THE HIDDEN COST OF PROMPT FORMATTING: MEASURING REPRESENTATIONAL SENSITIVITY IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompt formatting is often treated as a superficial implementation detail, yet meaning-preserving format changes can alter model behavior. We ask whether such changes also leave a systematic footprint inside the model before decoding, and whether this internal movement is associated with output instability. We introduce an activation-level audit for prompt-format sensitivity using matched one-shot prompt pairs in which the task instruction, demonstration, input, and answer semantics are fixed while exactly one structural formatting property is perturbed. Across 53 Super-NaturalInstructions tasks and four open-weight 7B-scale models—Falcon, LLaMA-2, Mistral, and Qwen-2.5—we analyze 1,739,400 atomic interventions per model spanning separator, enumeration, spacing, and casing edits. We find a stable hierarchy of representational sensitivity: separator edits induce the largest hidden-state shifts, enumeration edits form a consistent second tier, and spacing and casing produce smaller effects. We further evaluate activation–output coupling over 5,573,004 eligible base–perturbation pairs spanning 168 task–model strata. Controlled activation-distance coefficients are positive in 158/168 strata, while 167/168 within-stratum Spearman correlations are positive, with median $\rho = 0.412$, showing that larger activation shifts are generally associated with larger changes in semantic-label output distributions. Finally, we introduce PFSS, a compact 25-term interpretable predictor of activation movement from prompt-format metadata alone, achieving task-holdout raw-scale R^2 scores of 0.831, 0.773, 0.852, and 0.821. These results establish prompt formatting as a measurable component of the representational input space and activation movement as an informative, though heterogeneous, intermediate signal of output instability.

1 INTRODUCTION

A growing body of work has shown that large language models (LLMs) are remarkably sensitive to superficial variation in prompt construction (Zhao et al., 2021; Min et al., 2022; Lu et al., 2022; Sclar et al., 2024; Chatterjee et al., 2024). Modifications to separators, whitespace, enumeration schemes, labels, templates, and lexical form have been shown to alter predictions, calibration, and task accuracy even when the underlying instruction and task semantics are held fixed (Voronov et al., 2024; Zheng et al., 2024; Qian et al., 2025; Zhan et al., 2024; Errica et al., 2025). Collectively, these results indicate that prompt formatting cannot be regarded as a neutral presentation layer over semantic content. However, the overwhelming majority of this evidence is derived from model outputs alone. While it is well established that two meaning-preserving prompt formats can yield divergent answers or output distributions, considerably less is known about whether such changes systematically reshape the internal representations from which those outputs are ultimately produced.

Consider two prompts that share an identical task instruction, demonstration, evaluation input, candidate answer set, and answer semantics, differing only in a single field separator (e.g., a colon-space delimiter replaced by a tab). If both prompts happen to yield the same decoded answer, an output-only evaluation would treat them as functionally equivalent. Yet the model may nonetheless have routed the two prompts through substantially different internal states prior to decoding. If so, prompt-format sensitivity possesses a representational footprint that remains invisible to output-

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.

level analysis alone. Furthermore, if larger internal shifts are systematically associated with larger changes in task-relevant output distributions, such representations could serve as an intermediate signal for flagging formatting perturbations that merit closer behavioral scrutiny.

This question connects prompt robustness with representation-level analysis of neural language models. A separate line of work studies LLM internals directly, typically through probing classifiers, causal mediation analysis, or activation patching, to characterize what information hidden states encode and how that information is used downstream (Tenney et al., 2019; Vig et al., 2020; Meng et al., 2022). This interpretability literature has established that hidden representations are structured and often causally implicated in specific behaviors, but it has generally targeted semantic or factual content rather than surface-level formatting, and it rarely uses paired, meaning-preserving interventions of the kind required to isolate formatting as the manipulated variable. Related work on tokenization sensitivity has further shown that superficially similar strings can be tokenized differently depending on surrounding context or vocabulary construction, with downstream effects on model behavior (Rust et al., 2021; Singh & Strouse, 2024); our robustness check against invariant label-token mappings (Section 3.2) speaks to this concern but does not fully disentangle formatting from tokenization, as we discuss further in Section 4. Prompt-format sensitivity, in other words, has largely remained an output-side phenomenon: we know that formatting can change what models say, but we know much less about whether formatting systematically changes the internal state from which outputs are generated.

Motivated by this gap, we investigate whether meaning-preserving structural prompt perturbations systematically displace the pre-decoding representations of LLMs, and whether the magnitude of this displacement is associated with corresponding changes in model output. To this end, we construct matched one-shot prompt pairs drawn from Super-NaturalInstructions (Wang et al., 2022), in which the task instruction, demonstration, evaluation input, and answer semantics are held fixed while exactly one formatting property is perturbed. We consider four families of structural edits—separator, enumeration, spacing, and casing—and apply an identical perturbation procedure across four open-weight 7B-scale model families: Falcon, LLaMA-2, Mistral, and Qwen-2.5 (Almazrouei et al., 2023; Touvron et al., 2023; Jiang et al., 2023; Bai et al., 2023). For each base–perturbed pair, we quantify the Euclidean distance between final-layer hidden states at the last input token. In decoder-only Transformers, this state is formed once the entire prompt has been processed and is used directly to compute the next-token logits (Vaswani et al., 2017); we accordingly treat it as a compact diagnostic representation immediately preceding decoding. In doing so, we complement both lines of prior work by measuring how controlled, structural formatting edits displace the model's pre-decoding hidden representation, and by testing whether the magnitude of that displacement is systematically related to changes in the semantic output distribution—connecting representation-level and output-level accounts of prompt-format sensitivity.

Our analysis spans 53 tasks and 1,739,400 atomic formatting interventions per model. We find that formatting perturbations do not displace model representations uniformly. Instead, a stable hierarchy emerges consistently across all four model families: separator edits induce the largest hidden-state shifts, enumeration edits constitute a consistent second tier, and spacing and casing edits yield comparatively smaller displacements. This ordering suggests that structural features governing field boundaries and candidate-answer layout exert a disproportionately large influence on the internal state formed by the model.

We next examine whether this representational displacement relates to externally observable model behavior. For every task–model stratum in which the semantic label distribution can be unambiguously reconstructed, we normalize the model's next-token probabilities over the task's valid semantic labels and compare base and perturbed distributions via Jensen–Shannon divergence (JSD) (Lin, 1991). This analysis encompasses 42 tasks, four models, 168 task–model strata, and 5,573,004 base–perturbation pairs. After controlling for the exact formatting transition, absolute token-length change, and baseline semantic decision margin, the activation-distance coefficient is positive in 158 of 168 strata, and 167 of 168 within-stratum Spearman correlations are likewise positive, with a median $\rho = 0.412$. A random-effects synthesis (DerSimonian & Laird, 1986) yields a pooled standardized coefficient of 0.0153 (95% CI $[0.0143, 0.0162]$). These results indicate that larger formatting-induced shifts in hidden representations are generally accompanied by larger shifts in the task-relevant output distribution.

This relationship, however, is informative rather than deterministic: activation–output coupling varies substantially across tasks, and many distributional shifts do not culminate in discrete task failure. The median task–model answer-flip rate is 13.928%, while the median correct-to-incorrect transition rate is 6.631%. Under a grouped task-held-out evaluation protocol, augmenting formatting metadata with activation distance raises mean JSD R^2 from 0.0989 to 0.1085; for answer flips, AUC increases from 0.8082 to 0.8128 and log loss decreases from 0.5455 to 0.5399. These findings support treating activation movement as an intermediate audit signal that complements, rather than substitutes for, output-side robustness measures.

Finally, we examine whether the magnitude of representational displacement can itself be anticipated from the structure of the formatting intervention. We introduce the Pruned Feature-Specific Sensitivity predictor (PFSS), a compact 25-term model that predicts activation distance using only transparent prompt-format metadata—including separator type, enumeration structure, casing, and token-length change—while explicitly excluding raw prompt text, task identifiers, hidden states, logits, and generated outputs. Despite this restriction, PFSS achieves task-holdout raw-scale R^2 scores of 0.831, 0.773, 0.852, and 0.821 for Falcon, LLaMA-2, Mistral, and Qwen-2.5, respectively, indicating that a substantial portion of representational sensitivity is recoverable from the structural properties of the prompt alone.

Our contributions are summarized as follows.

- **Prompt formatting as a representational perturbation.** We introduce a controlled, paired auditing framework in which semantic content is held fixed while a single structural formatting property is varied, enabling measurement of the resulting displacement in a model's pre-decoding hidden representation across millions of interventions.
- **A stable hierarchy of representational sensitivity.** Across 53 tasks and four independently trained 7B-scale model families, we identify a consistent separator $>$ enumeration $>$ spacing/casing hierarchy, demonstrating that formatting features governing prompt structure induce disproportionately large internal shifts.
- **Activation movement is coupled to semantic output movement.** Across 168 eligible task–model strata and 5.57 million prompt pairs, larger activation distances are generally associated with larger shifts in semantic-label output distributions. This relationship is highly consistent in direction but heterogeneous in magnitude, supporting activation distance as an informative intermediate audit signal rather than a deterministic failure metric.
- **Representational sensitivity is predictable from prompt structure.** We introduce PFSS, a compact and interpretable predictor demonstrating that much of the observed activation movement is recoverable from prompt-format metadata alone, while generalizing across held-out semantic tasks.

2 Method: Activation-Level Prompt Perturbation Audit

We study prompt-format sensitivity through controlled paired comparisons. Each intervention instance consists of a base prompt and a meaning-preserving perturbed prompt that differ by exactly one structural formatting property. We measure how much this edit moves the model's internal representation, and whether that movement is associated with observable changes in model output.

2.1 Task Source and Prompt-Pair Construction

We construct matched one-shot prompt pairs from 53 Super-NaturalInstructions tasks (Wang et al., 2022), following the task split used in prior prompt-format sensitivity work (Sclar et al., 2024): 19 multiple-choice tasks and 34 classification tasks with basic descriptor fields (full task IDs in Appendix A). For each task, one instance is reserved as the one-shot demonstration and the remainder serve as evaluation inputs; a fixed random seed keeps the demonstration and evaluation instances identical across all model runs, so that activation differences arise from formatting alone.

Let z_i denote an evaluation instance and $c_0 = (c_1, \ldots, c_m)$ the canonical prompt-format configuration, where each c_j is one structural parameter. The base prompt is

$$p_i = B(z_i, c_0), \tag{1}$$

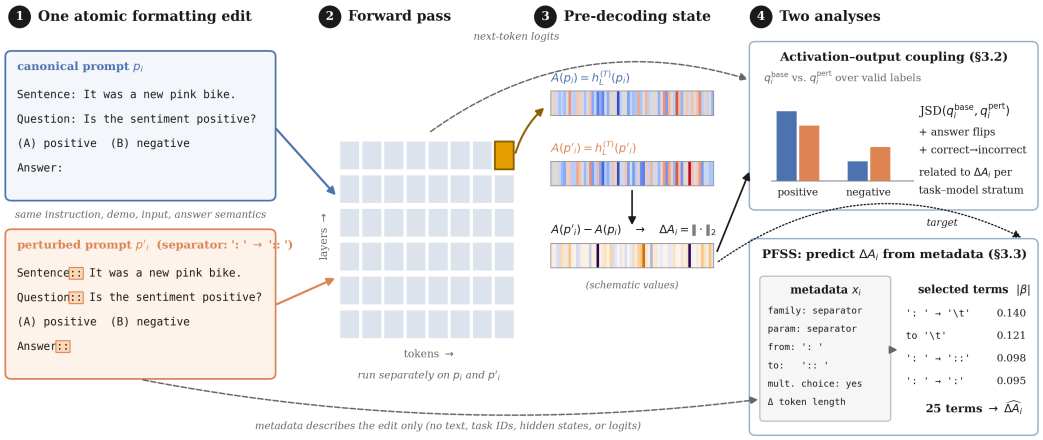


Figure 1: Anatomy of one intervention. (1) A canonical prompt and its perturbed counterpart differ by exactly one formatting parameter, here the field separator, applied consistently to every field. (2) Each prompt is run separately through the model. (3) The final-layer hidden state at the last input token is extracted for both prompts, and their Euclidean distance defines the activation displacement ΔA_i. (4) ΔA_i is then related to movement in the restricted semantic-label distribution (Section 3.2) and predicted from edit metadata alone by PFSS (Section 3.3). Activation strips and label-probability bars are schematic; the selected PFSS terms and $|\beta|$ values are taken from Table 1.

where $B(\cdot)$ maps a task instance and formatting configuration to a textual prompt. The canonical format uses colon-space field separators, title-cased labels, newline-separated fields, parenthesized uppercase enumeration labels, and newline-separated answer options; it serves as the control condition throughout.

2.2 Atomic Formatting Perturbations

A perturbed prompt changes exactly one formatting parameter k relative to c_0:

$$c'_j = \begin{cases} v'_k, & j = k, \\ c_j, & j \neq k, \end{cases} \qquad p'_i = B(z_i, c'), \tag{2}$$

so each pair (p_i, p'_i) preserves task instruction, demonstration, evaluation input, and answer semantics, differing only in the selected edit. We study four perturbation families: **separator** edits change the field-label delimiter (e.g., colon-space to tab, double colon, or newline variants); **enumeration** edits change option wrappers, index format (letters, numbers, Roman numerals), or option spacing, and apply only when a task has explicit option structure; **spacing** edits alter whitespace and field placement (e.g., same-line vs. separate-line fields); **casing** edits change label capitalization. Full parameter value sets and example renderings are given in Appendix B. For multiple-choice tasks, enumeration edits are propagated consistently: the prompt builder rewrites both the option list and any instruction text referencing those labels, so the edit changes structure without altering the answer mapping (Appendix A).

2.3 Model Coverage and Hidden-State Extraction

We apply this pipeline to four open-weight 7B-scale models: Falcon, LLaMA-2, Mistral, and Qwen-2.5 (Almazrouei et al., 2023; Touvron et al., 2023; Jiang et al., 2023; Bai et al., 2023). Each model-specific dataset contains 1,739,400 atomic interventions, yielding nearly seven million activation comparisons in total.

For a tokenized prompt $p = (x_1, \ldots, x_T)$, a Transformer-based model (Vaswani et al., 2017) with L layers and hidden dimension d produces layer-wise hidden states $H_\ell(p) \in \mathbb{R}^{T \times d}$. We define the

prompt-level diagnostic activation as the final-layer hidden state at the last input token,

$$A(p) = h_L^{(T)}(p), \tag{3}$$

the representation formed immediately after the full prompt is processed and used to compute the next-token logits. We use *hidden state* and *activation vector* interchangeably. Base and perturbed prompts are run separately; prompts are tokenized with the model's own tokenizer and truncated to 2048 tokens. Full extraction details are in Appendix G.

2.4 Activation-Distance Targets

For each pair, we compare $A(p_i)$ and $A(p_i')$ and interpret their distance as representational sensitivity to the formatting edit. Our primary target is Euclidean distance,

$$\Delta A_{2,i} = \|A(p_i') - A(p_i)\|_2 \,, \tag{4}$$

following standard vector-space distance definitions (Deza & Deza, 2009). To test whether results depend on this choice, we additionally compute Manhattan distance, cosine distance, and maximum absolute activation change (defined in Appendix C).

2.5 Prompt-Format Metadata

Each instance carries a metadata vector x_i describing the edit: perturbation family, parameter name, original and changed value, separator type, enumeration wrapper/index, casing pattern, multiple-choice status, and token lengths (base, perturbed, difference). This feature set deliberately excludes raw prompt text, prompt/task identifiers, hidden states, logits, generated outputs, and non-target activation metrics, so predictive models must estimate activation movement from transparent formatting metadata alone.

2.6 Activation–Output Coupling

We next test whether representational displacement is associated with observable output movement, using the next-token logits recorded during activation extraction. Of 212 possible task–model strata, 168 permit unambiguous reconstruction of the restricted semantic-label distribution (42 tasks, all four models, 41,995 evaluation examples, 5,573,004 pairs); the remainder are excluded because their label mappings cannot be reconstructed unambiguously (eligibility criteria in Appendix D.1). For each eligible pair, we renormalize the next-token distribution over the task's valid labels and quantify continuous output movement via Jensen–Shannon divergence (JSD), alongside two discrete outcomes: answer flips and correct-to-incorrect transitions.

We assess coupling along three axes: (i) within-stratum Spearman correlation between activation distance and JSD; (ii) a within-stratum regression of JSD on standardized log activation distance, controlling for the exact formatting transition, absolute token-length change, and baseline semantic decision margin, with example-clustered standard errors and random-effects synthesis across strata; and (iii) grouped task-held-out comparisons of metadata-only versus metadata-plus-activation-distance predictors, for both JSD and answer flips. Full formal definitions are given in Appendix D.

2.7 Pruned Feature-Specific Sensitivity Predictor

Finally, we test whether activation movement is itself predictable from prompt-format metadata alone. Let $y_i = \log(1 + \Delta A_i)$ denote the log-transformed target. PFSS is a compact selected-term predictor,

$$\hat{y}_i = \beta_0 + \sum_{j \in S} \beta_j \phi_j(x_i), \qquad |S| = 25,$$

where $\phi_j(\cdot)$ are readable transformations of the metadata (nonlinear expansions of continuous features; indicator terms for categorical ones). PFSS fits an ElasticNet selector (Zou & Hastie, 2005) over the full feature library, ranks terms by absolute coefficient magnitude, retains the top 25, and refits them via ordinary least squares (or Ridge (Hoerl & Kennard, 1970), when unstable). Full feature definitions, selection objective, and hyperparameters are in Appendix E.

3 RESULTS

We report results primarily as within-model sensitivity patterns because raw activation scales differ across model families. Across models, separator and enumeration edits produce the largest shifts, this movement is systematically coupled to changes in task-relevant output distributions, and PFSS shows that much of it is predictable from prompt-format metadata alone. Together, these results support the view that prompt formatting is part of the representational input space, not merely a surface wrapper around task meaning.

3.1 SENSITIVITY HIERARCHY

For each model m and perturbation family f (or parameter r), we compute the mean activation distance within that subgroup,

$$\overline{\Delta A}_2^{(m,f)} = \frac{1}{|\mathcal{I}_{m,f}|} \sum_{i \in \mathcal{I}_{m,f}} \Delta A_{2,i}. \tag{5}$$

A stable hierarchy emerges consistently across all four models: separator perturbations dominate, enumeration edits form a consistent second tier, and spacing and casing produce the smallest shifts. This ordering holds at the parameter level as well—separator changes remain the single largest driver, followed by enumeration wrapper and enumeration index, while field spacing, casing, and enumeration label spacing are consistently smaller—indicating that the effect is driven by specific boundary- and answer-structure markers rather than a broad formatting category. The consistency of this ordering across four independently trained model families suggests it reflects a structural property of the perturbations rather than a model-specific artifact.

This hierarchy is useful because it identifies which formatting dimensions should receive closer attention in prompt robustness audits: delimiter and enumeration choices should not be treated as incidental notation. A model evaluated under a single separator or answer-option format may appear more stable than it would under a broader set of semantically equivalent prompt structures, and this apparent stability is doubly misleading given that activation movement and output movement rise together (Section 3.2)—the formatting choices that most disturb the model's internal state are also the ones most likely to disturb its output distribution.

Raw activation magnitudes should not, however, be compared directly across models, since hidden-state scales depend on architecture, dimensionality, normalization, and implementation details. Our strongest cross-model claim is therefore not that the same numerical activation distance carries the same meaning across models, but rather the recurrence of a relative sensitivity hierarchy: separator and enumeration edits consistently produce larger internal shifts than spacing and casing edits across all four independently trained families, a pattern echoed both in the parameters PFSS most consistently selects (Section 3.4) and in the strata where activation–output coupling is strongest (Section 3.2).

3.2 ACTIVATION–OUTPUT COUPLING

Having established that formatting edits systematically move the pre-decoding representation, we ask whether this movement relates to the task-relevant output distribution. We restrict analysis to the 168 of 212 task–model strata in which the semantic label set maps unambiguously to next-token outputs (42 tasks, all four models, 5,573,004 pairs; eligibility criteria in Appendix D.1), and measure continuous output movement as Jensen–Shannon divergence (JSD) between base and perturbed semantic-label distributions.

Direction and magnitude. Activation distance and JSD are positively associated in nearly every stratum: 167 of 168 within-stratum Spearman correlations are positive (median $\rho = 0.412$), and grouping by within-stratum activation-distance decile shows JSD rising monotonically with decile (Appendix D.6). This association is not merely a byproduct of particular formatting transitions producing both large activation and output shifts together: after regressing JSD on standardized log activation distance while controlling for the exact formatting transition, token-length change, and baseline decision margin (example-clustered SEs), the activation-distance coefficient remains positive in 158 of 168 strata. A random-effects synthesis gives a pooled standardized coefficient $\hat{\beta} =$

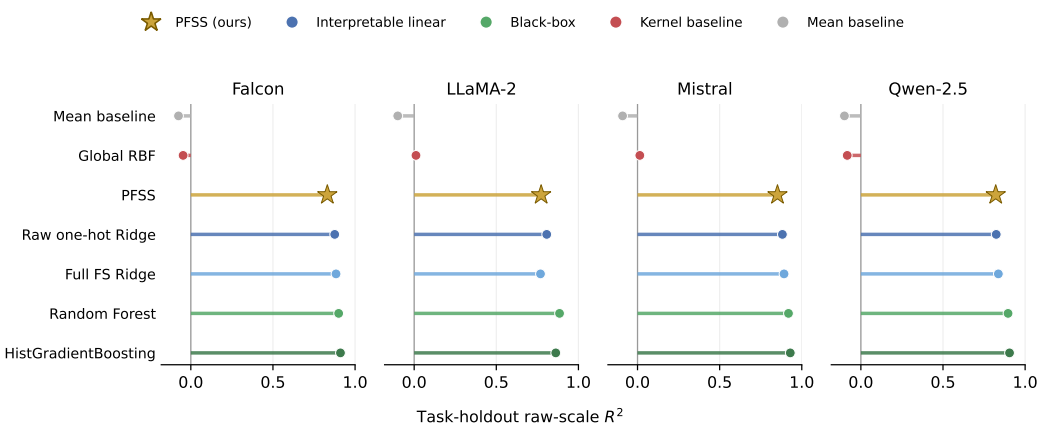


Figure 2: Task-holdout raw-scale R^2 for predicting Euclidean activation distance, across all methods and models. PFSS (gold star) uses only 25 selected structural terms yet sits close to the flexible black-box ceiling (Random Forest, HistGradientBoosting, green) and clearly above the Global RBF and mean baselines. Full numeric values, including random-split and log-scale results, are in Appendix E, Table 10.

0.0153 (95% CI $[0.0143, 0.0162]$), with substantial between-stratum heterogeneity ($I^2 \approx 99\%$). Full model specification and per-stratum results are in Appendix D.7–D.8.

Continuous movement vs. discrete failure. Output-distribution movement is far more common than answer changes: the median answer-flip rate is 13.928% and the median correct-to-incorrect transition rate is 6.631%, confirming that many perturbations shift label probabilities without changing the decoded answer. Activation distance also contributes modest held-out predictive value beyond formatting metadata alone: for JSD, mean R^2 rises from 0.0989 to 0.1085 under grouped task-holdout evaluation (MAE essentially unchanged); for answer flips, AUC rises from 0.8082 to 0.8128 and log loss falls from 0.5455 to 0.5399, improving in 9 of 10 folds (Appendix D.10–D.11). The relationship is not an artifact of formatting-induced tokenization changes to the labels themselves: restricting to interventions with an invariant semantic-label token mapping leaves 161 of 168 coefficients positive, strongly correlated with the full-sample coefficients ($\rho \approx 0.954$; Appendix D.12); we return to the broader question of formatting versus tokenization in Section 4.

3.3 Predicting Activation Movement

Having shown that activation movement is systematically coupled to output movement, we next test whether activation movement itself can be predicted from prompt-format metadata alone. Figure 2 reports PFSS and baseline predictive performance under task-holdout evaluation on Euclidean activation distance ΔA_2, the primary and most stringent setting: all examples from held-out tasks are assigned to the test set while the perturbation vocabulary remains fixed. Full results, including random-split and log-scale metrics for every method and model, are given in Appendix E (Table 10).

Two patterns stand out. First, raw one-hot Ridge already performs strongly: even after excluding raw prompt text, task identifiers, logits, outputs, and hidden states, transparent formatting metadata retains substantial predictive signal on its own. Second, black-box models (Random Forest, HistGradientBoosting) achieve the highest scores, as expected, but PFSS retains most of that signal using only 25 selected structural terms. The relevant comparison is therefore not whether PFSS beats the strongest black-box regressor, but how much predictive signal a compact, interpretable model can recover: on the raw activation-distance scale, PFSS reaches task-holdout R^2 of 0.831 (Falcon), 0.773 (LLaMA-2), 0.852 (Mistral), and 0.821 (Qwen-2.5)—competitive with Ridge-based baselines despite its far smaller term budget, and indicating that, within the observed perturbation vocabulary, there is a transferable relationship between structural formatting features and hidden-state displacement. PFSS is best understood as interpretable compression rather than a superior predictor: its

Selected term	Freq.	Models	Mean $\lvert\beta\rvert$
separator: : → \t	8	4	0.140
perturbed_separator \t	8	4	0.121
separator: : → : :	8	4	0.098
separator: : → :	8	4	0.095
to_value : :	8	4	0.094
to_value :	8	4	0.093
separator: : → ␣\n\t	8	4	0.085
to_value ␣\n\t	8	4	0.086

Table 1: Most stable PFSS selected terms across models and validation settings.

value is compressing most of the predictive signal into a small set of readable structural terms that make activation movement easy to inspect and compare across models. It should also be understood as a diagnostic predictor rather than a mechanistic explanation—its selected terms identify formatting features associated with activation movement, but they do not identify the neurons, attention heads, circuits, or causal pathways responsible for that movement.

3.4 Selected-Term Stability

Table 1 summarizes the most frequently selected PFSS terms across models and validation settings. Separator-related terms dominate—especially transitions from the canonical colon-space separator to tab, colon-only, double-colon, and space-prefixed newline-tab variants—including both transition features (encoding the edit itself) and resulting-value features (encoding the introduced delimiter). Their recurrence across all four models echoes the family- and parameter-level sensitivity hierarchy above, indicating that boundary-marker changes are both descriptively large and predictively stable.

That separator-related terms are simultaneously the largest source of activation displacement (Section 3.1) and the most stable predictors of that displacement reinforces the practical takeaway: deployed systems may benefit from deliberate standardization of separators, option labels, and field layouts, and from using activation distance as a cheap pre-screen to prioritize which formatting variants warrant full output-level testing before deployment.

3.5 Robustness Across Distance Targets

PFSS also predicts alternative activation-distance targets with comparable strength: random-split R^2 for cosine distance ranges from 0.825 (Falcon) to 0.928 (Mistral), and for maximum absolute activation change from 0.771 (Qwen-2.5) to 0.941 (Mistral) (full table in Appendix F). This indicates the metadata signal reflects hidden-state movement broadly—directional change and localized coordinate-wise shifts alike—rather than being an artifact of choosing Euclidean distance as the primary target.

4 Discussion and Conclusion

The results support the view that prompt formatting is part of the representational input space, not merely a surface wrapper around task meaning. In the controlled setting studied here, structure-bearing edits—especially delimiters and answer-option formatting—produce the largest internal shifts, while spacing and casing have smaller effects. This hierarchy is useful because it identifies which formatting dimensions should receive closer attention in prompt robustness audits.

Activation-level auditing complements output-level robustness evaluation. Output metrics show whether formatting changes alter answers, confidence, or accuracy, while activation metrics show whether formatting changes move the pre-decoding representation from which the next-token distribution is computed. Our coupling analysis links the two directly: larger activation distances are associated with larger semantic-label JSD in 167 of 168 eligible task–model strata (median $\rho = 0.412$), and this relationship survives controlling for the formatting transition, token-length change, and baseline decision margin (pooled $\hat{\beta} = 0.0153$, 95% CI $[0.0143, 0.0162]$). This signal is useful precisely because it is not redundant with output-level evaluation: two prompts may yield

the same decoded answer while still placing the model in different internal representational states, and many large representational shifts do not culminate in a changed decoded answer (the median answer-flip rate is 13.928%, well below the rate at which valid-label probabilities shift at all). At the same time, the relationship is far from deterministic—between-stratum heterogeneity is substantial ($I^2 \approx 99\%$)—so we treat activation distance as an intermediate audit signal rather than a direct proxy for answer changes: it reliably flags formatting edits that perturb the model's internal state and output distribution, but confirming whether that instability changes the final decision still requires output-level evaluation.

The predictive results further show that activation movement is not arbitrary. PFSS compresses much of the signal into a compact set of readable prompt-format terms while retaining strong task-holdout performance. This suggests that, within the observed perturbation vocabulary, there is a transferable relationship between structural formatting features and hidden-state displacement. At the same time, PFSS should be understood as a diagnostic predictor. Its selected terms identify formatting features associated with activation movement, but they do not identify the neurons, attention heads, circuits, or causal pathways responsible for that movement.

The practical implication is that delimiter and enumeration choices should not be treated as incidental notation. A model evaluated under a single separator or answer-option format may appear more stable than it would under a broader set of semantically equivalent prompt structures. Conversely, deployed systems may benefit from deliberate standardization of separators, option labels, and field layouts to reduce unnecessary representational drift across otherwise similar requests.

The scope of the study is intentionally controlled. We focus on structural rather than semantic perturbations because this allows paired comparisons in which the task instruction, demonstration, input, and answer semantics remain fixed. This design is useful for isolating formatting as the manipulated variable, but it also means that the results should not be generalized directly to semantic paraphrases, role prompts, long-context prompts, multi-turn dialogue, or instruction-strength changes. Those settings may introduce semantic, pragmatic, or discourse-level changes that require separate analysis.

Formatting edits are also realized through token sequences. We therefore do not attempt to separate formatting from tokenization entirely. This is not a flaw in the audit setting, because deployed prompts are ultimately processed as tokenized strings; tokenization is one legitimate pathway through which rendered format can affect internal representations. However, the present results should be interpreted as effects of rendered prompt surfaces, including their tokenization consequences, rather than as evidence for a tokenization-independent formatting mechanism.

Our activation measure uses the final-layer hidden state of the last input token. This choice provides a compact prompt-level representation immediately before next-token prediction in autoregressive language models, making it a useful diagnostic state for large-scale paired auditing. At the same time, it captures only one view of the model's internal computation. Layerwise states, pooled prompt representations, attention patterns, or output-token-specific states may reveal additional forms of sensitivity. Future work can therefore extend the audit from a single diagnostic state to a fuller representational trajectory across layers and token positions.

Finally, raw activation scales should not be compared directly across models because models differ in architecture, hidden-state dimensionality, normalization, and representation scale (Kornblith et al., 2019). For this reason, our strongest cross-model claim is not that the same numerical activation distance has the same meaning across models. Rather, the central cross-model result is the recurrence of a relative sensitivity hierarchy: separator and enumeration edits consistently produce larger internal shifts than spacing and casing edits.

Within these boundaries, the study supports a narrow but defensible conclusion: prompt-format sensitivity has a measurable representational footprint, and that footprint is systematically, if imperfectly, coupled to task-relevant output movement. In a controlled structural perturbation setting, activation movement can be measured, ranked, shown to relate to output-distribution change, and partially predicted from transparent prompt-format metadata alone. Prompt formatting is therefore not merely a surface-level presentation choice; it is part of the representational input space through which models form internal states before decoding.

AI USE STATEMENT

Large language model-based writing assistance was used during the preparation of this manuscript, limited to sentence-level editing, grammar and clarity suggestions, LaTeX formatting and figure-generation assistance (including diagram and chart code), citation verification, and section reorganization support. All experimental design, implementation, execution, and analysis were conducted by the authors. All research claims, numerical results, and interpretations reported in this paper were verified by the authors and reflect the authors' own findings; no AI system was used to generate, fabricate, or select experimental results, data, or citations without independent verification by the authors. The authors take full responsibility for the content of this paper, including any text, figures, or code assisted by AI tools, and confirm that this use is consistent with the ICLR 2027 AI Policy for Authors.

ETHICS STATEMENT

The authors have read the ICLR Code of Ethics and confirm that this submission adheres to it. This work audits activation-level sensitivity of open-weight language models to structural prompt formatting. All models used (Falcon, LLaMA-2, Mistral, and Qwen-2.5) are publicly released, open-weight models used under their respective licenses, and all tasks are drawn from the publicly available Super-NaturalInstructions benchmark (Wang et al., 2022); no new data involving human subjects, private information, or personally identifiable content was collected, and no user studies were conducted. The perturbations studied are meaning-preserving structural edits (separators, enumeration, spacing, casing) and do not involve adversarial content generation, jailbreaking, or the elicitation of harmful model outputs. We do not foresee direct harms from this work; rather, its intended contribution is diagnostic, helping practitioners identify formatting choices that induce larger internal representational shifts and potentially warrant closer output-level robustness testing before deployment. We are not aware of any conflicts of interest, dual use concerns, or funding sources that could be perceived as influencing the design, execution, or reporting of this study.

4.1 REPRODUCIBILITY PACKAGE

An anonymized reproducibility package is available at this link. The repository contains the analysis notebook, perturbation grammar, task identifiers, split assignments, processed prompt-format metadata, derived activation-distance targets, and PFSS selected terms, together with all derived activation–output coupling tables and figures. The full activation-extraction pipeline produced approximately 200GB of intermediate artifacts across four model-specific runs. These raw hidden-state artifacts are not required to reproduce the aggregate analyses once activation distances have been computed. The processed metadata and derived targets are therefore released rather than the complete raw hidden-state cache.

REFERENCES

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 26(2):211–243, 1964.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

David S. Broomhead and David Lowe. Multivariable functional interpolation and adaptive networks. *Complex Syst.*, 2, 1988. URL https://api.semanticscholar.org/CorpusID:3686496.

Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. POSIX: A prompt sensitivity index for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14550–14565, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.852. URL https://aclanthology.org/2024.findings-emnlp.852/.

Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7 (3):177–188, 1986.

Michel Marie Deza and Elena Deza. Encyclopedia of distances. In *Encyclopedia of distances*, pp. 1–583. Springer, 2009.

Federico Errica, Davide Sanvito, Giuseppe Siracusano, and Roberto Bifulco. What did I do wrong? quantifying LLMs' sensitivity and consistency to prompt engineering. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1543–1558, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.73. URL https://aclanthology.org/2025.naacl-long.73/.

Trevor Hastie. The elements of statistical learning: data mining, inference, and prediction, 2009.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning: with applications in R*, volume 103. Springer, 2013.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, 2022.

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

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

Junlang Qian, Zixiao Zhu, Hanzhang Zhou, Zijian Feng, Zepeng Zhai, and Kezhi Mao. Beyond the next token: Towards prompt-robust zero-shot classification via efficient multi-token prediction. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7093–7115, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.363. URL https://aclanthology.org/2025.naacl-long.363/.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3118–3135, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.243. URL https://aclanthology.org/2021.acl-long.243/.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, volume 2024, pp. 25055–25083, 2024.

Aaditya K. Singh and DJ Strouse. Tokenization counts: the impact of tokenization on arithmetic in frontier llms, 2024. URL https://arxiv.org/abs/2402.14903.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4593–4601, 2019.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*, 2020.

Anton Voronov, Lena Wolf, and Max Ryabinin. Mind your format: Towards consistent evaluation of in-context learning improvements. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6287–6310, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.375. URL https://aclanthology.org/2024.findings-acl.375/.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 5085–5109, 2022.

Pengwei Zhan, Zhen Xu, Qian Tan, Jie Song, and Ru Xie. Unveiling the lexical sensitivity of LLMs: Combinatorial optimization for prompt enhancement. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5128–5154, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.295. URL https://aclanthology.org/2024.emnlp-main.295/.

Task type	Task IDs
Multiple-choice (19)	task065, task069, task070, task190, task213, task214, task220, task296, task297, task385, task580, task1297, task1387, task1419, task1420, task1421, task1423, task1612, task1678
Classification (34)	task050, task114, task133, task155, task158, task161, task162, task163, task279, task280, task286, task316, task317, task319, task320, task322, task323, task325, task326, task327, task328, task335, task337, task607, task608, task609, task904, task905, task1186, task1283, task1284, task1347, task1502, task1724

Table 2: Super-NaturalInstructions task IDs used in the study, grouped by prompt structure type. The task pool contains 19 multiple-choice tasks and 34 classification tasks with basic descriptor fields.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pp. 12697–12706. Pmlr, 2021.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=shr9PXz7T0.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

A Dataset and Perturbation Construction

This appendix provides additional details on the construction of the prompt-pair intervention datasets. The goal of the dataset design is to isolate the representational effect of atomic prompt-format changes while holding task content fixed. Each retained row therefore corresponds to a controlled comparison between a canonical prompt and a structurally perturbed prompt that preserves the same task semantics.

A.1 Task Pool and One-Shot Setting

We construct matched one-shot prompt-pair intervention datasets from 53 tasks in Super-NaturalInstructions (Wang et al., 2022). The task pool contains 19 tasks with explicit multiple-choice structure and 34 classification tasks with basic descriptor fields. These tasks are suitable for the present study because they can be rendered as label-selection problems while allowing the same underlying instance to be expressed in multiple structurally equivalent prompt formats. For each task, one instance is reserved as the one-shot demonstration and the remaining instances are used as evaluation inputs. The demonstration example and evaluation examples are fixed across all model runs using a fixed random seed. This ensures that activation differences are attributable to prompt-format perturbations rather than changes in task examples, demonstrations, or input content.

A.2 Prompt-Pair Unit

The fundamental unit of analysis is a prompt-pair intervention instance:

$$(p_i, p_i', x_i, \Delta A_i), \tag{6}$$

where p_i is the canonical base prompt, p_i' is the structurally perturbed version of the same prompt, x_i is the metadata vector describing the formatting edit, and ΔA_i is the activation-distance target computed from model hidden states. For every intervention instance, the base and perturbed prompts contain the same task instruction, one-shot demonstration, evaluation input, candidate-answer semantics, and expected answer mapping. They differ by exactly one structural formatting parameter. This paired design allows each row to be interpreted as a controlled formatting intervention rather than a comparison between different tasks or examples.

A.3 Canonical Baseline Prompt

All prompt pairs are generated relative to a canonical baseline format. The canonical configuration uses:

- colon-space field separators;
- title-cased field labels;
- newline-separated fields;
- parenthesized uppercase enumeration labels for multiple-choice options;
- newline-separated answer options.

Let z_i denote an evaluation instance, and let

$$c_0 = (c_1, c_2, \ldots, c_m) \tag{7}$$

denote the canonical prompt-format configuration, where each c_j is one structural formatting parameter. The canonical base prompt is produced by a prompt-building function:

$$p_i = B(z_i, c_0), \tag{8}$$

where $B(\cdot)$ maps a task instance and formatting configuration into a textual prompt.

A.4 Atomic Perturbation Generation

Perturbed prompts are generated by changing exactly one formatting parameter relative to the canonical configuration. For a perturbation applied to parameter k, the perturbed configuration c' is defined as:

$$c'_j = \begin{cases} v'_k, & j = k, \\ c_j, & j \neq k, \end{cases} \tag{9}$$

where v'_k is the alternative value assigned to the selected formatting parameter. The perturbed prompt is then:

$$p'_i = B(z_i, c'). \tag{10}$$

This construction prevents a single prompt variant from simultaneously changing separators, spacing, casing, and enumeration structure. Each retained intervention is therefore attributable to one interpretable formatting contrast.

A.5 Perturbation Families

We study four families of structural formatting edits.

Separator edits. Separator edits modify the delimiter between field labels and field content. For example, a canonical colon-space separator may be changed to a colon-only separator, double colon, tab, newline, or newline-tab variant. These edits alter how boundaries between prompt fields are marked.

Enumeration edits. Enumeration edits modify the rendering of answer options or list structures. These include changes to option wrappers, such as `(A)` to `[A]` or `A.`; changes to enumeration indices, such as uppercase letters, lowercase letters, numbers, or Roman numerals; and changes to option spacing or option separators. Enumeration edits are applied only when a task contains explicit option structure.

Spacing edits. Spacing edits alter whitespace and field placement. For example, fields may be placed on separate lines or grouped on the same line, and newline boundaries may be replaced by spaces or other separators.

Casing edits. Casing edits modify capitalization patterns in prompt labels. For example, labels such as `Sentence`, `Question`, and `Answer` may be changed to lowercase or uppercase variants.

A.6 Multiple-Choice Rewriting

Multiple-choice tasks require additional handling because enumeration edits must preserve the mapping between option labels and candidate answers. When an enumeration perturbation changes option labels or wrappers, the prompt builder rewrites both the option list and any corresponding instruction text that refers to those labels. This ensures that the perturbation changes only the structural rendering of the options, not their semantic identity. For example, a canonical option list such as:

```
(A) positive
(B) negative
```

may be rewritten as:

```
[A] positive
[B] negative
```

or:

```
A. positive
B. negative
```

while preserving the original candidate-answer mapping. This check is important because otherwise an apparent formatting perturbation could accidentally become a semantic or label-mapping perturbation.

A.7 Dataset Granularity and Filtering

The dataset unit is one evaluation example paired with one valid atomic formatting perturbation. Therefore, for a given task, the number of rows is approximately the number of evaluation examples multiplied by the number of valid atomic interventions for that task. Multiple-choice tasks typically yield more intervention rows because they support additional enumeration-related edits. The same prompt-pair construction pipeline is applied separately to Falcon, LLaMA-2, Mistral, and Qwen-2.5. Each model-specific dataset contains 1,739,400 prompt-pair intervention instances. Across four models, this yields nearly seven million activation comparisons. The datasets are model-specific because hidden-state vectors and activation-distance targets are extracted separately for each model, while the perturbation schema and task setup are held constant. Two filtering steps are applied before predictive analysis. First, we retain only rows with `prompt_changed = 1`, excluding degenerate cases in which a nominal intervention does not alter the rendered prompt string. Second, we remove raw prompt text and output-side generation fields from the cleaned predictive-analysis tables. The predictive files therefore exclude base prompts, perturbed prompts, generated outputs, prediction-side diagnostics, and output-comparison fields. This prevents predictive models from using lexical prompt content, task identity, model outputs, or hidden states as inputs.

A.8 Saved Metadata

Each intervention instance stores metadata describing the task, prompt format, perturbation, tokenization, and activation-distance targets. The metadata includes task type, example index, multiple-choice status, perturbation family, parameter name, original value, changed value, separator type, enumeration wrapper, enumeration index, casing pattern, base prompt token length, perturbed prompt token length, token-length difference, and activation-distance metrics. The predictive feature set deliberately excludes raw prompt text, prompt identifiers, task identifiers, hidden states, logits, generated outputs, and non-target activation metrics. Thus, predictive models estimate activation movement from transparent prompt-format metadata alone.

B Prompt-Format Grammar and Examples

Table 3 summarizes the structural prompt-format grammar used to generate atomic perturbations. Each perturbation modifies one parameter relative to the canonical prompt format while preserving task content.

Family	Parameter	Full value set
Separator	separator	::: , :: , : , \n \t , \n , : , - , , \n , \n \t , :, ::, - , \t
Casing	casing	identity, title, upper, lower
Spacing	field_spacing	, , \n , \n , -- , , ; \n , || , <sep> , -- , , , \n , , , \n , . , ,
Enumeration	enum_separator	::: , :: , : , \n \t , \n , : , - , , \n , \n \t , :, ::, - , \t
Enumeration	enum_label_spacing	, , \t
Enumeration	enum_spacing	, , \n , \n , -- , , ; \n , || , <sep> , -- , , , \n , , , \n , . , ,
Enumeration	enum_wrapper	parens, dot, paren_right, space_paren, brackets, angles
Enumeration	enum_index	number, upper_alpha, lower_alpha, unicode_roman, roman_lower, roman_upper

Table 3: Prompt-format grammar used to generate atomic structural perturbations. Each edit modifies one structural parameter relative to the canonical prompt format while preserving task content. Value sets are preserved from the implemented perturbation grammar.

Family	Base snippet	Perturbed snippet
Separator	Sentence: It was a new pink bike.	Sentence:: It was a new pink bike.
Enumeration	(A) positive (B) negative	[A] positive [B] negative
Spacing	Sentence: ...\nQuestion: ...	Sentence: ... Question: ...
Casing	Sentence: / Question: / Answer:	sentence: / question: / answer:

Table 4: Representative examples of atomic formatting perturbations. These examples are simplified to show the structural edit; full prompts also include the task instruction, one-shot demonstration, evaluation input, and answer field.

C Activation-Distance Targets

This appendix defines the activation-distance targets used in the study. Let $A(p_i)$ denote the prompt-level activation representation for the base prompt and $A(p_i')$ denote the corresponding representation for the perturbed prompt. In the main experiments, $A(p)$ is the final-layer hidden state of the last input token. The distance definitions used below are standard vector-space measures; see Deza & Deza (2009).

C.1 Euclidean Activation Distance

The primary activation-distance target is Euclidean distance:

$$\Delta A_{2,i} = \|A(p_i') - A(p_i)\|_2 . \tag{11}$$

This target captures the overall magnitude of hidden-state displacement induced by a formatting edit.

C.2 Manhattan Activation Distance

As a complementary target, we compute Manhattan distance:

$$\Delta A_{1,i} = \|A(p_i') - A(p_i)\|_1 . \tag{12}$$

C.3 Cosine Distance

We also compute cosine distance:

$$\Delta A_{\cos,i} = 1 - \frac{A(p_i)^\top A(p_i')}{\|A(p_i)\|_2 \, \|A(p_i')\|_2}. \tag{13}$$

C.4 Maximum Absolute Activation Change

Finally, we compute the maximum absolute activation change:

$$\Delta A_{\max,i} = \max_{r\in\{1,\dots,d\}} |A_r(p_i') - A_r(p_i)| . \tag{14}$$

C.5 Log-Transformed Modeling Target

Activation-distance values are non-negative and may be right-skewed. We therefore train predictive models on a log-transformed target (Box & Cox, 1964):

$$y_i = \log(1 + \Delta A_i). \tag{15}$$

Predictions are mapped back to the raw activation-distance scale using:

$$\widehat{\Delta A}_i = \exp(\widehat{y}_i) - 1. \tag{16}$$

Unless otherwise specified, the main predictive experiments use $\Delta A_i = \Delta A_{2,i}$.

C.6 Cross-Model Scale Differences

Raw activation-distance magnitudes should not be interpreted as directly comparable across model families. Models differ in architecture, hidden-state dimensionality, normalization, and representation scale (Kornblith et al., 2019). Consequently, the main cross-model conclusions focus on relative sensitivity patterns within each model.

D Activation–Output Coupling

This appendix provides additional methodological and robustness details for the activation–output coupling analysis reported in the main text. The purpose of this analysis is to determine whether formatting-induced movement in the pre-decoding representation is systematically associated with movement in the task-relevant next-token output distribution. The analysis uses the final-token logits recorded by the original activation extraction pipeline. Importantly, output-side quantities are used only for this coupling analysis; they are not included as predictive inputs to PFSS.

D.1 Eligibility and Analysis Coverage

Output coupling can only be evaluated when the task's semantic answer space can be reconstructed unambiguously at the next-token output level. We therefore evaluate eligibility separately for every task–model stratum. A stratum is retained when the following conditions can be satisfied:

- the task has a well-defined finite semantic label set;
- the semantic labels can be mapped to the model's next-token output space without ambiguity for the analysis;
- base and perturbed intervention files can be aligned exactly;
- the corresponding model output head and recorded final-token logits are available.

Of 212 task–model strata considered, 168 are eligible and successfully analyzed. These strata span 42 tasks and all four models. No stratum deemed eligible subsequently fails analysis. The eligibility restriction is methodological rather than outcome-based: strata are excluded because a task-relevant semantic next-token distribution cannot be reconstructed unambiguously, not because of the observed strength or direction of activation–output coupling.

D.2 Restricted Semantic-Label Distribution

Let $\mathcal{Y}_s$ denote the valid semantic label set for task–model stratum s. For an intervention prompt p_i, let $\ell_i(y)$ denote the recorded next-token logit associated with valid semantic label $y \in \mathcal{Y}_s$. We form a restricted semantic-label distribution by normalizing only over the valid labels:

$$q_i(y) = \frac{\exp(\ell_i(y))}{\sum_{y'\in\mathcal{Y}_s} \exp(\ell_i(y'))}, \qquad y \in \mathcal{Y}_s. \tag{17}$$

Quantity	Value
Task–model strata considered	212
Eligible strata	168
Excluded strata	44
Eligible tasks	42
Models	4
Evaluation examples	41,995
Base–perturbation pairs	5,573,004

Table 5: Coverage of the activation–output coupling analysis. Eligibility is determined from the task label set, tokenizer mapping, exact file alignment, and availability of the corresponding local output head.

For each base–perturbation pair (p_i, p_i'), this yields two task-relevant distributions:

$$q_i^{\text{base}} \quad \text{and} \quad q_i^{\text{pert}}. \tag{18}$$

Restricting the distribution to valid semantic labels avoids conflating the coupling analysis with probability movement among arbitrary vocabulary tokens. The resulting quantity measures how formatting changes redistribute probability within the task's semantic decision space.

D.3 Jensen–Shannon Divergence

We quantify continuous output movement using Jensen–Shannon divergence (JSD). For distributions P and Q, let

$$M = \frac{1}{2}(P + Q). \tag{19}$$

The Jensen–Shannon divergence is

$$\text{JSD}(P, Q) = \frac{1}{2}\,\text{KL}(P\|M) + \frac{1}{2}\,\text{KL}(Q\|M). \tag{20}$$

For intervention i, our output-movement target is therefore

$$J_i = \text{JSD}\left(q_i^{\text{base}}, q_i^{\text{pert}}\right). \tag{21}$$

JSD captures changes in relative probability over valid semantic labels even when the highest-probability label remains unchanged. This is important because decoded-answer agreement can conceal substantial movement in the underlying decision distribution.

D.4 Discrete Behavioral Outcomes

We additionally compute discrete changes in semantic decisions. Let

$$\widehat{y}_i^{\text{base}} = \arg\max_{y \in \mathcal{Y}_s} q_i^{\text{base}}(y), \tag{22}$$

$$\widehat{y}_i^{\text{pert}} = \arg\max_{y \in \mathcal{Y}_s} q_i^{\text{pert}}(y). \tag{23}$$

The answer-flip indicator is

$$F_i = \mathbb{I}\left[\widehat{y}_i^{\text{base}} \neq \widehat{y}_i^{\text{pert}}\right]. \tag{24}$$

For examples with ground-truth label $y_i^\star$, we also define the correct-to-incorrect transition indicator

$$C_i = \mathbb{I}\left[\widehat{y}_i^{\text{base}} = y_i^\star \ \wedge\ \widehat{y}_i^{\text{pert}} \neq y_i^\star\right]. \tag{25}$$

Across task–model strata, the median answer-flip rate is 13.928%, while the median correct-to-incorrect transition rate is 6.631%. The lower incidence of discrete changes relative to continuous JSD movement motivates treating output-distribution displacement and final-answer changes as complementary quantities.

D.5 Within-Stratum Rank Association

Because activation magnitudes and output geometries differ substantially across tasks and model families, the primary nonparametric coupling analysis is performed separately within each task–model stratum. For each eligible stratum s, we compute Spearman rank correlation between Euclidean activation distance $\Delta A_{2,i}$ and semantic-label JSD J_i:

$$\rho_s = \mathrm{Spearman}\left(\Delta A_{2,i}, J_i\right), \qquad i \in s. \tag{26}$$

Of 168 eligible strata, 167 yield a positive Spearman correlation. The median within-stratum correlation is

$$\mathrm{median}_s(\rho_s) = 0.412. \tag{27}$$

This analysis establishes that the monotonic relationship between activation movement and semantic output movement is highly consistent in direction even though its strength differs across tasks and models.

D.6 Activation-Distance Decile Analysis

To visualize the relationship without imposing a parametric functional form, we rank interventions by activation distance within each task–model stratum and partition them into activation-distance deciles. Semantic-label JSD is then summarized within each decile. Across the eligible strata, higher activation-distance deciles generally correspond to larger semantic-label JSD. This analysis provides an intuitive view of the coupling result: interventions that move farther in representation space tend, on average, to induce larger changes in the task-relevant output distribution. The decile construction is performed within strata rather than globally so that model- and task-specific differences in raw activation scale do not determine the ordering.

D.7 Controlled Within-Stratum Regression

A positive raw association could arise because particular formatting transitions simultaneously induce large activation movement and large output movement. We therefore estimate the relationship between activation distance and JSD while controlling for observed properties of the perturbation. Within each eligible task–model stratum s, we fit a regression of the form

$$J_i = \beta_{0,s} + \beta_{A,s}\widetilde{D}_i + \Gamma_s^\top Z_i + \varepsilon_i, \tag{28}$$

where

$$\widetilde{D}_i = \mathrm{standardize}\left[\log(1 + \Delta A_{2,i})\right] \tag{29}$$

is standardized log activation distance and Z_i contains the control variables. The controls include:

- the exact formatting transition;
- absolute token-length change between the base and perturbed prompts;
- the baseline semantic decision margin.

The baseline decision margin captures how strongly the model prefers its leading semantic label before the formatting intervention. This matters because the same amount of output movement may be more likely to flip a decision when the baseline margin is small. Because multiple formatting interventions originate from the same evaluation example, standard errors are clustered by original evaluation example. The fitted activation-distance coefficient $\beta_{A,s}$ is positive in 158 of 168 eligible strata.

D.8 Random-Effects Synthesis

We summarize the controlled stratum-level activation-distance coefficients using a random-effects synthesis. This treatment allows the underlying coupling magnitude to vary across task–model strata rather than assuming one common effect. Of the 168 eligible strata, 165 yield estimable standard errors for inclusion in the pooled synthesis. The resulting pooled standardized activation-distance coefficient is

$$\widehat{\beta}_{\mathrm{pooled}} = 0.0153, \tag{30}$$

Subgroup	Strata	Pairs	$\beta_A > 0$	Pooled $\hat{\beta}$ [95% CI]	I^2 (%)	Median β_A	Flip rate	C→I rate
Falcon	42	1,393,251	42/42	0.0125 [0.0110, 0.0140]	99.2	0.0097	14.04%	4.13%
LLaMA-2	42	1,393,251	39/42	0.0103 [0.0087, 0.0119]	98.2	0.0089	12.98%	5.42%
Mistral	42	1,393,251	42/42	0.0279 [0.0247, 0.0311]	99.1	0.0200	12.89%	7.22%
Qwen-2.5	42	1,393,251	35/42	0.0143 [0.0115, 0.0170]	98.6	0.0122	16.67%	7.82%
Classification	100	–	94/100	0.0187 [0.0170, 0.0204]	98.7	0.0147	–	–
Multiple-choice	68	–	64/68	0.0112 [0.0100, 0.0123]	99.0	0.0091	–	–
All	**168**	**5,573,004**	**158/168**	**0.0153 [0.0143, 0.0162]**	**99.0**	**0.0120**	**13.93%**	**6.63%**

Table 6: Activation–output coupling by model and task type. β_A is the controlled within-stratum coefficient of semantic-label JSD on standardized log Euclidean activation distance, controlling for the exact formatting transition, absolute token-length change, and baseline decision margin. Pooled estimates come from a random-effects synthesis over strata with estimable standard errors (165 of 168; $k = 41$ for Falcon, LLaMA-2, and Mistral, $k = 42$ for Qwen-2.5). Flip and correct-to-incorrect (C→I) rates are medians across task–model strata. All ten negative coefficients occur in Qwen-2.5 (7) and LLaMA-2 (3).

with 95% confidence interval

$$[0.0143, 0.0162]. \tag{31}$$

Between-stratum heterogeneity is high:

$$I^2 \approx 99.0\%. \tag{32}$$

The combination of a consistently positive direction and high heterogeneity is central to our interpretation. Activation distance is informative about output movement across the studied tasks, but the numerical mapping from activation movement to semantic output movement is strongly task- and model-dependent.

D.9 Heterogeneity by Model and Task Type

The high between-stratum heterogeneity reported above motivates a breakdown of the controlled coupling coefficients by model and task type. Table 6 reports, for each subgroup, the number of positive controlled activation-distance coefficients, a random-effects pooled estimate, and the median answer-flip and correct-to-incorrect rates. Coupling is directionally consistent but model-dependent. All ten negative controlled coefficients occur in two models: Qwen-2.5 (7 of 42 strata) and LLaMA-2 (3 of 42 strata). Falcon and Mistral yield positive coefficients in all 42 strata each. Six of the ten negative coefficients are statistically significant at $\alpha = 0.05$; the remaining four are indistinguishable from zero. Pooled coefficients range from 0.0103 (LLaMA-2) to 0.0279 (Mistral), a roughly 2.7-fold difference, and the 95% confidence intervals of every model exclude zero. Coupling is also stronger for classification tasks than for multiple-choice tasks (pooled $\hat{\beta} = 0.0187$ vs. 0.0112). Heterogeneity remains high within every subgroup ($I^2 \geq 98.2\%$), indicating that the between-stratum variation is not explained by model family or task type alone. We therefore interpret the pooled coefficient as a summary of a consistently signed but task-specific relationship, not as a single common effect size.

D.10 Task-Held-Out Prediction of Semantic Output Movement

We next ask whether activation distance provides predictive information about semantic output movement beyond the structural metadata describing the formatting edit. We use grouped task-held-out evaluation, ensuring that interventions belonging to a held-out semantic task do not appear during training. We compare:

1. a metadata-only model using the available formatting metadata; and
2. the same feature set augmented with activation distance.

For semantic-label JSD prediction, the metadata-only model obtains mean

$$R^2 = 0.0989 \tag{33}$$

Input features	Mean R^2	Mean MAE
Formatting metadata only	0.0989	0.0259
Metadata + activation distance	0.1085	0.0260

Table 7: Grouped task-held-out prediction of restricted semantic-label Jensen–Shannon divergence. Adding activation distance increases explained variance, while MAE remains essentially unchanged.

Input features	AUC	Log loss
Formatting metadata only	0.8082	0.5455
Metadata + activation distance	0.8128	0.5399

Table 8: Grouped task-held-out prediction of answer flips. Adding activation distance provides a modest improvement in both discrimination and probabilistic prediction.

and

$$\mathrm{MAE} = 0.0259. \tag{34}$$

After adding activation distance, mean performance becomes

$$R^2 = 0.1085 \tag{35}$$

and

$$\mathrm{MAE} = 0.0260. \tag{36}$$

Thus, activation distance increases explained variance under task-held-out evaluation, although MAE does not improve. We therefore characterize the incremental predictive contribution as modest rather than uniformly beneficial across all metrics.

D.11 Task-Held-Out Prediction of Answer Flips

We conduct an analogous held-out analysis for the binary answer-flip target. Adding activation distance to formatting metadata increases mean AUC from

$$0.8082 \quad \text{to} \quad 0.8128, \tag{37}$$

and improves log loss from

$$0.5455 \quad \text{to} \quad 0.5399. \tag{38}$$

The primary held-out metric improves in 9 of 10 folds. These results show that activation movement contains information about answer instability that is not entirely captured by the structural description of the formatting intervention.

D.12 Robustness to Semantic-Label Token Mapping

Different formatting conditions may interact with tokenizer behavior. To test whether the observed coupling is driven by changes in the mapping between semantic labels and their output tokens, we repeat the controlled analysis on a restricted set of interventions for which the semantic-label token mapping is invariant between the compared conditions. Under this restriction, 161 of 168 controlled activation-distance coefficients remain positive. Furthermore, the restricted-analysis coefficients and the corresponding coefficients from the full analysis are strongly aligned:

$$\rho \approx 0.954. \tag{39}$$

This robustness check shows that the activation–output coupling result is not explained solely by changes in the semantic-label token mapping.

D.13 Summary of Coupling Evidence

Table 9 summarizes the main activation–output findings.

Analysis	Result
Eligible coverage	168/212 task–model strata; 42 tasks; 4 models; 5,573,004 prompt pairs; 41,995 examples
Within-stratum rank association	167/168 Spearman correlations positive; median $\rho = 0.412$
Controlled within-stratum regression	158/168 activation-distance coefficients positive
Random-effects synthesis	pooled standardized coefficient 0.0153, 95% CI $[0.0143, 0.0162]$; $I^2 \approx 99.0\%$
Median answer-flip rate	13.928%
Median correct-to-incorrect rate	6.631%
Task-held-out JSD prediction	R^2: $0.0989 \rightarrow 0.1085$; MAE: $0.0259 \rightarrow 0.0260$
Task-held-out answer-flip prediction	AUC: $0.8082 \rightarrow 0.8128$; log loss: $0.5455 \rightarrow 0.5399$; improvement in 9/10 folds
Invariant semantic-label token mapping	161/168 coefficients positive; full vs. restricted coefficient correlation $\rho \approx 0.954$

Table 9: Summary of the activation–output coupling evidence. The relationship is highly consistent in direction but heterogeneous in magnitude, motivating the interpretation of activation distance as an intermediate audit signal rather than a deterministic failure metric.

E PFSS Derivation and Predictive Modeling

This appendix provides the mathematical details of the Pruned Feature-Specific Sensitivity predictor (PFSS) and the predictive baselines used in the paper. The prediction task is to estimate activation movement from prompt-format metadata alone.

E.1 Prediction Problem

For each intervention instance i, let x_i denote the prompt-format metadata vector and let ΔA_i denote the activation-distance target. Predictive models are trained on:

$$y_i = \log(1 + \Delta A_i). \tag{40}$$

The prediction objective is:

$$\widehat{y}_i = f(x_i), \tag{41}$$

with raw-scale predictions recovered as:

$$\widehat{\Delta A}_i = \exp(\widehat{y}_i) - 1. \tag{42}$$

The input vector x_i contains only structural prompt-format metadata. It excludes raw prompt text, prompt identifiers, task identifiers, hidden states, logits, generated outputs, and activation targets other than the target being predicted.

E.2 Feature-Specific Expansion

For a continuous metadata feature x_{ij}, such as base prompt token length, perturbed prompt token length, or token-length difference, we standardize using training-set statistics (Hastie, 2009):

$$z_{ij} = \frac{x_{ij} - \mathrm{median}(x_j)}{\mathrm{sd}(x_j) + \epsilon}, \tag{43}$$

where ϵ is a small numerical stability constant. The standardized value is clipped:

$$\widetilde{z}_{ij} = \mathrm{clip}(z_{ij}, -8, 8). \tag{44}$$

For each continuous feature, the expanded library includes:

$$\phi_{j,1}(x_i) = \widetilde{z}_{ij}, \tag{45}$$
$$\phi_{j,2}(x_i) = \widetilde{z}_{ij}^2, \tag{46}$$
$$\phi_{j,3}(x_i) = \widetilde{z}_{ij}^3, \tag{47}$$
$$\phi_{j,4}(x_i) = \log(1 + |\widetilde{z}_{ij}|), \tag{48}$$
$$\phi_{j,5}(x_i) = \operatorname{sign}(\widetilde{z}_{ij})\sqrt{|\widetilde{z}_{ij}|}. \tag{49}$$

We also include local radial-basis terms (Broomhead & Lowe, 1988; Bishop & Nasrabadi, 2006):

$$\phi_{j,c}^{\mathrm{RBF}}(x_i) = \exp\left(-\gamma(\widetilde{z}_{ij} - c)^2\right), \qquad c \in \{-1, 0, 1\}. \tag{50}$$

Categorical variables are represented using one-hot indicators:

$$\phi_{j,c}(x_i) = \mathbb{I}[x_{ij} = c]. \tag{51}$$

After expansion, all terms are flattened into:

$$\Phi(x_i) = [\phi_1(x_i), \phi_2(x_i), \ldots, \phi_M(x_i)]. \tag{52}$$

E.3 PFSS Model

PFSS is a compact selected-term predictor:

$$\widehat{y}_i = \beta_0 + \sum_{j \in S} \beta_j \phi_j(x_i), \qquad |S| = K. \tag{53}$$

In the main experiments, $K = 25$.

E.4 ElasticNet Term Selection

PFSS first fits an ElasticNet selector over the full expanded library (Zou & Hastie, 2005):

$$\widehat{\beta}^{\mathrm{sel}} = \arg\min_{\beta} \left[\frac{1}{2n} \|y - \Phi\beta\|_2^2 + \lambda \left(\alpha\|\beta\|_1 + \frac{1-\alpha}{2}\|\beta\|_2^2 \right) \right]. \tag{54}$$

Terms are ranked by:

$$r_j = |\widehat{\beta}_j^{\mathrm{sel}}|, \tag{55}$$

and the top K are retained:

$$S = \mathrm{TopK}(r_1, \ldots, r_M). \tag{56}$$

E.5 Selected-Term Refit

If the selected design matrix is numerically stable, ordinary least squares is used:

$$\widehat{\beta}_S^{\mathrm{OLS}} = (\Phi_S^\top \Phi_S)^{-1} \Phi_S^\top y. \tag{57}$$

Otherwise, Ridge refitting is used:

$$\widehat{\beta}_S^{\mathrm{Ridge}} = (\Phi_S^\top \Phi_S + \lambda I)^{-1} \Phi_S^\top y. \tag{58}$$

E.6 Predictive Baselines

Mean baseline.

$$\widehat{y}_i = \overline{y}_{\mathrm{train}}. \tag{59}$$

Algorithm 1 PFSS Training Procedure

Require: Metadata matrix X, target vector y, feature library $\Phi(\cdot)$, number of selected terms K
Construct expanded feature matrix $\Phi(X)$
Fit ElasticNet on $\Phi(X)$ to obtain $\widehat{\beta}^{\mathrm{sel}}$
Compute selector scores $r_j = |\widehat{\beta}_j^{\mathrm{sel}}|$
Select the top K terms: $S = \mathrm{TopK}(r_1, \ldots, r_M)$
Construct selected design matrix $\Phi_S(X)$
if $\Phi_S^\top \Phi_S$ is numerically stable **then**
    Refit selected terms using ordinary least squares
else
    Refit selected terms using Ridge regression
end if
return selected terms S and fitted coefficients $\widehat{\beta}_S$

Raw one-hot Ridge.

$$\widehat{\beta}^{\mathrm{ridge}} = \arg\min_{\beta} \left[\|y - X_{\mathrm{raw}}\beta\|_2^2 + \lambda \|\beta\|_2^2 \right]. \tag{60}$$

Raw ElasticNet. Raw ElasticNet uses the same raw encoded design matrix with combined ℓ_1 and ℓ_2 regularization.

Global RBF Random-Feature Ridge. The raw metadata are mapped into random Fourier features approximating an RBF kernel (Rahimi & Recht, 2007), after which Ridge regression is fitted.

Full feature-specific Ridge. This baseline fits Ridge regression over the entire expanded feature-specific library.

Black-box regressors. Random Forest and Histogram Gradient Boosting are included as flexible nonlinear baselines (Breiman, 2001; Ke et al., 2017).

E.7 Validation Splits

Random split. The random split uses an 80/20 train–test division over intervention instances.

Task-holdout split. For task-holdout validation, all rows from held-out tasks are assigned to the test set. Task identifiers are used only to form the split and are not included as predictive features.

E.8 Evaluation Metrics

We report R^2, RMSE, and MAE (James et al., 2013) on both the log-transformed target and raw activation-distance scale.

E.9 Interpretation of PFSS

PFSS is intended as a compact diagnostic predictor rather than a mechanistic explanation. Its selected terms identify structural predictors of activation sensitivity but do not identify neurons, attention heads, circuits, or causal pathways.

E.10 Full Predictive Performance Table

Table 10 reports the complete set of results summarized in Figure 2 of the main text, including both random-split and task-holdout evaluation, and both raw-scale and log-scale R^2.

Model	Method	Random Raw R^2	Random Log R^2	Task Raw R^2	Task Log R^2
Falcon	Mean baseline	-0.090	0.000	-0.075	-0.004
Falcon	Raw one-hot Ridge	0.913	0.858	0.876	0.813
Falcon	Full FS Ridge	0.915	0.869	0.884	0.831
Falcon	PFSS	0.892	0.791	0.831	0.731
Falcon	Global RBF	0.077	-0.011	-0.047	-0.196
Falcon	Random Forest	0.960	0.912	0.900	0.826
Falcon	HistGradientBoosting	0.956	0.914	0.911	0.842
LLaMA-2	Mean baseline	-0.104	0.000	-0.100	-0.001
LLaMA-2	Raw one-hot Ridge	0.793	0.700	0.807	0.634
LLaMA-2	Full FS Ridge	0.818	0.732	0.769	0.715
LLaMA-2	PFSS	0.781	0.653	0.773	0.656
LLaMA-2	Global RBF	0.179	0.074	0.011	-0.002
LLaMA-2	Random Forest	0.896	0.801	0.885	0.685
LLaMA-2	HistGradientBoosting	0.889	0.806	0.862	0.716
Mistral	Mean baseline	-0.096	0.000	-0.091	0.000
Mistral	Raw one-hot Ridge	0.906	0.823	0.882	0.762
Mistral	Full FS Ridge	0.915	0.830	0.892	0.768
Mistral	PFSS	0.879	0.722	0.852	0.677
Mistral	Global RBF	0.170	-0.050	0.014	-0.059
Mistral	Random Forest	0.962	0.879	0.919	0.751
Mistral	HistGradientBoosting	0.956	0.877	0.930	0.763
Qwen-2.5	Mean baseline	-0.090	0.000	-0.100	-0.001
Qwen-2.5	Raw one-hot Ridge	0.852	0.762	0.824	0.717
Qwen-2.5	Full FS Ridge	0.871	0.789	0.837	0.742
Qwen-2.5	PFSS	0.860	0.730	0.821	0.700
Qwen-2.5	Global RBF	0.195	0.067	-0.083	-0.090
Qwen-2.5	Random Forest	0.936	0.860	0.896	0.764
Qwen-2.5	HistGradientBoosting	0.932	0.855	0.905	0.797

Table 10: Predictive performance on Euclidean activation distance ΔA_2 under random and task-holdout evaluation. Raw R^2 is computed on the original activation-distance scale, while log R^2 is computed on the log-transformed modeling target. Full FS Ridge denotes Ridge regression over the full feature-specific expanded library. PFSS denotes the 25-term Pruned Feature-Specific Sensitivity predictor. Global RBF denotes a random Fourier feature approximation to an RBF kernel followed by Ridge regression. Negative R^2 values indicate performance below the mean predictor of the evaluated split and can occur when a model trained on the training distribution generalizes poorly to the evaluated split. See Figure 2 in the main text for a visual summary of the task-holdout raw-scale column.

Model	Cosine R^2	Max-Abs R^2
Falcon	0.825	0.925
LLaMA-2	0.862	0.867
Mistral	0.928	0.941
Qwen-2.5	0.844	0.771

Table 11: PFSS robustness across alternative activation-distance targets under random-split evaluation.

F Additional Results and Robustness

F.1 Alternative Activation-Distance Targets

The main analysis uses Euclidean distance. We additionally evaluate PFSS on cosine distance and maximum absolute activation change.

F.2 Selected-Term Stability

Separator-related terms appear prominently among stable selected features. The same structural features that induce large mean activation shifts also recur as stable predictors in PFSS, reinforcing the conclusion that activation movement is partially recoverable from transparent structural features.

F.3 Activation–Output Robustness Summary

The activation–output findings are supported by several complementary analyses rather than a single statistic. The near-universal positive within-stratum rank association demonstrates monotonic coupling, while the controlled regressions show that the relationship persists after accounting for the formatting transition, token-length change, and baseline decision margin. Task-held-out prediction tests whether activation distance adds information beyond formatting metadata, and the invariant-token-mapping restriction checks that the main relationship is not attributable solely to changes in semantic label tokenization. Full details are given in Appendix D.

G Implementation and Reproducibility

G.1 Models and Tokenization

The experiments use four open-weight 7B-scale language models: Falcon, LLaMA-2, Mistral, and Qwen-2.5. Prompts are tokenized with each model's corresponding tokenizer. Base and perturbed prompts are processed separately so that each activation vector is extracted from the tokenization induced by that specific prompt format. Prompts are truncated to a maximum sequence length of 2048 tokens.

G.2 Hidden-State and Logit Extraction

For each prompt $p = (w_1, \ldots, w_T)$, we tokenize the prompt with the corresponding model tokenizer and run a forward pass with hidden-state outputs enabled. The base and perturbed prompts are processed independently. A Transformer language model with L layers and hidden dimension d produces:

$$H_\ell(p) \in \mathbb{R}^{T \times d}, \qquad \ell \in \{1, \ldots, L\}. \tag{61}$$

The primary prompt representation is:

$$A(p) = h_L^{(T)}(p). \tag{62}$$

The pipeline additionally records final-token logits. These logits are not used as PFSS inputs; they are retained for output-side analyses and are used to construct the restricted semantic-label distributions in Appendix D.

G.3 Preprocessing and Hyperparameters

Before model fitting, rows with non-finite values for the current activation-distance target are removed. Numeric predictors are preprocessed using median imputation followed by standardization. Categorical predictors use most-frequent imputation followed by one-hot encoding. Unseen test categories are ignored, and rare categorical values are grouped using a minimum frequency threshold of 5. Raw one-hot Ridge uses `alpha=1.0`. Raw ElasticNet uses `alpha=0.002`, ℓ_1 ratio 0.3, and a maximum of 5000 iterations. The global RBF baseline uses $\gamma = 0.5$ and 300 random Fourier components, followed by Ridge regression with `alpha=1.0`. PFSS uses an ElasticNet selector with `alpha=0.002`, ℓ_1 ratio 0.75, and a maximum of 5000 iterations. The top $K = 25$ terms are retained and refit using OLS when numerically stable and Ridge with `alpha=1.0` otherwise.

G.4 Split Realization

Random-split evaluation uses an 80/20 train–test partition with fixed seed 42. Task-holdout evaluation uses group-based splitting so that all rows associated with a held-out task are assigned to the test set, preventing within-task leakage. The activation–output predictive analysis similarly uses

grouped task-held-out evaluation, ensuring that the test tasks are not represented in the corresponding training partition.

G.5 Reproducibility Package

An anonymized reproducibility package is available at this link. The repository contains the Colab notebook, perturbation grammar, task identifiers, split assignments, processed prompt-format metadata, derived activation-distance targets, and PFSS selected terms. The full activation-extraction pipeline produced approximately 200GB of intermediate artifacts across four model-specific runs. These raw hidden-state artifacts are not required to reproduce the aggregate analyses once activation distances have been computed. The processed metadata and derived targets are therefore released rather than the complete raw hidden-state cache.