset, in bits (318 contexts/60 types per model; 95% expression-bootstrap intervals). C_I and C_L subtract same-reading from cross-reading effects within idiomatic and literal recipients, so $C = C_I - C_L$; positive values worsen prediction. Random-expression columns give raw recipient changes. Full rows are in Table 16.

	Cross minus same		Random-expression donor	
Model	C_I (idiomatic)	C_L (literal)	Idiomatic	Literal
GPT-2-Large	$1.11\ [0.35, 1.86]$	$-1.02\ [-1.87, -0.11]$	$14.16\ [12.72, 15.54]$	$12.07\ [10.60, 13.58]$
Llama-3.1-8B	$1.53\ [0.92, 2.18]$	$-1.58\ [-2.17, -1.02]$	$14.75\ [13.42, 16.09]$	$12.09\ [10.95, 13.20]$
Qwen2.5-14B	$1.52\ [0.83, 2.26]$	$-1.59\ [-2.29, -0.93]$	$13.05\ [11.67, 14.45]$	$10.80\ [9.51, 12.14]$

Table 4. Subsequent context controls: AUROC [95% CI] on the common 2,032-context/402-type sample in each model. Selection uses grouped inner folds; uncertainty resamples expressions. High scores persist without the original MWE prefix.

Representation	Llama-3.1-8B	Qwen2.5-14B
Intact predictor	$0.984\ [0.979, 0.989]$	$0.989\ [0.984, 0.992]$
Pre-MWE context	$0.989\ [0.986, 0.993]$	$0.989\ [0.985, 0.993]$
Neutral prefix: something	$0.985\ [0.980, 0.990]$	$0.991\ [0.987, 0.994]$
Neutral prefix: thing	$0.986\ [0.981, 0.991]$	$0.987\ [0.982, 0.991]$

Prefix controls show where this information is available (Table 4). Llama and Qwen are compared on the same 2,032 contexts/402 types per model using nested five-outer/three-inner grouped folds. Pre-MWE context alone reaches AUROC 0.989 in both models. Replacing the visible MWE prefix with neutral token IDs while preserving its length, position, and preceding context produces similarly high scores. Original-minus-neutral paired AUROC differences range from -0.002 to 0.001, and all their intervals include zero. The discriminative information is thus largely present in the preceding context. Consistent with their different sources, idiomatic DICE contexts have longer prefixes than literal ones (about 16 vs. 9 tokens in the causal subset; Appendix Table 23), so part of this information is available from length alone. Neutral replacements are token corruptions, and the original span-mean representation also sees the final word.

Held-out attention and head-output interventions. We select the top five heads using training expression types and evaluate the unchanged composite on held-out types. Attention is measured from the final-word predictor query to preceding/available MWE keys. Paired d_z is 0.672, 0.828, and 1.063 for GPT-2-Large, Llama, and Qwen, with positive intervals in every model. All-head BH-FDR tests find significant effects of both signs and negative median effects, indicating selective, heterogeneous routing (Appendix B.4). Later interventions replace selected head-output slices before the attention output projection. On the 60 held-out DICE types, the predeclared top-five cross-minus-same DiD exceeds the mean contrast for five layer/count-matched random sets by $0.083\ [0.033, 0.139]$ bits in Llama and $0.082\ [0.041, 0.129]$ in Qwen. These component effects are small relative to residual patching; Appendix B.5 reports single-head, negative, and direct-route results.

4.3 Behavioral Model Survey

A raw-drop survey spans GPT-2, Llama, and Qwen checkpoints up to 122B parameters (Appendix Table 33). AUROC ranges from 0.616 to 0.652 in the autoregressive rows, with positive pooled

Table 5. Surprisal drop as a signal of lexical fixedness across figurative language types. Phrase-level: multi-token spans. Sentence-level: full tweet/sentence. Metaphor rows use VUA20 multi-word spans (≥ 2 consecutive metaphor-labeled tokens).

Type	Level	N	Mean Δ_{drop}	Std	Cohen's d
Literal	phrase	3,497	1.006	3.43	(ref)
Idiomatic	phrase	18,236	4.115	3.39	0.912
VUA20 literal spans	phrase	19,231	0.275	4.30	(ref)
VUA20 metaphor spans	phrase	2,823	2.214	4.21	0.452
Sarcastic	sentence	1,444	2.313	4.39	−0.013
Non-sarcastic	sentence	1,415	2.368	4.38	(ref)

Notes. Cohen's d is computed relative to the reference row within each block; *(ref)* denotes that reference distribution. We report effect sizes rather than p-values because all per-group sample sizes are large. Dataset sources are described in Section 4.

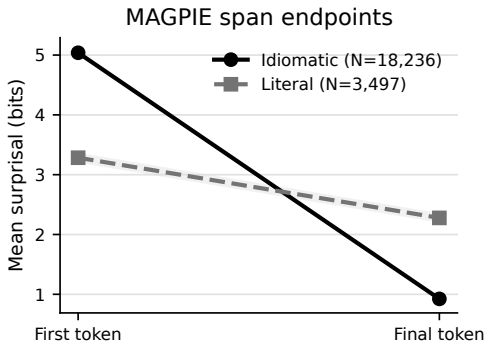

Figure 1. First-to-last surprisal endpoints for MAGPIE idiomatic and literal spans. Idiomatic spans collapse from approximately 5.0 to 1.0 bits; literal spans decline more gently, from approximately 3.3 to 2.3 bits. The diagnostic signal is the difference in endpoint drop.

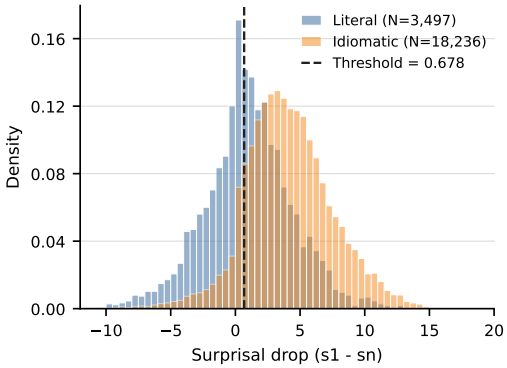

Figure 2. MAGPIE drop distribution ($N = 21{,}733$). Dashed line: dev-tuned threshold $\tau = 0.678$; Cohen's $d = 0.91$. Idiomatic mass is shifted to the right.

effects. Per-model threshold sweeps are descriptive, and masked-LM pseudo-likelihood rows measure a different scoring operation.

4.4 Figurative Language Types

Table 5 reports phrase-level idiom and multi-word metaphor spans together with the sentence-level sarcasm control. Token-level VUA20 metaphoric tokens are less predictable than literal tokens (4.822 vs. 4.342 bits; $d = 0.156$), with POS-matched gaps of $+0.613$ to $+0.628$ bits (Appendix Table 7); Section 5.2 interprets the ordering.

5 Analysis

5.1 Why Trajectory, Not Mean?

On MAGPIE ($N = 21{,}733$), mean surprisal differs little between idioms and literals (2.46 vs. 2.43 bits; $p = 0.18$), while drop is 4.12 vs. 1.01 bits ($d = 0.91$). Figures 1–2 show that idiomatic trajectories decline sharply whereas literal trajectories are flatter.

The discriminative information in these measurements is concentrated in trajectory shape. A partial regression compares the measured predictors: after standardizing all predictors on MAGPIE train+dev, drop remains the dominant coefficient even when mean surprisal, within-span variance, and span length are included in the same model (Table 6, left).

Table 6. Trajectory-shape controls on MAGPIE. **Left:** standardized logistic-regression coefficients on train+dev with bootstrap 95% CIs. **Right:** pooled-corpus drop effect by span length; N_i and N_l are counts of idiomatic and literal spans, D_i and D_l are mean surprisal drops, and d is Cohen's d comparing idiomatic vs. literal Δ_{drop} within each length stratum. We report effect sizes rather than significance markers because the per-stratum sample sizes are large ($N_i + N_l \geq 4{,}229$ at every length).

Predictor	$\hat{\beta}$	95% CI
Δ_{drop}	$+0.889$	$[+0.834, +0.952]$
Mean surprisal $\bar{s}$	$+0.062$	$[+0.016, +0.115]$
Within-span std σ_s	$+0.021$	$[-0.052, +0.094]$
Span length n	$+0.455$	$[+0.406, +0.501]$

Len	N_i	N_l	$\mathbf{D}_i$	$\mathbf{D}_l$	d
2	3,296	933	3.33	1.03	0.638
3	8,148	2,131	3.97	0.65	0.945
4	5,136	414	4.49	2.65	0.713

Span length robustness. The span-length panel in Table 6 shows that the effect is robust for 2–4 token spans, with the strongest effect at length 3 ($d = 0.945$). We omit the unstable 5-token case because it has only six literal examples.

5.2 Surprisal Drop as a Lexical Fixedness Signal

Table 5 and Figure 5 delimit the scope of the drop signal. It is strongest for idioms ($d = 0.912$), weaker for multi-word metaphor spans ($d = 0.452$), and absent for sentence-level sarcasm ($d \approx 0$), consistent with a phrase-level fixedness signal. The rows come from different corpora, annotation processes, and linguistic levels, so the ordering is qualitative.

The full-tweet sarcasm contrast is near null (2.31 vs. 2.37 bits; $p = 0.74$).

Surface controls quantify how much of the behavioral effect is shared with surface statistics. MAGPIE's control-pipeline raw $d = 0.916$ falls to residual $d = 0.143$ and matched $d = 0.250$; DICE matched effects remain 0.392–0.524, and same-expression pairwise accuracy is 0.654–0.716 (Appendix Table 10). Because the controls include final-token surprisal, one endpoint of the drop, the adjusted effects measure sensitivity to these covariates (Appendix A.5). Frequency and collocation can themselves participate in prediction, so the matched interventions in Section 4.2 provide the functional test.

5.3 Model Coverage, Tokenization, and Context Transfer

The DICE raw-drop survey is non-monotonic with parameter count (Appendix Figure 6). Architecture, training, and checkpoint selection vary jointly across these rows. The full matched DICE battery covers three models (0.77B–14B); the 70B checkpoint retains behavioral/probing coverage and the 122B checkpoint its behavioral coverage.

Endpoint robustness. Four audited definitions use the first/final span subtokens, the onset subtokens of the first/final lexical words, sums over those words, or their mean subtoken surprisals. Type-paired d_z ranges are 0.596–0.618, 0.697–0.720, and 0.578–0.583 for GPT-2-Large, Llama, and Qwen. The word-onset effects remain positive when both endpoint words have one subtoken. GPT-2-Large and Qwen exclude 20 sentence-initial cases lacking a preceding predictor; Llama excludes none. Appendix B.3 reports every endpoint, sample denominator, and alignment restriction.

Exploratory transfer to new contexts. We subsequently evaluated 20 idiomatic/literal and 20 compositional/compositional expression pairs (80 sentences drafted with AI assistance and accepted after the researcher's batch-level verification). These are new contexts, with possible expression overlap in head-selection training. For each pair, we average the absolute onset-surprisal change across donor directions. The idiom-minus-compositional difference in top-five-minus-random head sensitivity is 0.068 $[0.025, 0.114]$ bits in Llama and 0.040 $[0.015, 0.068]$ in Qwen. Within-idiom intervals include zero; the group interaction is partly driven by negative compositional-group contrasts.

An exploratory extension evaluates 34 further checkpoints, including multiple Qwen3.5 variants, on the same 40 pairs. After Holm correction, positive contrasts occur in 0/34 residual-minus-noise and 4/8 head tests (Appendix B.6).

5.4 Secondary Predictive and Calibration Analyses

Classification characterizes the trajectory diagnostic. The auxiliary tables use surface-form-grouped DICE folds; these differ from the canonical-type nested probe protocol. Gradient-boosted trajectory classifiers give modest gains over the raw-drop threshold (Appendix Tables 11 and 30); Appendix B.7 documents the feature audit, stored-endpoint discrepancies, and the scope of these auxiliary results. Calibration and ablation summaries are in Appendices A.4–A.7.

5.5 Failure Analysis

The threshold classifier is sensitive to weak-collapse idioms, rare forms, short spans, and context-dependent reversals of the same expression, and local collocation is associated with drop on MAGPIE literal uses (Pearson $r = 0.176$; Appendix A.8).

6 Discussion

The results separate information availability from prediction sensitivity. Reading labels are highly predictable from the preceding text, and neutral-token prefixes leave decodability essentially unchanged; in DICE, this predictability can also reflect collection source. Swapping the predictor state between opposite readings of the same expression nonetheless shifts final-word surprisal in opposite directions for idiomatic and literal recipients, while earlier and pre-expression replacements have near-zero or negative effects. Models thus carry label-predictive information in the preceding context and use context-dependent states at the point of local continuation. Trajectory signals, probe accuracy, and intervention magnitude each measure a different part of this process.

7 Limitations

Evidence covers English and aligned contiguous spans; inflection uses the annotated observed form, and discontinuous or unaligned uses fall outside the endpoint analysis. Productive compounding and character-heavy segmentation require language-specific units and resources. The matched battery covers three models. Cross-minus-random intervals include zero. Cross-reading donors differ systematically from their recipients in position, length, clean surprisal, and collection source, and a donor-gap-adjusted sensitivity model leaves the Llama cross-minus-same interval including zero. Context-only probes recover reading labels nearly as well as intact prefixes, so contextual, source, and predictability cues remain candidate explanations for decodability. Attention effects are heterogeneous, and head interventions identify small component contributions. The 40-pair pilot is purposively constructed, lacks item-level ratings and inter-annotator agreement, and retains position/length imbalance after matching; reusing it across checkpoints keeps the linguistic sample fixed, and head count, precision, architecture, and training vary across models. GPT-2-Large gives the widest primary causal intervals. Full head-to-head and head-to-MLP circuit tracing remains future work.

8 Conclusion

Surprisal dynamics distinguish idiomatic and literal uses, with substantially smaller effects under surface controls. Exact-expression interventions support a label-contingent internal contribution to continuation, beyond same-label replacement, that moves idiomatic and literal recipients in opposite directions. Expression-disjoint probing establishes contextual availability, and context-only probes show that this information is largely present before the expression. Held-out attention and small head-output effects add heterogeneous component contributions. Together, the results indicate context-dependent internal support for continuing the tested fixed expressions, which we characterize functionally rather than as an atomic token. Separating reading from donor-context and source differences, for example with minimally paired contexts, is the next step toward an idiom-specific mechanism.

AI USE STATEMENT

Generative AI tools were used to assist with writing and editing, literature retrieval and discovery, and aspects of research ideation and execution, including experimental code and analysis. AI assistance was also used to draft materials for the exploratory context-pair pilot, which were reviewed by the authors before use. The authors take responsibility for the final manuscript, analyses, claims, citations, and research artifacts.

ETHICS STATEMENT

This study analyzes existing English idiomaticity datasets and pretrained language models; no human-subject data were collected. The exploratory pilot uses 80 sentences drafted with AI assistance and accepted after researcher batch verification. The diagnostic is an analysis tool rather than a deployed language-assessment system, and its findings should not be generalized to untested languages or varieties of English (Appendix A.12). Existing datasets and checkpoints remain subject to their original terms of use (Appendix A.13).

REPRODUCIBILITY STATEMENT

Appendix B specifies the analysis populations, intervention hooks, donor construction, bootstrap units, probing splits, and alignment rules; Appendix A.14 lists training settings, seeds, and hardware. An anonymized numerical supplement contains saved result rows and summaries, the reanalysis scripts, and SHA-256 manifests (Appendix B.8). Model weights and raw activation caches are not included.

REFERENCES

Diego Alves, Sergei Bagdasarov, and Elke Teich. Surprisal dynamics for the detection of multi-word expressions in English. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 1185–1194, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-303-6. doi: 10.18653/v1/2025.findings-ijcnlp.72. URL https://aclanthology.org/2025.findings-ijcnlp.72/.

Timothy Baldwin and Su Nam Kim. Multiword expressions. In *Handbook of Natural Language Processing*, pp. 267–292. CRC Press, 2nd edition, 2010.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1644–1650, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.148. URL https://aclanthology.org/2020.findings-emnlp.148/.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features.

CreativeLang. CreativeLang/VUA20. Hugging Face Dataset, 2024. URL https://huggingface.co/datasets/CreativeLang/vua20_metaphor.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac

Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103, 2009.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. Probing for idiomaticity in vector space models. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3551–3564, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.310. URL https://aclanthology.org/2021.eacl-main.310/.

Andrew Gomes. Anatomy of an idiom: Tracing non-compositionality in language models. arXiv:2511.16467, 2025. URL https://arxiv.org/abs/2511.16467.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Hessel Haagsma, Johan Bos, and Malvina Nissim. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of LREC*, 2020.

John Hale. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL*, pp. 159–166, 2001.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of ACL*, pp. 1725–1744, 2020.

Albert Q. Jiang et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. Automatic sarcasm detection: A survey. *ACM Computing Surveys*, 50(5), 2017.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pp. 18–29, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.figlang-1.3. URL https://aclanthology.org/2020.figlang-1.3/.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. Rolling the DICE on idiomaticity: How LLMs fail to grasp context. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7314–7332, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.362. URL https://aclanthology.org/2025.acl-long.362/.

Maggie Mi, Golzar Atefi, Atsuki Yamaguchi, Felix Gers, Aline Villavicencio, and Nafise Sadat Moosavi. Rethinking the idiomaticity decomposability hypothesis: Evidence from distributional learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 34208–34229. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1579. URL https://aclanthology.org/2026.acl-long.1579/.

Soyoung Oh, Xinting Huang, Mathis Pink, Michael Hahn, and Vera Demberg. Tug-of-war between idioms' figurative and literal interpretations in LLMs. arXiv:2506.01723, 2025. URL https://arxiv.org/abs/2506.01723.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI technical report, 2019.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In *Proceedings of CICLing*, pp. 1–15, 2002.

Bahar Salehi, Paul Cook, and Timothy Baldwin. A word embedding approach to predicting the compositionality of multiword expressions. In Rada Mihalcea, Joyce Chai, and Anoop Sarkar (eds.), *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 977–983, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1099. URL https://aclanthology.org/N15-1099/.

Giancarlo D. Salton, Robert J. Ross, and John D. Kelleher. Idiom token classification using sentential distributed semantics. In *Proceedings of ACL*, pp. 194–204, 2016.

Ekaterina Shutova. Models of metaphor in NLP. In *Proceedings of ACL*, pp. 688–697, 2010.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3464–3477, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.294. URL https://aclanthology.org/2021.findings-emnlp.294/.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pp. 107–121. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.semeval-1.13. URL https://aclanthology.org/2022.semeval-1.13/.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. Metaphor detection with cross-lingual model transfer. In *Proceedings of ACL*, pp. 248–258, 2014.

Cynthia Van Hee, Els Lefever, and Veronique Hoste. SemEval-2018 task 3: Irony detection in english tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pp. 39–50, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1005. URL https://aclanthology.org/S18-1005/.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020.

An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.16042.

A Supplementary Material

A.1 Dataset Statistics

The profiled MAGPIE splits contain 17,386 train examples (14,588 idiomatic / 2,798 literal), 2,173 development examples (1,824 / 349), and 2,174 test examples (1,824 / 350), matching the aggregate main-text counts (18,236 idiomatic / 3,497 literal). DICE contains 2,066 balanced examples (1,033 / 1,033), SemEval dev contains 739 examples (403 / 336), and the aligned sarcasm subset contains 2,859 tweets (1,444 sarcastic / 1,415 non-sarcastic).

A.2 VUA20 Token-Level Metaphor Control

Table 7 records the token-level VUA20 statistics summarized in Section 4. These numbers are reported as a control for token-level metaphoricality before extracting multi-word metaphor spans.

Table 7. VUA20 token-level metaphor control. POS-matched rows report the range of metaphor-minus-literal surprisal gaps across matched POS controls.

Statistic	Value	Interpretation
Metaphoric token mean surprisal	4.822 bits	Less predictable tokens
Literal token mean surprisal	4.342 bits	Reference condition
Cohen's d	0.156	Small token-level gap
POS-matched gap range	$+0.613$–$+0.628$ bits	Robust to POS matching

A.3 Trajectory-Shape Family Comparison

Table 8 compares raw drop with related endpoint and boundary scores. The trajectory-family implementation uses relative drop $(f-l)/(f+l+1)$, collapse ratio $1-l/\max(f,1)$, and boundary concentration $f-\bar{s}$. These are distinct implementation-specific scores, separate from the normalized drop of Table 1, which coincides with $1-l/\max(f,\epsilon)$ when $f \geq \epsilon$ (Appendix B.7). Dashes indicate unmeasured entries, not zero effects.

A.4 Threshold Calibration Details

Table 9 reports threshold, calibration, and label-free thresholding results. These results are relevant for deployment, but they are secondary to the paper's mechanistic claim. The main takeaway is that drop is separable across settings, while absolute threshold calibration is dataset-dependent.

Appendix Figure 4 shows the corresponding threshold sweep: MAGPIE development thresholds transfer within MAGPIE but not to DICE.

Table 8. Trajectory-shape family comparison. MAGPIE values are held-out test results; DICE values summarize autoregressive scaling runs.

Score	MAG. AUROC	MAG. d	DICE AUROC	DICE d
Drop	0.756	0.909	0.634	0.495
Slope	0.712	0.648	–	–
Relative drop	0.783	1.208	0.639	0.497
Collapse ratio	0.784	0.989	0.650	0.424
Boundary concentration	0.745	0.821	–	–
Context delta	0.502	-0.012	–	–
Span mean	0.502	0.025	–	–

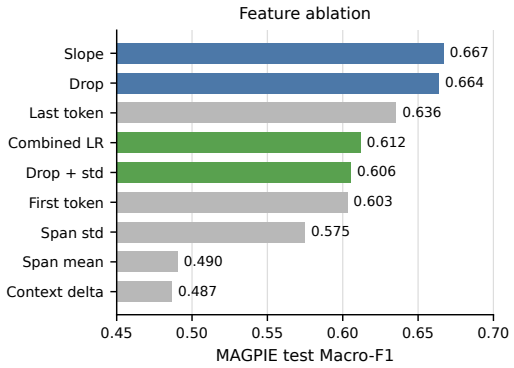

Figure 3. MAGPIE feature ablation. Slope and drop dominate mean surprisal.

Table 9. Threshold, calibration, and label-free thresholding results. MAGPIE uses dev-tuned thresholds and held-out test evaluation; DICE-L3.1 uses Llama-3.1-8B. *Traj-core* is logistic regression on trajectory features; centered drop subtracts a leave-one-out per-expression mean. Label-free rows use only target-corpus drop values.

Data	Setting	τ	Macro-F1	Rec	AUROC
MAGPIE test	drop, τ_{F1}	0.678	0.664	0.874	0.756
MAGPIE test	drop, rec$\geq$ 0.90	–	**0.671**	**0.914**	0.756
MAGPIE test	traj-core	–	0.671	0.854	**0.772**
DICE-L3.1	drop, τ_{F1}	2.80	0.611	0.601	0.652
DICE-L3.1	drop, rec$\geq$ 0.90	–	0.518	**0.906**	0.652
DICE-L3.1	centered drop	–	**0.673**	0.751	**0.720**
DICE-L3.1	z-score ($z = 0$)	2.95	0.606	–	–
DICE-L3.1	Otsu	3.37	0.603	–	–
DICE-L3.1	GMM	4.24	0.581	–	–
DICE-L3.1	MAGPIE τ transfer	0.678	0.544	–	–
MAGPIE test	z-score ($z = 0$)	3.62	0.539	–	–
MAGPIE test	Otsu	3.78	0.530	–	–
MAGPIE test	GMM	4.82	0.455	–	–

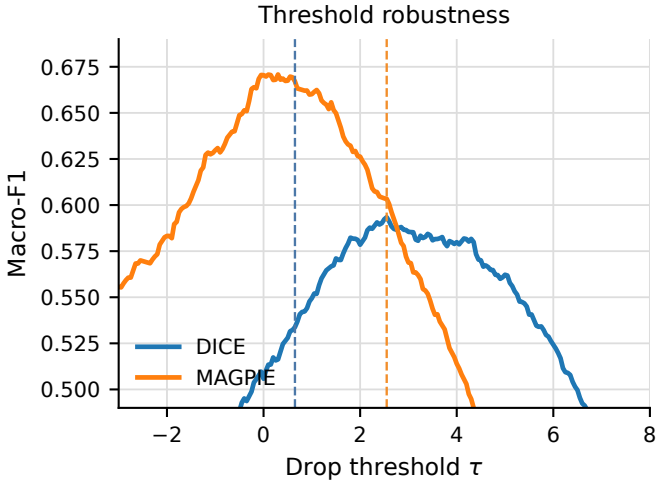

Figure 4. Threshold sweep. MAGPIE dev transfers to MAGPIE test but not DICE.

A.5 Robustness Controls

Table 10 summarizes two surface-statistical controls. The archived control pipeline operates on stored feature rows. Its 16 controls are expression word length, character length, GPT-2 subtoken length, sentence word length, left- and right-context word lengths, relative expression position, log phrase count, log expression count, mean and minimum log word count, mean bigram log count, mean adjacent PMI, final surprisal, and two stored surprisal-summary statistics (sentence mean surprisal and its standard deviation in the DICE loader). Frequency and association counts are estimated from the local MAGPIE and DICE texts rather than an external frequency corpus.

Residualization standardizes the controls and fits ordinary least squares to drop over the complete analyzed sample; Table 10 reports this global version. An out-of-fold variant that fits the scaler and regression on training partitions and forms residuals on held-out rows gives DICE effects of 0.183–0.367 (global: 0.190–0.383). Its folds group the observed MWE field (591 groups on DICE), not the 402 canonical types of the new probes and interventions. Neither version identifies a neural mediator. Final surprisal is itself a component of drop: if drop equals first minus final surprisal, residualized drop coincides with residualized first-token surprisal (stored drop deviates from this identity in 12 DICE rows per audited model; Appendix B.7). The adjusted effects are therefore sensitivity analyses rather than decompositions of the drop.

Matching operates within expression word-length strata. The minority label supplies queries and the other label supplies candidates; the remaining covariates are standardized on candidate rows, up to ten Euclidean nearest neighbors are retrieved per query, and unused candidates are assigned greedily, visiting queries in order of increasing nearest-neighbor distance. Matching enforces neither equal expression identity nor a distance caliper, and imbalance remains: on DICE, the maximum absolute post-match standardized mean difference is 0.380–0.519 across checkpoints, with 1,328–1,372 matched contexts. Same-expression ranking removes the dataset-level threshold by comparing every idiomatic use with every literal use of the same observed surface expression, with ties weighted 0.5.

Table 10. Robustness controls. Coefficient intervals are bootstrap 95% CIs for standardized drop with all controls. The MAGPIE raw $d = 0.916$ is the control-pipeline estimate; Table 5 reports a separately generated hierarchy-summary estimate ($d = 0.912$), differing by 0.004. Replacing inconsistent stored-drop rows with endpoint differences changes the DICE effect sizes by at most 0.024 (Table 31).

Dataset/model	Raw d	Resid. d	Matched d	Coef. CI	Pair acc.
MAGPIE	0.916	0.143	0.250	[0.429,1.162]	0.689
DICE GPT-2-Large	0.475	0.383	0.465	[0.651,1.008]	0.713
DICE GPT-2-XL	0.506	0.376	0.504	[0.633,0.979]	0.713
DICE Llama-3.1-8B	0.553	0.366	0.524	[0.587,0.931]	0.715
DICE Llama-3.1-70B	0.540	0.348	0.497	[0.580,0.871]	0.716
DICE Qwen2.5-14B	0.478	0.262	0.431	[0.433,0.807]	0.667
DICE Qwen3-32B	0.418	0.190	0.392	[0.269,0.609]	0.654
DICE Qwen3.5-122B-A10B	0.526	0.334	0.475	[0.621,0.944]	0.698

A.6 Secondary Predictive Validation Details

Table 11 reports the compact trajectory-classifier comparison summarized in Section 5.4. These numbers validate that trajectory features carry information beyond the raw endpoint difference, but they are not presented as state-of-the-art idiom detection.

Table 11. Secondary predictive validation. DICE uses the historical observed-surface MWE grouping; canonical-type separation is established only for the new nested probes. MAGPIE uses held-out test evaluation. Bold marks the best value within each dataset row. For DICE, a 10-seed grouped-CV robustness run gives GB-enhanced macro-F1 0.663 ± 0.006 vs. raw 0.604 ± 0.003 and AUROC 0.720 ± 0.004 vs. 0.652; GB-enhanced is higher in all 10 seed pairs (two-sided paired sign test, $p = 0.002$). MAGPIE uses a fixed train/dev/test split, so no seed-level marker is shown.

Dataset	Best trajectory model	Raw F1	Best F1	ΔF1	Raw AUROC	Best AUROC
DICE grouped	GB, enhanced features	0.605	**0.666**	**+0.061**	0.652	**0.720**
MAGPIE held-out	GB, base features	0.664	**0.686**	**+0.022**	0.756	**0.805**

A.7 GROUPED DICE FOLLOW-UPS

Table 12 provides the full model-by-model summary of grouped STRATIFIEDGROUPKFOLD ($k = 5$) evaluation on DICE using the historical observed MWE field; this is distinct from the 402-type canonical grouping of the new probes. Numbers for the base LR/GB trajectory rows correspond to Table 11. Agreement and semantic-gate rows are additional follow-ups not shown in the main paper.

Table 12. Grouped DICE follow-up results (Llama-3.1-8B unless noted). Rows marked † are Qwen-family backbone results.

Method	Macro-F1	AUROC
Raw drop threshold	0.605	0.652
LR, base features	0.608	0.655
LR, enhanced features	0.630	0.673
GB, base features	0.629	0.682
GB, enhanced features	**0.666**	**0.720**
Cross-model agreement (Llama+Qwen)	0.632	–
Model stacking	0.623	–
One-sided semantic gate	0.634	–
Qwen3-32B trajectory model†	0.609	–
Qwen3.5-122B-A10B trajectory model†	0.611	–
GPT-2-XL + soft semantic	0.613	–

These are frozen historical comparisons, including the old duplicate feature. The deduplicated rerun is reported in Appendix B.7. Without paired uncertainty for every auxiliary method, small differences in this table are descriptive.

A.8 FAILURE ANALYSIS DETAILS

The threshold diagnostic has difficult cases among lexicalized noun compounds such as *night owl* and *loan shark*, rare idiomatic expressions whose drop pattern may not have formed during pretraining, ascending-surprisal literal spikes, and single-word figurative uses, which fall outside the phrase-level diagnostic. **Ascending surprisal (literal spikes).** The 17% of literal expressions with $\Delta_{\text{drop}} < -2$ bits are not necessarily classification errors: they indicate contextually surprising literal continuations, plausibly arising from low co-occurrence between the words. Negative literal drops are predicted as literal by a positive-drop threshold; their prevalence alone is not evidence of classification failure. On MAGPIE literal uses, raw drop is positively correlated with local adjacent-PMI estimates computed from MAGPIE+DICE counts ($N = 3{,}497$; Pearson $r = 0.176$, $p = 1.3{\times}10^{-25}$; Spearman $\rho = 0.115$, $p = 8.2{\times}10^{-12}$), matching the expectation that lower co-occurrence corresponds to lower, more negative drops. **Short spans.** For $n = 1$, endpoint drop is identically zero and uninformative; for $n = 2$, only one transition is observed, making the statistic sensitive to either endpoint. Individual token surprisal has no vocabulary-entropy ceiling. This motivates the length-stratified analysis in Table 6, where the strongest separation is at $n = 3$ ($d = 0.945$). Low-margin MAGPIE predictions are 2.46× more error-dense than the full test set and are dominated by idioms with weak collapse, including *from scratch*, *off the hook*, and *by heart*. DICE errors instead reflect literal/idiomatic inversions of repeated surface forms, explaining why expression-centered calibration helps there.

Using a conservative inversion criterion–a mistaken literal must exceed the idiom mean drop for the same MWE, or a mistaken idiom must fall below the literal mean–310/803 DICE raw-drop errors (38.6%; 46.5% of comparable same-MWE errors) are repeated-surface-form inversions. The matching MAGPIE fraction is 96/415 (23.1%; 50.3% among comparable errors), because only 89 MAGPIE expressions in the held-out split contain both labels.

A.9 AUXILIARY CLASSIFIER DETAILS

The supervised baseline fine-tunes RoBERTa-Large (Liu et al., 2019) on SemEval-2022 idiom-detection training data, following the standard task setup (Tayyar Madabushi et al., 2022). The model takes [CLS, sentence, SEP] as input and predicts idiomatic (1) or literal (0). The hybrid system combines supervised encoder scores with surprisal-derived trajectory features through logistic regression on aligned SemEval development examples. The surprisal side includes span surprisal, context modulation, slope, and drop; the encoder side contributes its logit and probability. These

systems are documented for completeness and are not load-bearing evidence for the mechanistic claim.

A.10 Cross-Lingual PARSEME Audit

We treat cross-lingual transfer as out of scope for the present mechanistic claim. Cross-lingual MWE detection introduces annotation-guideline differences across PARSEME languages, tokenization mismatches, and translation artifacts that would require separate study to disentangle from the trajectory signal itself. As an indicative audit using cached XLM-R span features for positive English/Portuguese VMWEs, EN→PT reaches macro-F1 $= 0.253$ and PT→EN reaches 0.355, with similarly weak monolingual cross-validation; because this setup classifies VMWE types among positive spans only, it is not directly comparable to MAGPIE/DICE idiom-vs-literal detection. We therefore restrict claims in this paper to English.

A.11 Additional Scope Notes

Our evidence is strongest for English idioms and related phrase-level fixed expressions; single-token figurative uses and sentence-level pragmatic phenomena fall outside the span-local diagnostic. Threshold calibration also remains dataset-dependent: MAGPIE dev thresholds transfer within MAGPIE but not to DICE, so deployment requires either target-domain calibration or label-free alternatives such as the z-score heuristic.

The three-model DICE specificity tests and the later 36-checkpoint pilot have different samples and outcomes. The pilot does not extend the 60-type DICE causal estimate to every checkpoint. See Section 7.

A.12 Broader Impacts

This work is a diagnostic analysis of pretrained autoregressive language models. The later pilot adds 80 sentences drafted with AI assistance and accepted after researcher batch verification; it does not introduce a deployed decision system. Potential positive uses include improved interpretability of phrase-level linguistic knowledge in language models and clearer evaluation of figurative-language behavior. Potential risks are indirect: idiomaticity diagnostics could be incorporated into generated-text or classification systems without adequate attention to language, dialect, or domain differences. We frame the method as an analysis tool rather than a deployment-ready detector.

A.13 Licenses for Existing Assets

Datasets. MAGPIE (Haagsma et al., 2020) is distributed under CC-BY 4.0. DICE is attributed to Mi et al. (2025); its source is the repository accompanying that publication. Use and redistribution remain subject to the source dataset terms. SemEval-2022 Task 2 data (Tayyar Madabushi et al., 2022) is distributed under the original SemEval task license. VUA20 (CreativeLang, 2024) is distributed via the Hugging Face Datasets hub under the dataset card's stated license. TweetEval (Barbieri et al., 2020) is distributed under CC-BY 4.0; the underlying SemEval-2018 Task 3 corpus (Van Hee et al., 2018) retains its original task license.

Models. GPT-2 family checkpoints (Radford et al., 2019) are distributed under the modified MIT license on the original release. Llama-3.1 and Llama-3.2 checkpoints (Grattafiori et al., 2024) are distributed under the Llama 3.1/3.2 Community License. Qwen2.5 (Yang et al., 2024), Qwen3 (Yang et al., 2025), and Qwen3.5 (Qwen Team, 2026) checkpoints are used under the license stated in the corresponding model card for each released checkpoint. Mistral-7B (Jiang et al., 2023) is distributed under Apache 2.0. RoBERTa-Large (Liu et al., 2019) and DeBERTa-v3-Large (He et al., 2021) are distributed under MIT.

Software. We use Hugging Face Transformers (Wolf et al., 2020) (Apache 2.0), PyTorch (BSD-style), scikit-learn (BSD), and NumPy (BSD). All software is used in accordance with its license.

A.14 REPRODUCIBILITY

The SemEval RoBERTa-Large baseline uses learning rate 2×10^{-5}, 3 epochs, batch size 16/32 (train/eval), weight decay 0.01, fp16, and maximum sequence length 512. GPT-2-Large is the default surprisal profiler ($T_{\text{max}} = 512$, batch size 8, bfloat16 on CUDA). For the MAGPIE threshold baseline, the threshold is selected on the development split and applied unchanged to the test split; DICE threshold sweeps are descriptive, and nested probe selection and follow-up analyses are specified separately. An anonymized numerical supplement accompanies this submission (Appendix B.8).

Archived Slurm logs for the large-model runs record 4× NVIDIA GH200 120GB GPUs, NVIDIA driver 565.57.01, CUDA 12.7, conda environment `py10`, and Python 3.10.19. The saved code specifies `torch>=2.0.0`, `transformers>=4.30.0`, `scikit-learn>=1.3.0`, and `numpy>=1.24.0`.

Classifier experiments use random state 42 for the fixed MAGPIE split and DICE grouped CV; the DICE robustness run uses seeds {1, 2, 3, 4, 5, 13, 21, 34, 42, 55}. Collocation bootstrap/control analyses use seed 42.

B Controlled Experiments and Subsequent Extensions

B.1 Matched Causal Protocol and Specificity Controls

The unit of inference is normalized canonical DICE expression type, not the observed inflected string. Five grouped folds use random state 42. Within the test fold, an eligible exact-surface/token-signature group has at least two idiomatic and two literal contexts, allowing distinct-context donors under both label relations. Targets and donors must differ. The retained population is 318 contexts/60 types for each of the three models. Mean replacement uses label-blind training-fold activations matched by hook position and expression word length, with a position-specific fallback. The hook is the decoder-block output after attention and MLP residual updates; layer indices are zero-based.

The norm-matched-noise control replaces the state by an isotropic vector with the clean residual's norm. Random-expression donors have the recipient label but a different canonical expression, selected deterministically using word length, subtoken length, sentence-length bin, baseline-surprisal bin, and sentence token length. Thus the noise control, random-expression replacement, and same-expression replacement perturb different properties. Every intervention recomputes the forward pass with `use_cache=False`. Final-word onset, final span subtoken, and summed final-word surprisal are stored in negative-log2 bits; independent clean-endpoint checks agree exactly.

Table 13. Final-word-onset effects. Idiom and literal columns average context-level surprisal changes within canonical type, then weight types equally. DiD is their difference; 60 types/318 contexts per model. Noise and replacement controls are different interventions, not interchangeable null models.

Model	Intervention	δ_I	δ_L	DiD [95% CI]
GPT-2-Large	Cross-label	0.78	-1.21	1.99 $[0.26, 3.61]$
GPT-2-Large	Same-label	-0.34	-0.19	-0.15 $[-0.45, 0.16]$
GPT-2-Large	Random-expression	14.16	12.07	2.09 $[0.68, 3.50]$
GPT-2-Large	Train-mean	7.87	6.52	1.35 $[0.66, 2.01]$
GPT-2-Large	Norm-matched noise	14.75	13.70	1.05 $[0.29, 1.79]$
GPT-2-Large	Zero	16.34	15.36	0.98 $[-0.15, 2.08]$
Llama-3.1-8B	Cross-label	1.32	-1.92	3.24 $[2.04, 4.50]$
Llama-3.1-8B	Same-label	-0.21	-0.34	0.13 $[-0.17, 0.44]$
Llama-3.1-8B	Random-expression	14.75	12.09	2.66 $[1.32, 4.07]$
Llama-3.1-8B	Train-mean	6.84	5.13	1.71 $[1.07, 2.36]$
Llama-3.1-8B	Norm-matched noise	16.56	14.60	1.96 $[1.22, 2.73]$
Llama-3.1-8B	Zero	15.88	13.87	2.00 $[1.11, 2.90]$
Qwen2.5-14B	Cross-label	0.98	-2.32	3.30 $[1.93, 4.76]$
Qwen2.5-14B	Same-label	-0.55	-0.73	0.18 $[-0.15, 0.50]$
Qwen2.5-14B	Random-expression	13.05	10.80	2.25 $[0.73, 3.90]$
Qwen2.5-14B	Train-mean	5.65	3.79	1.86 $[1.16, 2.59]$
Qwen2.5-14B	Norm-matched noise	15.59	13.81	1.78 $[0.95, 2.61]$
Qwen2.5-14B	Zero	15.20	13.22	1.98 $[0.93, 3.03]$

Table 14. Cross-label-minus-norm-matched-noise DiD at final-word onset (bits). GPT-2-Large retains an interval spanning zero. These values are traced to the stored control-contrast artifact.

Model	Contrast [95% CI]
GPT-2-Large	0.94 $[-0.08, 1.92]$
Llama-3.1-8B	1.28 $[0.47, 2.12]$
Qwen2.5-14B	1.52 $[0.60, 2.50]$

Table 15. Sensitivity of the cross-label causal DiD to alternative continuation outcomes, in bits (60 types per model). The GPT-2-Large final-subtoken interval includes zero.

Model	Outcome	Cross-label DiD [95% CI]
GPT-2-Large	Final span subtoken	$1.53\ [-0.05, 3.03]$
GPT-2-Large	Final-word sum	$1.99\ [0.29, 3.61]$
Llama-3.1-8B	Final span subtoken	$2.86\ [1.72, 4.11]$
Llama-3.1-8B	Final-word sum	$3.24\ [2.04, 4.53]$
Qwen2.5-14B	Final span subtoken	$2.92\ [1.61, 4.29]$
Qwen2.5-14B	Final-word sum	$3.30\ [1.95, 4.79]$

Table 13 reports both recipient labels and their DiD. Cross-minus-same is the primary label-contingent comparison in Table 2. Cross-minus-random is $-0.10\ [-1.42, 1.20]$, $0.57\ [-0.42, 1.60]$, and $1.04\ [-0.28, 2.34]$ bits for GPT-2-Large, Llama, and Qwen. These null contrasts constrain causal specificity. The noise contrast is supported by the stored causal-control artifact (Table 14); it is not inferred from mean-ablation values. Alternative outcomes retain their own uncertainty (Table 15).

Recipient decomposition and donor audit. Recomputing from the saved initial causal rows reproduces all 216 checked point estimates and intervals to floating-point tolerance. Table 16 separates the two recipient readings. Contexts are averaged within type and reading, types are weighted equally, and all quantities share 10,000 type-resampling draws with recorded seeds, so $C = C_I - C_L$ holds in every draw. Historical contrast intervals retain their original seeds; the new joint-resampling intervals can differ slightly without changing the estimand. Neither the sign nor the magnitude directly measures semantic transfer: the outcome scores the observed final-word-onset token. Table 17 reports absolute donor–recipient gaps for predictor position, sentence length, and clean onset surprisal. The paired cross-minus-same mismatch contrasts are all positive, so exact expression and token identity does not fix these contextual properties.

Table 16. Recipient-specific final-word-onset effects in bits on the same 318 contexts/60 types per model. Intervals use joint expression-type resampling. Positive values worsen prediction; negative values improve it. C_I and C_L subtract same-reading from cross-reading effects within each recipient reading, and $C = C_I - C_L$. The unadjusted GPT-2-Large idiomatic cross-reading interval includes zero, while its paired C_I interval does not.

Condition	Idiomatic recipients [95% CI]	Literal recipients [95% CI]
GPT-2-Large		
Cross-reading	$0.78\ [-0.02, 1.57]$	$-1.21\ [-2.09, -0.28]$
Same-reading	$-0.34\ [-0.56, -0.12]$	$-0.19\ [-0.38, 0.01]$
Random-expression	$14.16\ [12.72, 15.54]$	$12.07\ [10.60, 13.58]$
Cross minus same (C_I, C_L)	$1.11\ [0.35, 1.86]$	$-1.02\ [-1.87, -0.11]$
Llama-3.1-8B		
Cross-reading	$1.32\ [0.67, 2.02]$	$-1.92\ [-2.55, -1.32]$
Same-reading	$-0.21\ [-0.38, -0.04]$	$-0.34\ [-0.57, -0.13]$
Random-expression	$14.75\ [13.42, 16.09]$	$12.09\ [10.95, 13.20]$
Cross minus same (C_I, C_L)	$1.53\ [0.92, 2.18]$	$-1.58\ [-2.17, -1.02]$
Qwen2.5-14B		
Cross-reading	$0.98\ [0.27, 1.74]$	$-2.32\ [-3.08, -1.60]$
Same-reading	$-0.55\ [-0.76, -0.35]$	$-0.73\ [-1.01, -0.47]$
Random-expression	$13.05\ [11.67, 14.45]$	$10.80\ [9.51, 12.14]$
Cross minus same (C_I, C_L)	$1.52\ [0.83, 2.26]$	$-1.59\ [-2.29, -0.93]$

Table 17. Cross-minus-same differences in mean absolute donor–recipient mismatch. Each mismatch is averaged within recipient and then canonical type. Positive values indicate greater mismatch for cross-reading donors. Intervals jointly resample types; these are descriptive donor audits.

Model	Position (tokens)	Sentence (tokens)	Clean onset (bits)
GPT-2-Large	1.96 [0.75, 3.31]	5.35 [3.24, 7.65]	0.73 [0.32, 1.17]
Llama-3.1-8B	1.91 [0.71, 3.25]	5.15 [3.10, 7.37]	0.76 [0.28, 1.26]
Qwen2.5-14B	1.97 [0.72, 3.32]	5.34 [3.23, 7.73]	0.85 [0.32, 1.44]

Position sweeps and repeated estimates. Train-mean and matched-donor sweeps are analyzed separately (Tables 18 and 19). A post-MWE intervention cannot alter an earlier next-token prediction under causal masking; the exact zero is an implementation check. It is not evidence that the final predictor is uniquely responsible among upstream positions. The matched sweep reuses the same final-predictor contrast as the primary analysis. We report its canonical primary confidence interval consistently, rather than presenting bootstrap seed variation as independent confirmation. GPT-2-Large's middle-MWE interval includes zero, and some pre-MWE matched effects are negative.

Nominal positions are not uniformly upstream of the predictor. In every model, the first span token coincides with the final-word predictor in 52 contexts/10 types; the middle token coincides with it in 227 contexts/43 types, follows it in 69 contexts/13 types, and strictly precedes it in only 22 contexts/4 types. A position after the predictor cannot affect the onset prediction. Restricting the mean-ablation sweep to strictly preceding first-token positions (50 types/266 contexts), the predictor-minus-first contrast remains positive: 1.06 [0.33, 1.74], 1.25 [0.57, 1.93], and 1.20 [0.50, 1.93] bits for GPT-2-Large, Llama, and Qwen. The strictly preceding middle-position subset (4 types/22 contexts) is inconclusive. The matched-donor position rows are not in the numerical supplement, so the matched sweep could not be restricted in the same way.

Table 18. Train-mean ablation position sweep: idiom-minus-literal DiD [95% CI], bits, at each model's fixed layer (60 types). The final-predictor row reuses the condition estimate and interval in Table 13. The post-MWE effect is exactly zero by causal masking. Nominal first/middle positions can coincide with or follow the predictor (Appendix B.1).

Position	GPT-2-Large	Llama-3.1-8B	Qwen2.5-14B
Earlier context	0.00 [−0.11, 0.11]	0.03 [−0.19, 0.20]	−0.05 [−0.24, 0.16]
Pre-MWE	0.02 [−0.08, 0.12]	0.01 [−0.01, 0.04]	−0.01 [−0.21, 0.19]
First MWE token	0.46 [0.08, 0.89]	0.66 [0.24, 1.14]	0.84 [0.31, 1.46]
Middle MWE token	0.67 [0.10, 1.23]	0.88 [0.37, 1.40]	1.09 [0.63, 1.59]
Final-word predictor	1.35 [0.66, 2.01]	1.71 [1.07, 2.36]	1.86 [1.16, 2.59]
Post-MWE sanity	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]

Table 19. Matched cross-minus-same-label position contrasts [95% CI], bits (60 types per model). The final-predictor row reuses the canonical interval from Table 2; it is the same estimate, not an independent replication. Earlier and middle-position null/negative effects are retained. Nominal first/middle positions can coincide with or follow the predictor (Appendix B.1).

Position	GPT-2-Large	Llama-3.1-8B	Qwen2.5-14B
Earlier context	−0.11 [−0.27, 0.01]	−0.04 [−0.11, 0.02]	−0.19 [−0.32, −0.07]
Pre-MWE	−0.07 [−0.20, 0.01]	−0.07 [−0.15, −0.01]	−0.21 [−0.36, −0.08]
First MWE token	0.82 [0.10, 1.68]	0.91 [0.22, 1.72]	0.99 [0.10, 2.07]
Middle MWE token	0.89 [−0.42, 2.09]	1.83 [0.98, 2.82]	1.74 [0.85, 2.72]
Final-word predictor	2.13 [0.59, 3.63]	3.11 [1.99, 4.24]	3.12 [1.88, 4.45]

Predictability-conditioned sensitivity. Within exact surface/token groups, idiomatic and literal recipients are matched without replacement with clean onset-surprisal differences at most one bit and predictor-position differences at most 32 tokens. Retained types/contexts are 45/138, 43/142, and 46/148 for GPT-2-Large, Llama, and Qwen. Cross-minus-same effects are 1.143 [−0.563, 2.821], 2.231 [0.967, 3.551], and 2.039 [0.753, 3.336] bits. Cross-minus-random effects

are -0.148 $[-2.110, 1.770]$, 0.762 $[-0.618, 2.100]$, and 0.780 $[-0.752, 2.318]$. Predictability can be a mediator, and conditioning reduces the sample; these are sensitivity estimates rather than an unconfounded mechanism estimate. A further type-fixed-effects model includes recipient baseline and length covariates plus donor–recipient baseline gaps. Its cross-minus-same estimates are 0.637 $[0.287, 1.036]$, 0.187 $[-0.075, 0.453]$, and 0.532 $[0.110, 0.928]$ bits; the Llama interval includes zero. This adjustment may condition on mediating predictability, so it is not an alternative identified causal effect.

B.2 Probing: Expression Identity and Context Information

The initial nested protocol groups all idiomatic and literal uses of each of 402 canonical expressions into one outer fold. Each eligible example receives one OOF prediction per representation. Logistic-regression regularization and layer are selected using grouped inner folds; scalers are fitted only on the corresponding training partitions. The integrity audit checks fold separation, feature row alignment, duplicate sentences, and absence of bracketed idiom/literal label markers.

Surface-only features and neutral-template hidden states quantify measured lexical-form alternatives. The latter use one extraction per observed surface (573 rows for GPT-2-Large/Qwen and 590 for Llama), mapped to the aligned OOF population. They do not have the same extraction count as contextual features. Table 20 reports evaluated examples/types. Random sentence-level splits and expression-majority controls are historical comparisons; they do not replace expression-disjoint testing. Random-minus-grouped intervals include zero, which does not exclude other shortcuts.

Table 20. Initial nested-probe audit. OOF AUROC and expression-bootstrap intervals; sample counts are examples/types in the evaluated population. Context-free features are extracted once per observed surface and mapped back to this population. Random sentence-level results are auxiliary.

Model	Evaluation	N	AUROC [95% CI]
GPT-2-Large	Grouped span mean	2046/402	0.981 $[0.975, 0.987]$
GPT-2-Large	Grouped predictor	2046/402	0.965 $[0.957, 0.973]$
GPT-2-Large	Grouped first token	2046/402	0.979 $[0.974, 0.984]$
GPT-2-Large	Surface-only	2046/402	0.537 $[0.523, 0.553]$
GPT-2-Large	Context-free	2046/402	0.553 $[0.538, 0.569]$
GPT-2-Large	Random-split span mean	2046/402	0.984 $[0.979, 0.989]$
Llama-3.1-8B	Grouped span mean	2066/402	0.991 $[0.988, 0.994]$
Llama-3.1-8B	Grouped predictor	2066/402	0.986 $[0.982, 0.990]$
Llama-3.1-8B	Grouped first token	2066/402	0.986 $[0.982, 0.990]$
Llama-3.1-8B	Surface-only	2066/402	0.543 $[0.526, 0.560]$
Llama-3.1-8B	Context-free	2066/402	0.564 $[0.548, 0.581]$
Llama-3.1-8B	Random-split span mean	2066/402	0.992 $[0.989, 0.995]$
Qwen2.5-14B	Grouped span mean	2046/402	0.991 $[0.987, 0.994]$
Qwen2.5-14B	Grouped predictor	2046/402	0.988 $[0.984, 0.992]$
Qwen2.5-14B	Grouped first token	2046/402	0.989 $[0.985, 0.992]$
Qwen2.5-14B	Surface-only	2046/402	0.535 $[0.519, 0.551]$
Qwen2.5-14B	Context-free	2046/402	0.563 $[0.546, 0.581]$
Qwen2.5-14B	Random-split span mean	2046/402	0.992 $[0.988, 0.995]$

Subsequent context-only protocol. This follow-up was designed after examining the initial probe results; it is not a previously untouched confirmatory dataset. New representations exclude the final lexical word and right context. The pre-MWE control ends before the expression. Neutral-prefix controls replace each visible MWE-prefix token by a repeated token for `something` or `thing`, preserving the pre-expression IDs, token count, and positions. They are artificial corruptions, not grammatical paraphrases. Original predictor, first-token, and span-mean features are reused from the initial cache; span mean sees the final word and is only an auxiliary comparison.

The common population is 2,032 contexts/402 types per model. Llama excludes 34 further rows from 2,066 aligned examples, and Qwen 14 from 2,046, for missing textual left context or an empty visible prefix. Layer search uses nine predeclared depths and $C \in \{0.01, 0.1, 1\}$, with five outer and three

inner grouped folds. All final L2/liblinear fits use the same primal objective with explicit convergence checks; two initial iteration-limited Llama fits were superseded without re-extracting GPU features. Training-label-permuted fixed-layer sanity AUROCs are 0.473 and 0.496.

Table 21. Full six-representation context-control comparison. Each model uses the same 2,032 contexts/402 types for all representations. Span mean is not an onset-only representation.

Representation	Llama-3.1-8B	Qwen2.5-14B
Span mean (sees final word)	$0.990\ [0.986, 0.994]$	$0.993\ [0.990, 0.995]$
First MWE token	$0.986\ [0.981, 0.990]$	$0.987\ [0.982, 0.991]$
Intact predictor	$0.984\ [0.979, 0.989]$	$0.989\ [0.984, 0.992]$
Pre-MWE context	$0.989\ [0.986, 0.993]$	$0.989\ [0.985, 0.993]$
Neutral prefix: something	$0.985\ [0.980, 0.990]$	$0.991\ [0.987, 0.994]$
Neutral prefix: thing	$0.986\ [0.981, 0.991]$	$0.987\ [0.982, 0.991]$

Table 22. Paired AUROC differences: intact predictor minus each control. Each interval resamples the same expression clusters jointly. All original-minus-neutral intervals include zero.

Control	Llama-3.1-8B	Qwen2.5-14B
Neutral: something	$-0.001\ [-0.006, 0.004]$	$-0.002\ [-0.006, 0.001]$
Neutral: thing	$-0.002\ [-0.007, 0.002]$	$0.001\ [-0.003, 0.005]$
Pre-MWE	$-0.005\ [-0.009, -0.001]$	$-0.001\ [-0.005, 0.003]$

The paired original-minus-neutral effects in Table 22 are the direct incremental comparisons. Near-ceiling absolute AUROC alone does not show that the MWE prefix creates the information. As an explicitly exploratory diagnostic added after the high context-only result, prefix-only TF-IDF reaches AUROC $0.927\ [0.914, 0.939]$ on both model populations; six numeric prefix statistics (log-transformed token, character, and word counts, mean word length, and digit and punctuation fractions of the pre-expression text) reach $0.771\ [0.750, 0.792]$ and $0.769\ [0.747, 0.789]$. Their transformations and regularization are still fitted inside training folds. These comparisons demonstrate substantial information in pre-expression text, not an absence of lexicalized prediction. In the causal subset, idiomatic prefixes and sentences are also longer than literal ones (Table 23); these univariate AUROCs are pooled descriptive ranks, not cross-validated classification.

Table 23. Descriptive statistics on the 318-context causal subset: reading-specific mean (SD), pooled standardized mean difference (idiomatic minus literal), and univariate AUROC with larger values predicting idiomatic use. These pooled AUROCs are descriptive ranks, not held-out classifier performance or results for the larger context-control population.

Feature	Idiomatic	Literal	SMD	AUROC
GPT-2-Large				
Prefix tokens	15.66 (13.71)	8.53 (3.94)	0.718	0.630
Prefix chars.	73.22 (63.28)	42.42 (19.61)	0.668	0.611
Predictor position	16.99 (13.74)	9.81 (4.29)	0.717	0.637
Sentence tokens	30.39 (16.27)	17.88 (4.72)	1.061	0.736
Expression subtokens	3.42 (1.11)	3.36 (1.08)	0.056	0.515
Llama-3.1-8B				
Prefix tokens	16.55 (13.61)	9.53 (3.95)	0.712	0.628
Prefix chars.	73.22 (63.28)	42.42 (19.61)	0.668	0.611
Predictor position	17.86 (13.63)	10.79 (4.29)	0.712	0.635
Sentence tokens	31.16 (16.02)	18.84 (4.60)	1.063	0.735
Expression subtokens	3.41 (1.11)	3.35 (1.07)	0.055	0.514
Qwen2.5-14B				
Prefix tokens	15.69 (13.75)	8.53 (3.95)	0.720	0.629
Prefix chars.	73.22 (63.28)	42.42 (19.61)	0.668	0.611
Predictor position	17.00 (13.75)	9.79 (4.29)	0.719	0.637
Sentence tokens	30.44 (16.24)	17.84 (4.60)	1.074	0.738
Expression subtokens	3.41 (1.11)	3.35 (1.07)	0.055	0.514

B.3 Span Alignment and Tokenization Endpoints

Alignment uses the annotated in-context lexical-word form, including inflection, before mapping character/word spans to model subtokens. Punctuation outside the annotation is excluded. A contiguous aligned span is required; the result does not quantify unrestricted discontinuous or non-canonical realizations. The four drop definitions subtract final from first endpoint surprisal, using (i) first/final span subtokens, (ii) first subtokens of the endpoint words, (iii) sums over endpoint-word subtokens, or (iv) their means. Both endpoint words must be single subtokens for the no-fragmentation subset.

GPT-2-Large and Qwen align 2,046 examples across 402 types, excluding exactly 20 sentence-initial examples (11 idiomatic, nine literal; 19 types) with no preceding predictor. Llama aligns all 2,066 examples/402 types. Requiring both labels within type for the paired endpoint estimand leaves 2,043/399 for GPT-2-Large and Qwen, and 2,066/402 for Llama. The no-fragmentation aligned counts are 1,818/363, 1,844/365, and 1,826/365; the corresponding paired-eligible counts are smaller (Table 25).

Table 24. All aligned endpoint variants. Counts are paired-eligible examples/types after within-type label requirements. The bit contrast averages contexts within label/type; $d_z = \mathrm{mean}(d_e)/\,\mathrm{SD}(d_e)$ for type-paired differences d_e, rather than pooled Cohen's d.

Model	Endpoint	N	Difference bits [95% CI]	d_z [95% CI]
GPT-2-Large	Span subtoken	2043/399	3.10 [2.58, 3.61]	0.602 [0.495, 0.713]
GPT-2-Large	Word onset	2043/399	3.11 [2.61, 3.62]	0.612 [0.510, 0.721]
GPT-2-Large	Word sum	2043/399	3.20 [2.69, 3.71]	0.618 [0.514, 0.725]
GPT-2-Large	Word mean	2043/399	3.00 [2.51, 3.49]	0.596 [0.493, 0.704]
Llama-3.1-8B	Span subtoken	2066/402	3.53 [3.04, 4.03]	0.697 [0.601, 0.799]
Llama-3.1-8B	Word onset	2066/402	3.63 [3.14, 4.12]	0.710 [0.615, 0.810]
Llama-3.1-8B	Word sum	2066/402	3.70 [3.19, 4.20]	0.720 [0.624, 0.823]
Llama-3.1-8B	Word mean	2066/402	3.51 [3.04, 4.01]	0.703 [0.606, 0.803]
Qwen2.5-14B	Span subtoken	2043/399	3.35 [2.78, 3.92]	0.578 [0.472, 0.690]
Qwen2.5-14B	Word onset	2043/399	3.42 [2.84, 4.00]	0.583 [0.478, 0.691]
Qwen2.5-14B	Word sum	2043/399	3.44 [2.86, 4.04]	0.582 [0.476, 0.694]
Qwen2.5-14B	Word mean	2043/399	3.29 [2.74, 3.85]	0.579 [0.470, 0.693]

Table 25. Word-onset diagnostic when both endpoint words have exactly one subtoken. Paired-eligible sample counts; intervals resample canonical types.

Model	N	Difference bits [95% CI]	d_z [95% CI]
GPT-2-Large	1805/356	3.15 [2.61, 3.69]	0.609 [0.498, 0.726]
Llama-3.1-8B	1829/359	3.67 [3.17, 4.18]	0.736 [0.634, 0.846]
Qwen2.5-14B	1810/357	3.43 [2.86, 4.02]	0.612 [0.494, 0.737]

We first average contexts within expression and label, then form a paired type difference. The paired effect is $d_z = \bar{d}_e / \mathrm{SD}(d_e)$. It is a paired estimator, distinct from the pooled example-level Cohen's d in the behavioral tables. All intervals use 10,000 type resamples. The saved 144 fragmentation-stratum by endpoint rows retain label-blind model-specific bins and mark strata with fewer than 20 eligible types as descriptive. Thus the result establishes robustness to tested endpoints within aligned spans, not a tokenizer-independent atomic object or invariance to unrestricted morphological variation.

B.4 Held-Out Attention Selection

The primary attention statistic averages attention from the final-word predictor query to MWE subtokens at or before that predictor. A sensitivity statistic averages strict earlier-key causal pairs within the span. In each expression-grouped fold, five heads are selected on training types and averaged unchanged on held-out types. The paired d_z is computed from OOF type-level idiom-minus-literal differences. Separate all-head tests receive BH-FDR correction within model/statistic.

Table 26. Held-out top-five attention and all-head characterization. Positive/negative counts use BH-FDR $q < 0.05$ within model/statistic. Top-five d_z is type-paired; the median is the stored all-head pooled Cohen's d, a different statistic. Each composite is positive in 5/5 held-out folds.

Model	N	Heads	Top-five d_z [95% CI]	FDR + / –	Median d
GPT-2-Large	2023/397	720	0.672 [0.584, 0.771]	95/356	-0.090
Llama-3.1-8B	2049/401	1024	0.828 [0.737, 0.926]	184/522	-0.084
Qwen2.5-14B	2023/397	1920	1.063 [0.951, 1.191]	439/829	-0.057

The positive selected composite is consistent across all five folds, but the full head distributions include more negative than positive significant heads and negative medians. This is selective heterogeneous routing rather than a global idiom-attention increase or evidence for one uniquely storing head. Attention requires at least one prior available MWE key. GPT-2-Large and Qwen retain 2,032 valid examples before paired-type filtering and 2,023/397 after it; Llama retains 2,052 before and 2,049/401 after. These denominators differ from the tokenization and probing populations.

B.5 Head-Output Interventions

The follow-up selects attention candidates only on DICE folds 2–4 and freezes top-1, top-3, and top-5 sets before the fold-0 intervention (60 types/318 contexts). We patch per-query-head output slices immediately before the attention output projection and compare cross-label with same-label donors. Five random sets match each selected set's layer and head count; low-bias heads provide an additional comparison. Query heads are distinguished from shared GQA key/value groups. Each model retains 13,992 outcome records.

The predeclared primary onset contrast is top-five cross-minus-same DiD minus the mean of five random-set DiDs: Llama 0.083 [0.033, 0.139] bits and Qwen 0.082 [0.041, 0.129]. The top-1 and restoration outcomes in Table 27 prevent interpreting the joint effect as a sum of positive individual head effects.

Table 27. Head-output cross-minus-same DiD at onset, bits, on 60 types/318 contexts. These are pointwise conditional intervals given the frozen head selection. Auxiliary groups are not all multiplicity-adjusted; negative and null single-head effects are retained.

Head group	Llama-3.1-8B	Qwen2.5-14B
Top 1	-0.011 $[-0.020, -0.003]$	0.021 $[-0.001, 0.044]$
Top 3	0.018 $[0.001, 0.036]$	0.037 $[0.004, 0.072]$
Top 5	0.099 $[0.054, 0.148]$	0.092 $[0.051, 0.139]$
Low bias	-0.001 $[-0.010, 0.008]$	0.014 $[-0.022, 0.048]$
Top 5, restore top 1	0.105 $[0.059, 0.155]$	0.084 $[0.050, 0.123]$
Top 5, direct residual route	0.089 $[0.034, 0.158]$	0.051 $[0.022, 0.084]$

Restoring top 1 means restoring that head to its clean recipient value after corrupting the top-five set. Restoring all selected heads gives the self check. For the direct-residual-route test, other predictor-position head and MLP outputs are clamped to clean recipient values. The selected contributions reach onset prediction through the residual stream, final normalization, and unembedding. This is a restricted direct-route intervention, not complete head-to-head or head-to-MLP circuit discovery. All auxiliary head groups, restoration tests, and alternative word outcomes are retained in the saved results; their pointwise intervals are not jointly multiplicity-adjusted and condition on the frozen head selection.

B.6 New-Context Pilot and Exploratory Checkpoint Extension

The pilot contains 20 idiomatic/literal and 20 compositional/compositional expression pairs, each with two contexts (80 sentences). The items were drafted with AI assistance and accepted after the researcher's batch-level confirmation that all pairs and labels were verified unchanged. Item-level ratings and inter-annotator agreement were not recorded. The contexts may differ in multiple places, so these are same-expression context pairs, not strict single-edit minimal pairs. Frozen DICE head sets are used. Seven of the 40 normalized expressions/surfaces overlap the head-selection training rows in the two-model analysis, with possible further inflectional overlap; this is transfer to new contexts, not expression-disjoint generalization.

Pilot estimand. Let $A_e(S)$ be the mean absolute change in final-word-onset surprisal across the two donor directions when head set S is replaced. For five matched random sets R_j, define

$$H_e = A_e(S_{\text{top5}}) - \tfrac{1}{5}\sum_{j=1}^{5} A_e(R_j), \qquad G = \text{mean}_{e\in I}\, H_e - \text{mean}_{e\in C}\, H_e.$$

Here I contains idiomatic/literal pairs and C compositional/compositional pairs. This absolute-effect group interaction is distinct from the signed DICE cross-minus-same DiD. The pilot has only one context per reading, so self-patching checks implementation and does not substitute for an independent same-reading donor.

Llama has $G = 0.068$ $[0.025, 0.114]$ bits and Qwen 0.040 $[0.015, 0.068]$, with Holm-adjusted approximate centered-bootstrap $p = 0.0066$ for each of the two initial model tests. Within-idiom selected-minus-random estimates are 0.040 $[-0.000, 0.084]$ and 0.017 $[-0.003, 0.039]$, both including zero. Compositional-group estimates are -0.028 $[-0.043, -0.014]$ and -0.023 $[-0.040, -0.008]$. Thus the interaction reflects relative group selectivity, partly driven by reduced selected-head sensitivity in the compositional group; it does not establish a significant selected-head advantage within idioms.

Matching and annotation limits. Matching on pair-mean clean onset surprisal, predictor position, and span subtoken count uses calipers of one bit, 32 tokens, and one token. Each model retains 16 idiom/control pair-pairs. Matched G is 0.077 $[0.024, 0.131]$ for Llama and 0.030 $[0.009, 0.056]$ for Qwen. Post-match standardized mean differences for surprisal/position/span length are $-0.100/0.693/1.061$ and $-0.030/0.339/0.923$. Caliper eligibility therefore does not imply balance. Matching is a sensitivity analysis on the same purposive sample.

Checkpoint extension. After inspecting the first two models, we evaluated 34 additional cached checkpoints on the same 40 pairs. This is exploratory repeated model evaluation on one linguistic sample. Residual replacement uses the post-block predictor state at zero-based layer $\lfloor 0.75L \rfloor - 1$, with the other same-expression, same-token-sequence context as donor. Three Gaussian controls match the norm of the recipient-to-donor displacement. The residual outcome subtracts their mean absolute surprisal effect before comparing idiom and compositional groups. This noise definition differs from the clean-state-norm replacement in Table 13.

Eight additional checkpoints with inspected compatible attention implementations also receive the full head analysis. Each selects its own heads on DICE folds 2–4 and freezes them before pilot evaluation. Tables 28 and 29 retain all successful checkpoints, including negative and null effects. Ten thousand pair-cluster resamples yield pointwise intervals; approximate centered-bootstrap p values receive separate Holm corrections over 34 new residual tests and eight new head tests. Corrected positive counts are 0/34 and 4/8. The tables also report the sensitivity correction including the earlier models. These counts are not prevalence estimates over a model population, and a positive interaction does not by itself imply a positive within-idiom selected-minus-random effect.

All checkpoints score plain sentences without chat templates; multimodal checkpoints use their text forward route. Quantized checkpoints and library compatibility changes are recorded in per-model metadata, rather than treated as native BF16 replications. Exact self-replacement, sentence completeness, pair construction, and token signatures pass for successful runs. No unresolved failure or timeout remains among attempted checkpoints; incomplete caches, encoder-only models, and weights beyond the single-GPU resident budget are excluded by the saved inventory rather than by effect direction. Differences in context change, predictability, architecture, training, numerical precision, and the fraction of heads represented by a fixed top-five set prevent a causal model-size interpretation.

Table 28. Residual-minus-noise group contrast on the same 20 idiom/literal and 20 compositional context pairs, in bits. Intervals are pointwise 95%; Holm columns adjust approximate centered-bootstrap p within intervention family, over the new checkpoints or all successful checkpoints including the two earlier models. Missing head tests are not zero effects.

Checkpoint	Contrast [95% CI]	Holm new	Holm all
Qwen/Qwen2.5-0.5B	1.005 $[-0.067, 2.168]$	1.0000	1.0000
Qwen/Qwen2.5-1.5B	1.315 $[0.138, 2.677]$	0.8210	0.8601
Qwen/Qwen2.5-7B	1.570 $[0.499, 2.795]$	0.2376	0.2464
allenai/Olmo-3-1025-7B	0.865 $[-0.282, 2.181]$	1.0000	1.0000
NousResearch/Meta-Llama-3.1-8B-Instruct	1.386 $[0.204, 2.919]$	0.8210	0.8601
gpt2	0.908 $[0.284, 1.577]$	0.1827	0.1890
Qwen/Qwen3.5-0.8B	-0.179 $[-0.915, 0.619]$	1.0000	1.0000
Qwen/Qwen3.5-0.8B-Base	-0.698 $[-1.542, 0.209]$	1.0000	1.0000
reasonwang/ToolGen-Qwen2.5-1.5B-Tool-Retriever	1.152 $[0.197, 2.194]$	0.5567	0.5799
gpt2-large	1.333 $[0.533, 2.206]$	0.0782	0.0828
Qwen/Qwen3.5-2B-Base	0.165 $[-0.778, 1.302]$	1.0000	1.0000
ibm-granite/granite-4.2-3b	0.837 $[0.068, 1.661]$	0.8210	0.8601
mistralai/Ministral-3-3B-Base-2512	1.083 $[-0.006, 2.382]$	1.0000	1.0000
Qwen/Qwen3.5-4B-Base	0.952 $[-0.054, 2.206]$	1.0000	1.0000
openai/gpt-oss-20b	-1.648 $[-2.888, -0.565]$	0.1960	0.2030
stabletoolbench/MirrorAPI-Cache	1.745 $[0.586, 3.009]$	0.1560	0.1664
meta-llama/Llama-3.1-8B	1.171 $[0.109, 2.432]$	Earlier	0.8639
ibm-granite/granite-4.2-8b	0.982 $[0.064, 2.027]$	0.8414	0.8639
mistralai/Ministral-3-8B-Base-2512	1.130 $[-0.045, 2.614]$	1.0000	1.0000
zai-org/glm-4-9b-chat-1m-hf	1.683 $[0.208, 3.493]$	0.8210	0.8601
Qwen/Qwen3.5-9B	0.946 $[-0.191, 2.288]$	1.0000	1.0000
Qwen/Qwen3.5-9B-Base	0.786 $[-0.235, 2.061]$	1.0000	1.0000
Qwen/Qwen2.5-14B	1.699 $[0.598, 2.970]$	Earlier	0.1736
deepseek-ai/DeepSeek-V2-Lite-Chat	1.678 $[0.363, 3.150]$	0.5017	0.5210
Qwen/Qwen3.6-35B-A3B-FP8	0.620 $[0.038, 1.340]$	0.8789	0.8789

Checkpoint	Contrast [95% CI]	Holm new	Holm all
google/gemma-4-26B-A4B	-2.335 $[-5.139, 0.396]$	1.0000	1.0000
google/gemma-4-26B-A4B-it	6.889 $[2.120, 11.703]$	0.1426	0.1518
Qwen/Qwen3.5-27B	0.394 $[0.030, 0.826]$	0.8414	0.8639
ibm-granite/granite-4.2-30b	0.680 $[-0.260, 1.977]$	1.0000	1.0000
zai-org/GLM-4.7-Flash	0.731 $[0.123, 1.428]$	0.6071	0.6335
google/gemma-4-31B-it	1.138 $[-2.409, 4.431]$	1.0000	1.0000
allenai/Olmo-3-1125-32B	1.549 $[0.545, 2.642]$	0.1216	0.1292
openai/gpt-oss-120b	1.521 $[0.544, 2.547]$	0.1023	0.1085
Qwen/Qwen3.5-35B-A3B	0.939 $[0.187, 1.914]$	0.7479	0.7819
Qwen/Qwen3.5-35B-A3B-Base	1.000 $[0.289, 1.926]$	0.5017	0.5210
Qwen/Qwen3.6-35B-A3B	0.515 $[-0.064, 1.222]$	1.0000	1.0000

Table 29. Selected-minus-random heads group contrast on the same 20 idiom/literal and 20 compositional context pairs, in bits. Intervals are pointwise 95%; Holm columns adjust approximate centered-bootstrap p within intervention family, over the new checkpoints or all successful checkpoints including the two earlier models. Missing head tests are not zero effects.

Checkpoint	Contrast [95% CI]	Holm new	Holm all
Qwen/Qwen2.5-0.5B	0.174 $[0.075, 0.289]$	0.0133	0.0171
Qwen/Qwen2.5-1.5B	0.065 $[-0.016, 0.152]$	0.2524	0.2524
Qwen/Qwen2.5-7B	0.096 $[0.045, 0.148]$	0.0016	0.0020
allenai/Olmo-3-1025-7B	0.095 $[0.008, 0.195]$	0.1700	0.1700
NousResearch/Meta-Llama-3.1-8B-Instruct	0.057 $[0.006, 0.116]$	0.1700	0.1700
reasonwang/ToolGen-Qwen2.5-1.5B-Tool-Retriever	0.127 $[0.036, 0.217]$	0.0360	0.0360
stabletoolbench/MirrorAPI-Cache	0.110 $[0.038, 0.193]$	0.0370	0.0370
meta-llama/Llama-3.1-8B	0.068 $[0.025, 0.114]$	Earlier	0.0264
Qwen/Qwen2.5-14B	0.040 $[0.015, 0.068]$	Earlier	0.0301
allenai/Olmo-3-1125-32B	0.002 $[-0.022, 0.022]$	0.8541	0.8541

B.7 Feature-Formula and Auxiliary-Result Audit

The historical normalized-drop and collapse-ratio definitions coincide when their endpoint and denominator definitions agree and the clipping branch is inactive, so they are one feature. The historical enhanced runner used stored drop and a 0.5-bit denominator floor. The rows below that floor, identified by first-token surprisal, number 16, 41, and 56 for GPT-2-Large, Llama-3.1-8B, and Qwen2.5-14B. In each of these models, 12 of 2,066 stored DICE drops differ from the stored first-minus-last endpoints, with maximum absolute differences of 11.507, 13.842, and 12.352 bits. Table 31 replaces only inconsistent rows with endpoint differences for all seven DICE checkpoints while holding covariates, grouping, and matching fixed; the DICE effect sizes in Table 10 change by at most 0.024. Table 10 reports the archived values. The new tokenization endpoint analysis is recomputed from aligned token-level surprisals and does not depend on that historical stored-drop column.

The deduplicated auxiliary rerun removes collapse ratio while retaining the historical inputs and split protocol (Table 30). It is a feature-deduplication check, not a correction of every historical endpoint discrepancy. DICE groups the historical observed MWE field, which can split inflectional variants across groups; it is distinct from the canonical-type probe evaluation. The separate trajectory-family implementation uses relative drop $\Delta_{\text{drop}}/(f + l + 1)$ and collapse ratio $1 - l/\max(f, 1)$, which are not algebraically identical (MAGPIE-test correlation 0.814).

The earlier auxiliary feature-set results remain explicitly historical in Tables 11 and 12. Their ten-seed uncertainty cannot be transferred to the deduplicated run. The main mechanistic argument uses the audited endpoint, probe, and intervention outputs, not differences between redundant feature formulas.

Table 30. Audited auxiliary rerun after removing collapse ratio from the learned feature list. DICE uses historical observed-surface grouping; MAGPIE uses train/dev/test. These are distinct from canonical-expression nested probes. The earlier ten-seed uncertainty applies to the historical feature set, not this rerun.

Dataset	Method	Macro-F1	AUROC
DICE	raw_drop_threshold	0.605	0.652
DICE	lr_trajectory	0.608	0.655
DICE	lr_enhanced	0.628	0.670
DICE	gb_trajectory	0.629	0.682
DICE	gb_enhanced	0.675	0.724
MAGPIE	raw_drop_threshold	0.664	0.756
MAGPIE	lr_trajectory	0.641	0.774
MAGPIE	lr_enhanced	0.649	0.780
MAGPIE	gb_trajectory	0.686	0.805
MAGPIE	gb_enhanced	0.677	0.806

Table 31. Stored-drop sensitivity for the DICE rows of Table 10. Only stored-drop rows inconsistent with the stored first-minus-last endpoints are replaced by the endpoint difference; covariates, grouping, and matching membership are held fixed. These are descriptive effect sizes, not verified corrections of the extraction record.

Dataset/model	Raw d	Resid. d	Matched d
DICE GPT-2-Large	$0.475 \to 0.498$	$0.383 \to 0.398$	$0.465 \to 0.476$
DICE GPT-2-XL	$0.506 \to 0.530$	$0.376 \to 0.389$	$0.504 \to 0.515$
DICE Llama-3.1-8B	$0.553 \to 0.576$	$0.366 \to 0.378$	$0.524 \to 0.534$
DICE Llama-3.1-70B	$0.540 \to 0.562$	$0.348 \to 0.360$	$0.497 \to 0.508$
DICE Qwen2.5-14B	$0.478 \to 0.499$	$0.262 \to 0.268$	$0.431 \to 0.440$
DICE Qwen3-32B	$0.418 \to 0.438$	$0.190 \to 0.198$	$0.392 \to 0.400$
DICE Qwen3.5-122B-A10B	$0.526 \to 0.548$	$0.334 \to 0.344$	$0.475 \to 0.485$

B.8 Artifact Provenance and Reproduction Scope

The anonymous numerical supplement supplies the initial causal outcome rows under `results/initial/raw/`, frozen DICE covariate rows, and selected summaries. CPU reanalysis outputs, including the recipient decomposition, donor audit, position-overlap audit, prefix statistics, and stored-drop sensitivity, are under `results/iclr_revision/`. SHA-256 manifests, deterministic seeds, statistical implementation excerpts, and package versions accompany the data. The exported scripts were run independently and reproduce the corresponding numerical tables. The statistical bootstrap unit is canonical expression type for DICE and expression pair for the pilot; models are not independent linguistic samples.

The package is a numerical reanalysis subset, not a complete GPU experiment archive. Follow-up matched-position recipient rows, common context-control manifests, extension per-pair outcomes, and full per-checkpoint layer metadata are not included; those results retain their saved summaries. The two-model follow-up used BF16 on a single GH200. The larger exploratory sweep used the documented resident-weight budget and per-checkpoint precision and compatibility settings. The pilot candidate sentences were generated with AI assistance, and acceptance relies on researcher batch verification; the study does not add unrecorded item-level ratings or agreement measures. Raw activation caches and model weights are not included. The available records identify the hardware and partial environment information, but do not provide a complete per-run compute ledger or exact environment lock for every historical run.

B.9 Initial Exploratory Mechanistic Diagnostics

The initial ten-model survey is reported in Table 32. Its selected heads and sentence-level probes are descriptive. The matched three-model tests in Section 4.2 carry the revised causal argument; their effects, units, samples, and uncertainty are not interchangeable with these absolute zero-ablation numbers.

Table 32. Initial exploratory localization diagnostics across ten models. Probe F1 is the initial sentence-level result, not nested canonical-expression-disjoint AUROC. Ablation is the absolute final-subtoken effect of zeroing the predictor; it is not an idiom–literal DiD. Attention reports a selected maximum without held-out selection; d is pooled Cohen's d and N is the attention-example count. Depth is best layer divided by $L - 1$; dashes mean not measured.

Model	Fam.	Probing			Ablation		Attention		
		Best L	Depth	F1	Δ	+%	Head	d	N
GPT-2 (124M)	GPT-2	10/11	91%	0.840	+5.23	94%	L5H7	+0.727	100
GPT-2-M (345M)	GPT-2	15/23	65%	0.949	+5.76	100%	L8H2	+0.681	200
GPT-2-L (774M)	GPT-2	21/35	60%	0.957	+3.06	67%	L13H1	+0.629	200
GPT-2-XL (1.5B)	GPT-2	25/47	53%	0.945	+2.61	53%	L9H2	+0.531	200
Qwen2.5-0.5B	Qwen	17/23	74%	0.872	+7.42	97%	L3H8	+0.577	200
Qwen2.5-14B	Qwen	15/47	32%	0.919	+6.45	93%	L11H20	+1.074	200
Llama-3.2-3B	Llama	19/27	70%	0.899	+8.82	97%	L17H19	+0.766	200
Llama-3.1-8B	Llama	15/31	48%	0.903	+1.53	63%	L19H26	+0.826	200
Mistral-7B	Mistral	11/31	35%	0.902	+2.75	67%	L5H8	+0.652	200
Llama-3.1-70B[†]	Llama	13/79	16%	0.904	–	–	–	–	–

[†]Residual-stream ablation and attention sweeps were not run for Llama-3.1-70B because of compute and memory constraints.

Table 33. Raw-drop architecture diagnostic on balanced DICE ($N = 2{,}066$). Model rows use $\Delta_{\text{drop}} > \tau$ with per-model threshold sweeps, so they diagnose separability rather than deployable calibration. The constant-class baseline has no continuous score distribution, so Cohen's d is undefined.

Model	Params	Macro-F1	AUROC	Cohen's d
GPT-2-Large	774M	0.593	0.628	0.475
GPT-2-XL	1.5B	0.597	0.636	0.506
Llama-3.1-8B	8B	**0.611**	**0.652**	**0.553**
Llama-3.1-70B	70B	0.602	0.645	0.540
Qwen2.5-14B	14B	0.588	0.629	0.478
Qwen3-30B-A3B (Instruct)	30B	0.599	0.638	0.508
Qwen3-32B (base)	32B	0.588	0.616	0.418
Qwen3.5-122B-A10B	122B	0.600	0.642	0.526
RoBERTa-Large (PLL)	355M	0.525	0.523	0.078
DeBERTa-v3-Large (PLL)	435M	0.489	0.482	-0.054
Constant-class baseline		0.333	0.500	–

B.10 Behavioral Model Survey

Table 33 reports the raw-drop survey summarized in Section 4.3. The new causal battery has different coverage and estimands.

B.11 Behavioral Boundary and Scaling Figures

Figures 5–7 report the behavioral summaries. They do not establish causal model-size effects.

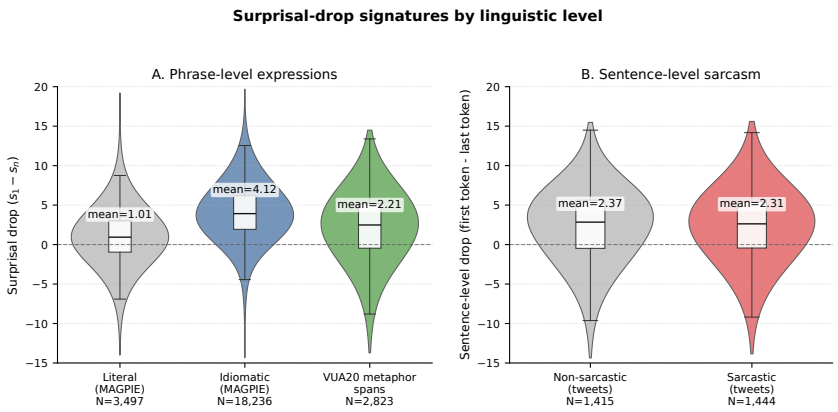

Figure 5. Surprisal-drop signatures by linguistic level. **Left:** phrase-level surprisal drop distributions for MAGPIE literal ($N = 3{,}497$, $\mu = 1.01$), MAGPIE idiomatic ($N = 18{,}236$, $\mu = 4.12$), and VUA20 multi-word metaphor spans ($N = 2{,}823$, $\mu = 2.21$). The ordering is consistent with the within-block effect sizes in Table 5; because the rows come from different corpora, the figure is a qualitative boundary visualization rather than a common-scale benchmark. **Right:** sentence-level surprisal drop distributions for sarcastic ($N = 1{,}444$, $\mu = 2.31$) and non-sarcastic ($N = 1{,}415$, $\mu = 2.37$) tweets show near-overlapping distributions. Violins show kernel-density estimates of the surprisal-drop distribution; inner boxes show median and interquartile range; mean values are annotated. Violin widths are normalized within panel and should not be interpreted as sample size.

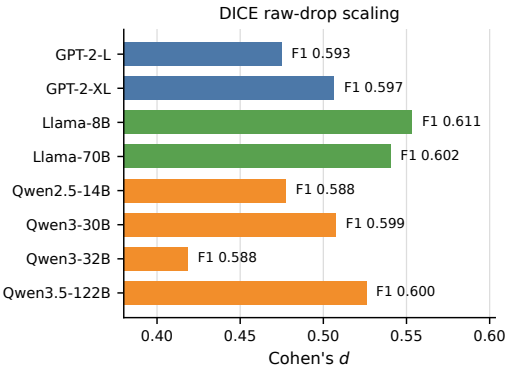

Figure 6. DICE raw-drop scaling summary across autoregressive model families. Larger checkpoints do not monotonically improve raw-drop macro-F1 or effect size.

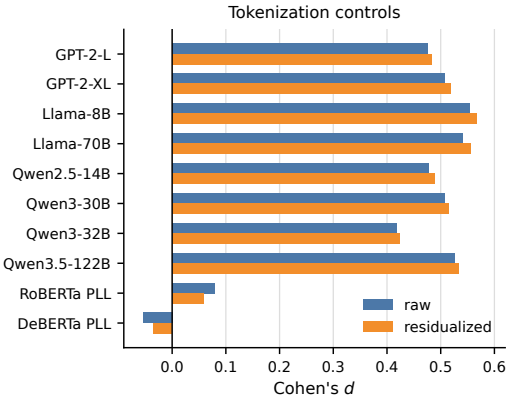

Figure 7. DICE effect size before and after residualizing raw drop on tokenization features. AR-model effects remain positive after segmentation controls.