

# Measuring the Measure: Mechanistic Prompt Sensitivity for LLM-Based Populism Coding

Stanford CS224N Custom Project

**Jiehan Liu**

Department of Political Science  
Stanford University  
jiehanl@stanford.edu

## Abstract

Large language models are increasingly used as text coders, but the resulting measurement can depend on prompt wording. I study that dependence in binary populism coding of speeches from the Global Populism Database using 93 meaning-preserving rewrites of a fixed rubric and an open-weight Llama 3 model forced to answer with a single binary answer. This design lets me audit prompt sensitivity at the level of hard labels, probability scores, prompt-only priors, and hidden-state drift. Hard label flips are rare: only one speech in the analyzed sample changes label under any non-baseline prompt. Score-level movement is much more common: 56.9% of speeches show at least a one-point spread in  $p(y=1)$  across prompts, 47.1% at least two points, 25.5% at least five points, and 9.8% at least ten points. Meaning-preserving rewrites shift the model's null-input prior by several tenths of a logit, indicating that wording changes the measurement instrument before the speech is read. Prompt sensitivity is concentrated among borderline speeches with low baseline margins—cases where the baseline prompt  $z_0$  produces a logit margin near zero and the model is only weakly committed to either label. Ranking speeches by the absolute baseline margin  $|\delta(x, z_0)|$  shows that auditing roughly the lowest-margin 20% recovers essentially all top-decile sensitive cases. Internal representation drift is weak as an average-case ranking signal, although the highest-risk speeches still display structured rewrite-specific drift patterns. In an auxiliary speech-level auditing benchmark, simple baseline-prompt features, especially margin, probability, entropy, and related prior/evidence summaries under  $z_0$ , outperform internal-only predictors based on representation drift (ROC-AUC 0.62 vs. 0.24).

## 1 Key Information to include

- Mentor: Nevin George
- External Collaborators: None
- Sharing project: POLISCI 450D Political Methodology IV
- Late Days: None

## 2 Introduction

Large language models (LLMs) are increasingly used as measurement tools in the social sciences. Instead of training a task-specific classifier, a researcher can write a natural-language coding rubric and ask a model to assign a label or score to a text. That workflow is attractive because it is flexible, cheap, and easy to align with substantive coding rules. It also introduces a reliability problem. If two prompts mean the same thing to a human reader, then small wording edits should not substantially

change the measurement. When they do, the wording of the rubric becomes part of the measurement error.

This paper studies that problem in a substantively meaningful setting: binary coding of populist discourse in political speeches. I treat prompt sensitivity not just as a nuisance of prompting, but as a local failure of measurement invariance. The question is not only whether prompt wording changes the output, but how it changes the output. Does it primarily shift the model’s class prior before the speech is read? Does it change the text-conditioned evidence extracted from the speech? Or does it alter internal representations in ways that can help identify fragile cases before the full prompt bank is run?

The answer is mixed. Meaning-preserving rewrites rarely flip the final hard label in the experiment samples, but they often move the underlying probability. Some of that movement is a prompt-only prior shift: rewrites change the model’s null-input logit even before any speech content is shown. The speeches that move the most are concentrated among low-margin cases, so a cheap output-based audit rule is already highly effective. Internal drift is more informative qualitatively than quantitatively in this run: it reveals structured prompt-specific signatures for high-risk speeches, but it is weaker than simple output-only features as a ranking signal.

The paper makes three contributions. First, it introduces a controlled prompt audit, following [1] for LLM-based populism coding using a factorial neighborhood of 93 meaning-preserving prompt variants. Second, it decomposes prompt sensitivity into prompt-only bias, text-conditioned evidence, and representation drift using a weight-access open model. Third, it evaluates whether inexpensive audit signals can prioritize the speeches most worth rechecking, and shows that baseline output geometry is already a strong triage signal.

### 3 Related Work

This project is motivated most directly by work on prompt perturbation [1, 2]. [1] show that LLM behavior can vary substantially across meaning-preserving prompt formatting changes, and argue that prompt choice should be treated as a source of uncertainty rather than ignored. Their perspective is primarily black-box and output-focused. I keep the same reliability concern, but use an open-weight model so that I can inspect logit margins, prompt-only bias terms, and hidden-state drift in addition to final labels.

The project also connects to political science work that treats populism as a measurable discourse rather than a bag of keywords. [3] develops the ideational approach to populist discourse and shows that human coders can evaluate speeches holistically. The Global Populism Database extends that logic to a cross-national corpus of chief-executive speeches and provides the substantive task context for this study [3]. My goal is not to replace that human infrastructure with an unaudited LLM. Instead, I ask how stable an LLM-based rubric is under local meaning-preserving rewrites and what signals identify speeches that deserve additional audit.

At the methodological level, the paper is also related to hidden-state literature which show that model internals can reveal information that is not obvious from the final output alone [4, 5]. Rather than treating hidden states as a new predictor to optimize for task accuracy, I use them as a diagnostic: if prompt sensitivity corresponds to structured changes in internal representations, then those changes may help distinguish fragile cases from stable ones. This diagnostic perspective is especially useful in measurement settings, where the goal is not only to maximize performance but also to understand when the measurement procedure itself is unstable.

## 4 Approach

### 4.1 Task formulation

Let  $x$  denote a speech excerpt and let  $y \in \{0, 1\}$  indicate whether the excerpt contains a coherent populist frame. The prompt  $z$  serves as a coding rubric. I force the model to output exactly one digit 0 or 1, at a fixed `Answer :` position, which yields the logit margin

$$\delta(x, z) = \log p_{\theta}(y=1 \mid x, z) - \log p_{\theta}(y=0 \mid x, z). \quad (1)$$

The corresponding forced-choice probability is  $p_\theta(y=1 \mid x, z) = \sigma(\delta(x, z))$ . Because the label is read directly from logits at a fixed answer slot, the main results are not driven by sampling temperature or decoding randomness.

## 4.2 Meaning-preserving prompt neighborhood

The prompt bank is constructed by applying controlled rewrites to a fixed populism-coding template (see Appendix for the full prompt). The runner defines eight rewrite sites, each corresponding to a local verbalization choice inside the instruction: *community/group*, *ordinary/common*, *explicit/direct*, *morally bad/morally corrupt*, *struggle/conflict*, *good vs evil/good versus evil*, *opposed to/set against*, and *shows/contains*. Each prompt variant is represented by a bitmask over those sites. The analyzed run includes all variants with up to three simultaneous rewrites,

$$\sum_{k=0}^3 \binom{8}{k} = 93, \quad (2)$$

producing a local neighborhood of prompts intended to preserve the underlying construct while perturbing the wording of the rubric.

## 4.3 Prior-evidence decomposition

To separate prompt sensitivity into prompt-only and text-conditioned components, I estimate the prompt-only bias on a null input  $x_{\text{null}}$  (implemented as N/A):

$$b(z) = \delta(x_{\text{null}}, z). \quad (3)$$

I then define the residual evidence term as

$$e(x, z) = \delta(x, z) - b(z). \quad (4)$$

A useful centered quantity is the prior shift relative to the baseline prompt  $z_0$ ,

$$\text{bias\_shift}(z) = b(z) - b(z_0). \quad (5)$$

If prompt rewrites were purely stylistic in the strongest sense, then  $\text{bias\_shift}(z)$  would be close to zero for all variants.

## 4.4 Representation drift and audit signals

For a baseline prompt  $z_0$  and a perturbed prompt  $z$ , I extract mean-pooled hidden representations over the speech span and compute layer-wise drift

$$d_\ell(x, z) = 1 - \cos(h_\ell(x, z_0), h_\ell(x, z)). \quad (6)$$

I summarize drift using an AUC-style mean over layers together with mid-layer and last-layer summaries. These internal features support a second question: can prompt sensitivity be predicted without exhaustively running the full prompt bank? To test that, I compare output-only audit signals (such as signed and absolute margin, baseline  $p(y = 1)$ , entropy,) against internal-feature predictors built from a small set of rewrite prompt comparisons.

# 5 Experiments

## 5.1 Data

The source corpus is the Global Populism Database (GPD), which contains chief-executive speeches coded for populist discourse [3]. The analyzed dataset contains 50 speech texts and 93 prompt variants, for 4,650 speech-prompt evaluations. The rubric maps a speech to  $y=1$  when it contains a coherent populist frame: a unified and morally virtuous people opposed to a broadly bad elite in moral good-evil terms. Otherwise the speech receives  $y=0$ .

## 5.2 Evaluation method

The main paper uses two speech-level sensitivity summaries. The first is the probability spread across prompt variants,

$$S_{\text{spread}}(x) = \max_z p_{\theta}(y=1 \mid x, z) - \min_z p_{\theta}(y=1 \mid x, z), \quad (7)$$

which captures how much the score moves even when the hard label stays fixed. The second is a hard-flip indicator,

$$S_{\text{flip}}(x) = \mathbb{1}\{\exists z : \hat{y}(x, z) \neq \hat{y}(x, z_0)\}, \quad (8)$$

which records whether any non-baseline prompt changes the baseline decision. I also report variance-based summaries such as  $\text{Var}_z[\delta(x, z)]$  and  $\text{Var}_z[e(x, z)]$  for decomposition analyses.

For targeted auditing, I use two complementary evaluations. The audit-lift curves in Figure 3 target text in the top decile of the probability-spread distribution. The auxiliary cross-validated predictor benchmark in Figure 5 uses a denser binary target because hard flips are too sparse for stable model comparison: a speech is labeled *sensitive* if its prompt-induced margin variance lies in the top 20% of the sample (equivalently, `variance_above_quantile` with `delta_variance` at the 80th percentile). I report out-of-fold ROC-AUC, average precision, and precision at fixed audit budgets.

## 5.3 Experimental details

Inference uses `meta-llama/Meta-Llama-3-8B-Instruct` through a local factorial prompt runner. The model is queried in chat mode and evaluated deterministically by reading logits at the fixed `Answer:` position rather than by sampling free-form continuations. For each prompt, the runner first computes the prompt-only bias  $b(z)$  on the null input, then evaluates every speech-prompt pair to obtain  $\delta(x, z)$ ,  $e(x, z)$ , forced-choice probabilities and entropies, predicted labels, and flip indicators. It also extracts mean-pooled hidden states across layers to compute drift relative to the baseline prompt. All runner and analysis code used in the paper was written for this project.

# 6 Results

## 6.1 Hard flips are rare, but score movement is common

Figure 1 shows the central empirical pattern. Hard label flips are rare in this run: only one speech has any non-baseline prompt that changes the baseline label. At the same time, score-level movement is common. The median speech has a 1.7 percentage-point max-min spread in  $p(y=1)$  across prompts, 56.9% of speeches have at least a one-point spread, 47.1% have at least two points, 25.5% have at least five points, and 9.8% have at least ten points. The distribution has a clear long tail extending to roughly 15 percentage points.

This distinction matters. A flip-rate summary alone would suggest that the prompt bank barely changes the measurement. The probability spread shows otherwise: many speeches move materially even when they remain on the same side of the decision boundary. In a measurement pipeline, that movement can affect thresholding, ranking, downstream aggregation, and the decision of which speeches deserve human review.

## 6.2 Meaning-preserving rewrites shift the prompt-only prior

Figure 2 shows that prompt rewrites move the null-input prior relative to the baseline prompt. The distribution of `bias_shift(z)` is centered near zero but not degenerate; visually, shifts span roughly  $[-0.2, 0.2]$  logit units. That is large enough to matter for borderline speeches. Some multi-rewrite prompts show wider dispersion than one-rewrite prompts, suggesting that local wording changes can accumulate.

Prompt variation is not only changing how the model reads the speech. It is also changing the measurement instrument itself before the speech is read. Two rubric variants that a human coder would regard as equivalent can place the same speech on slightly different baseline scales.

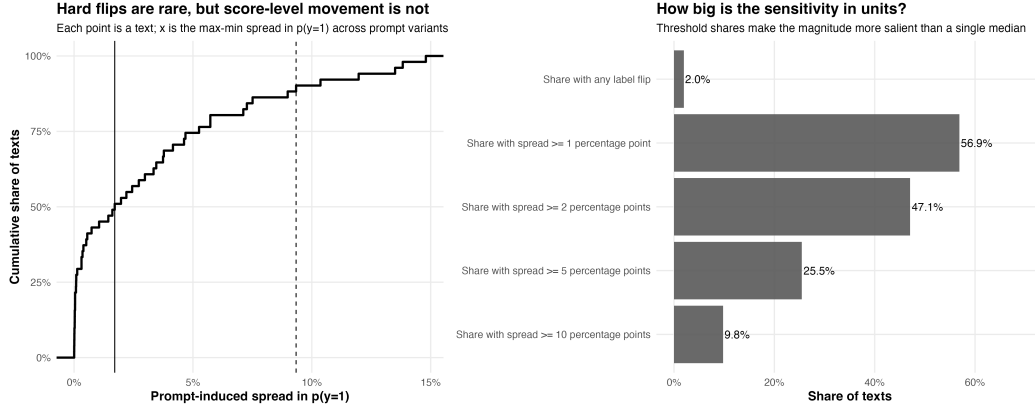


Figure 1: Prompt sensitivity is uncommon on the hard-label scale but common on the score scale. Left: empirical CDF of max-min spread in  $p(y=1)$  across prompt variants for each speech. Right: threshold shares that make the magnitude of score movement more interpretable.

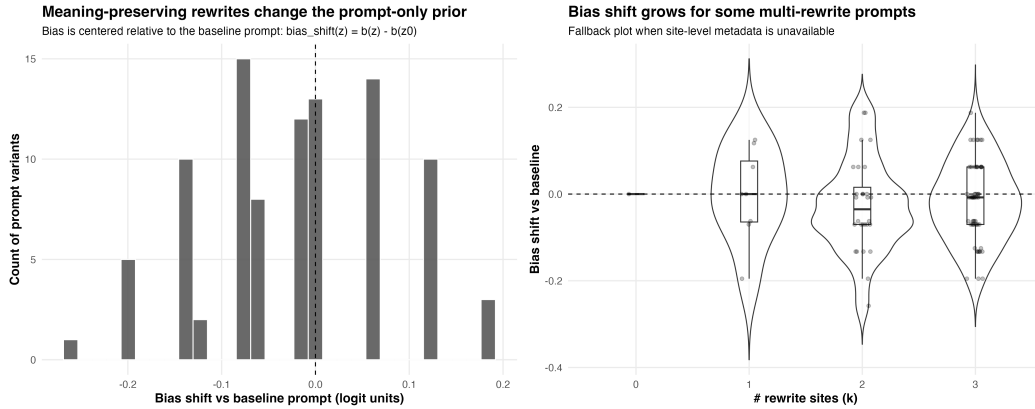


Figure 2: Meaning-preserving prompt rewrites change the prompt-only prior. Left: histogram of bias shifts relative to the baseline prompt. Right: bias shifts stratified by the number of rewrite sites changed.

### 6.3 Prompt sensitivity concentrates on knife-edge speeches

Figure 3 shows that low-margin speeches (speeches for which the baseline prompt yields a logit margin close to zero) absorb most of the sensitivity. In the lowest-margin decile, the 90th percentile of prompt-induced score spread is about 11%; by the most confident deciles, the corresponding spread is essentially zero. If we audit the speeches that the model is least confident about under the baseline prompt first, we find the most prompt-sensitive cases much faster than if we prioritize speeches by internal drift or entropy. In this run, checking roughly the 20% of speeches with the smallest absolute baseline margins was enough to identify essentially all speeches in the top 10% of prompt sensitivity.

This is the most directly useful practical result in the paper. Even without a complete mechanistic account, a cheap output-only statistic already tells the researcher where additional prompt checks are most valuable.

### 6.4 Internal drift is weak on average, but structured in high-risk cases

The main drift summary in Figure 4 is modest. For one-atom prompt variants, the rank correlation between drift and absolute probability movement is Spearman  $\rho = -0.06$ , which is effectively no monotonic relationship. The decile plot nevertheless hints that the highest-drift prompt comparisons

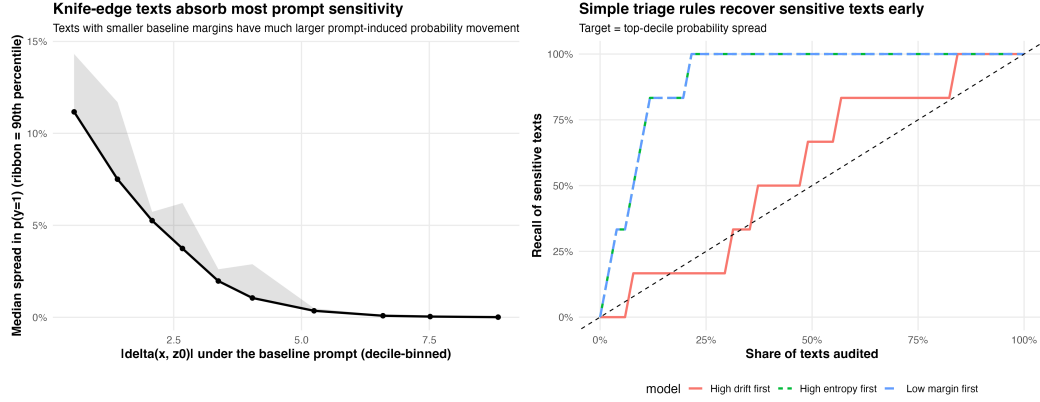


Figure 3: Low-margin speeches are the most prompt-sensitive. Left: decile-binned baseline margin versus prompt-induced probability spread. Right: audit-lift curves for simple triage rules, where the target is membership in the top decile of the spread distribution.

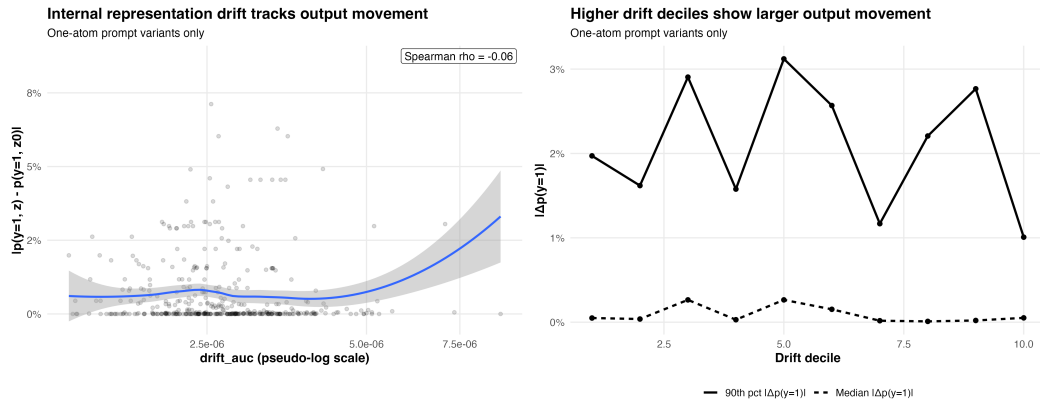


Figure 4: Average-case association between internal drift and output movement is weak. The left panel shows the one-atom scatter with Spearman  $\rho = -0.06$ ; the right panel summarizes median and 90th-percentile output movement by drift decile.

have somewhat larger tail movements even if the median movement stays near zero. I therefore interpret internal drift as a weak average-case predictor but a potentially useful tail-risk descriptor.

That interpretation is supported qualitatively by the rewrite-drift heatmap in Appendix Figure 8. The highest-risk speeches do not show uniform elevation across all rewrite prompts. Instead they exhibit structured, prompt-specific drift signatures: some speeches light up strongly on a small subset of rewrites, while others are more diffuse. This pattern suggests that the internal mechanism of sensitivity is not simply “more drift everywhere.” It is often tied to particular wording changes.

## 6.5 Output-only predictors outperform internal-only predictors in the auditing benchmark

Figure 5 summarizes the auxiliary speech-level targeted-auditing benchmark. On the top-quintile margin-variance target described in Section 4.2, the best-performing family is `output_only` with ROC-AUC 0.62 and average precision 0.26. The constant-prevalence baseline is at chance (ROC-AUC 0.50), while `internal_only` falls well below chance (0.24). The combined `output_plus_internal` model also underperforms the simpler output-only model.

In this run, output geometry around the baseline prompt, especially margin and entropy, is more useful for ranking audit risk than the current rewrite-based internal features. That does not make

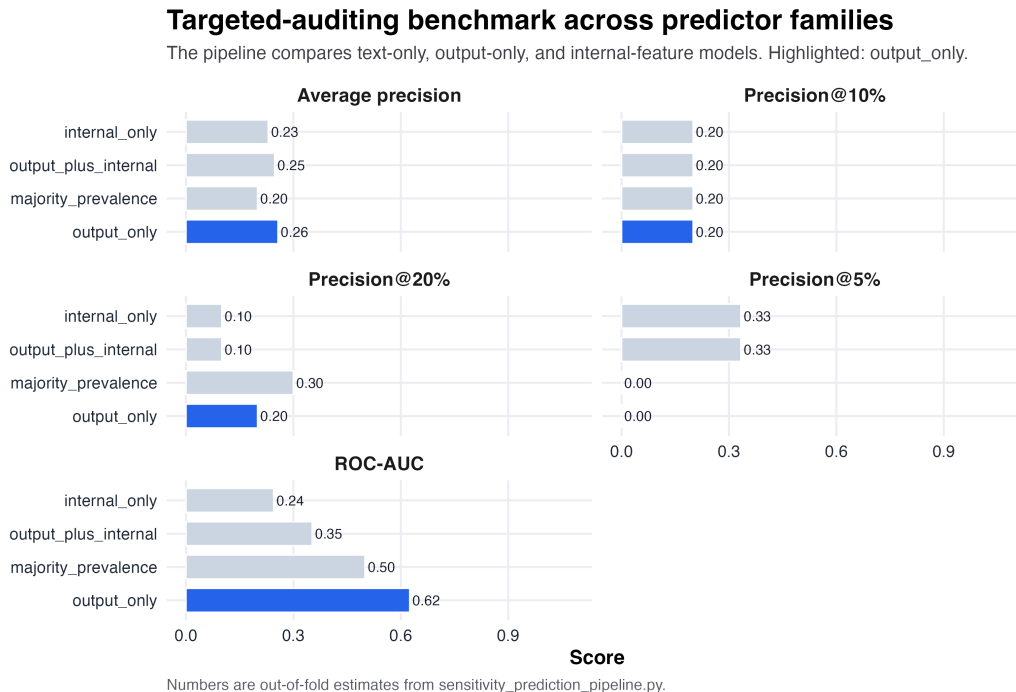


Figure 5: Auxiliary speech-level targeted-auditing benchmark. A speech is labeled sensitive if its prompt-induced margin variance falls in the top 20% of the sample. The optional text-only TF-IDF baseline defined in the pipeline is not shown because the exported run did not include raw speech text for every speech.

the drift analysis irrelevant, but it does mean that internal features are presently more valuable as a qualitative diagnostic than as a stand-alone predictor.

## 7 Analysis

Three conclusions are robust across the main figures.

First, prompt sensitivity in this setting is mainly a *score movement* problem rather than a hard-label instability problem. The 2.0% flip rate sounds reassuring, but it hides the fact that one quarter of speeches move by at least five percentage points across prompt variants.

Second, part of this movement is structured as a prompt-only prior shift. Figure 2 shows that the prompt wording changes the null-input logit even before the speech is read. This supports the broader measurement perspective: prompt wording is not just an interface choice; it is part of the measurement instrument. At the same time, the decomposition analysis in Appendix Figure 6 suggests that prompt-only bias is not the whole story. For many unstable speeches, substantial across-prompt variation remains after removing the null-input bias, indicating that prompt sensitivity cannot be reduced to a speech-invariant prior shift alone<sup>1</sup>.

Third, simple output-level signals dominate the audit-triage story in the current run. The low-margin heuristic is highly effective, and the output-only predictor family beats the internal-only alternatives. That is a useful negative result as well as a practical positive one. It suggests that the first stage of an audit pipeline should be cheap and black-box: use baseline margin and entropy to prioritize

<sup>1</sup>Think about the variance decomposition  $\text{Var}_z[\delta(x, z)] = \text{Var}_z[b(z)] + \text{Var}_z[e(x, z)] + 2\text{Cov}_z(b(z), e(x, z))$ . Because of that covariance term, subtracting  $b(z)$  does not necessarily have to reduce variance. If the prompt-only bias and the residual evidence move in opposite directions across prompts, then  $\text{Var}_z[e(x, z)]$  can even be larger than  $\text{Var}_z[\delta(x, z)]$ .

cases. Mechanistic features can then be added as a secondary diagnostic rather than assumed to be universally superior.

First, the strongest evidence that prompt sensitivity is concentrated in borderline speeches comes from the probability-spread definition of sensitivity in Figure 3; when sensitivity is instead measured by margin variance, the pattern is weaker and less linear, showed in Appendix Figure 7. Second, the prior-correction plot in Appendix Figure 6 is useful for description, but its summary shrink statistic is not yet trustworthy enough to support a strong numerical claim. So the main practical result is solid, but some of the more ambitious mechanistic and decomposition claims should be stated more cautiously.

## 8 Conclusion

This paper studies prompt sensitivity as a measurement problem for LLM-based populism coding. In the analyzed run, meaning-preserving prompt rewrites rarely flip the final hard label, but they often move the underlying probability. Those movements are not purely random: rewrites shift the prompt-only prior, and the speeches that move the most are concentrated among low-margin cases. For practice, the clearest result is that simple output-level features already support targeted auditing. For mechanism, the result is more modest: internal drift exposes structured prompt-specific signatures for risky speeches, but it is not yet a better ranking signal than baseline output features.

In terms of limitations, this paper analyzes a relatively small exported run rather than the full intended corpus, due to resources and time constraints. The study focuses on one open-weight model family and one prompt family, so it does not establish cross-model robustness. Some auxiliary analyses, especially the source-level heterogeneity summary (Appendix Figure 9) should be treated as descriptive because several groups are tiny.

This project should be read as an auditing and reliability study, not as a claim that LLM outputs are ready to replace human political-text coding. The point of the analysis is precisely that prompt wording impact measurement, and that such instability should be quantified before these systems are used in substantive research pipelines.

The broader implication is that prompt variation should be treated as a measurable source of uncertainty when LLMs are used as coders. The next step is to scale the cleaned audit to the full intended sample and additional model families while keeping the same measurement-first perspective.

## Team contributions

This was a single-author project. Jiehan Liu completed the runner implementation, analysis scripts, figure refinement, interpretation, and write-up.

## References

- [1] 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. *arXiv preprint arXiv:2310.11324*, 2023.
- [2] Anwoy Chatterjee, HSVNS Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. Posix: A prompt sensitivity index for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14550–14565, 2024.
- [3] Kirk A. Hawkins, Rosario Aguilar, Bruno Castanho Silva, Erin K. Jenne, Bojana Kocijan, and Cristóbal Rovira Kaltwasser. Global Populism Database, 2019.
- [4] Amos Azaria and Tom Mitchell. The internal state of an llm knows when it’s lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, 2023.
- [5] Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. The curious case of hallucinatory (un) answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625, 2023.

## A Additional analyses

### A.1 Speech-level prior-correction map

Figure 6 adds a speech-level decomposition view. Each point is one speech, with total margin variance on the  $x$ -axis and residual variance after subtracting the prompt-only bias on the  $y$ -axis. Color marks the chosen speech-level sensitivity metric and point size shows mean rewrite drift.

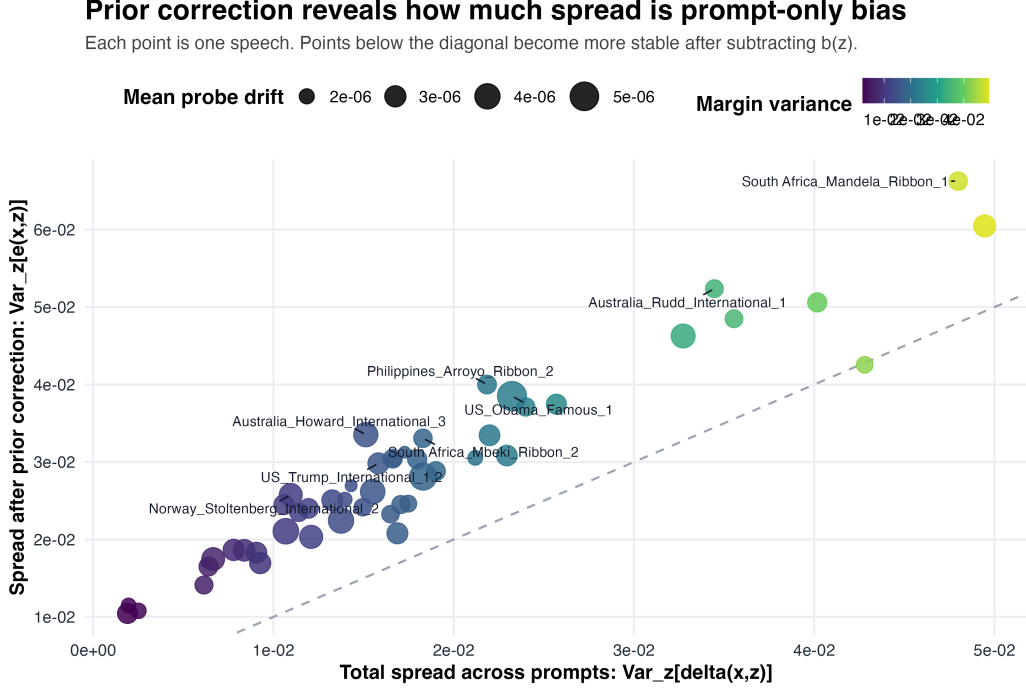


Figure 6: Speech-level decomposition map.

### A.2 Corner-case diagnostic under a variance-based sensitivity metric

Figure 7 revisits the knife-edge hypothesis using speech-level margin variance as the outcome. The relationship is nonlinear and weakly positive overall (Spearman  $\rho = 0.09$ ), which is much less decisive than the max-min probability-spread result in Figure 3. This is useful evidence that the choice of sensitivity metric matters: the baseline-margin story is strongest when the target is large score movement, not every variance-style summary.

### A.3 rewrite-specific drift signatures for the highest-risk speeches

Figure 8 shows that high-risk speeches often have distinct rewrite-specific drift profiles. The brightest cells are not uniformly spread across all rewrites; rather, sensitivity often concentrates on a few prompt comparisons.

### A.4 Exploratory group-level heterogeneity

Figure 9 summarizes prompt sensitivity across each group, where group is defined by country and speech types. Some group show near-zero median spread while others reach roughly 15 percentage points; hard flips are even more concentrated.

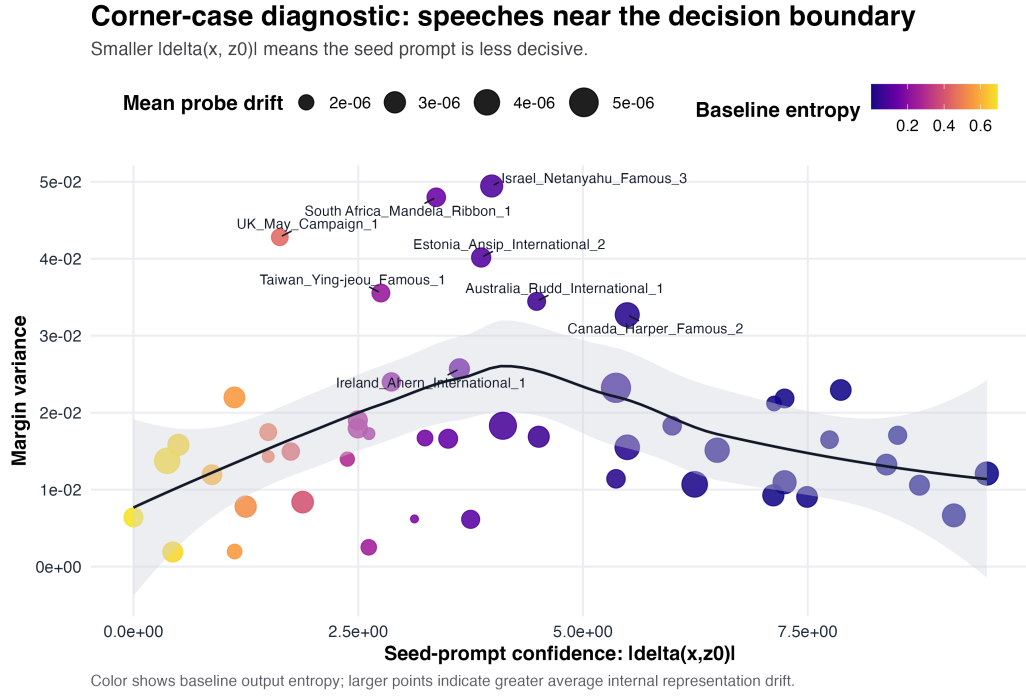


Figure 7: Corner-case diagnostic using speech-level margin variance rather than max–min probability spread. Colors show baseline entropy and point size shows mean rewrite drift.

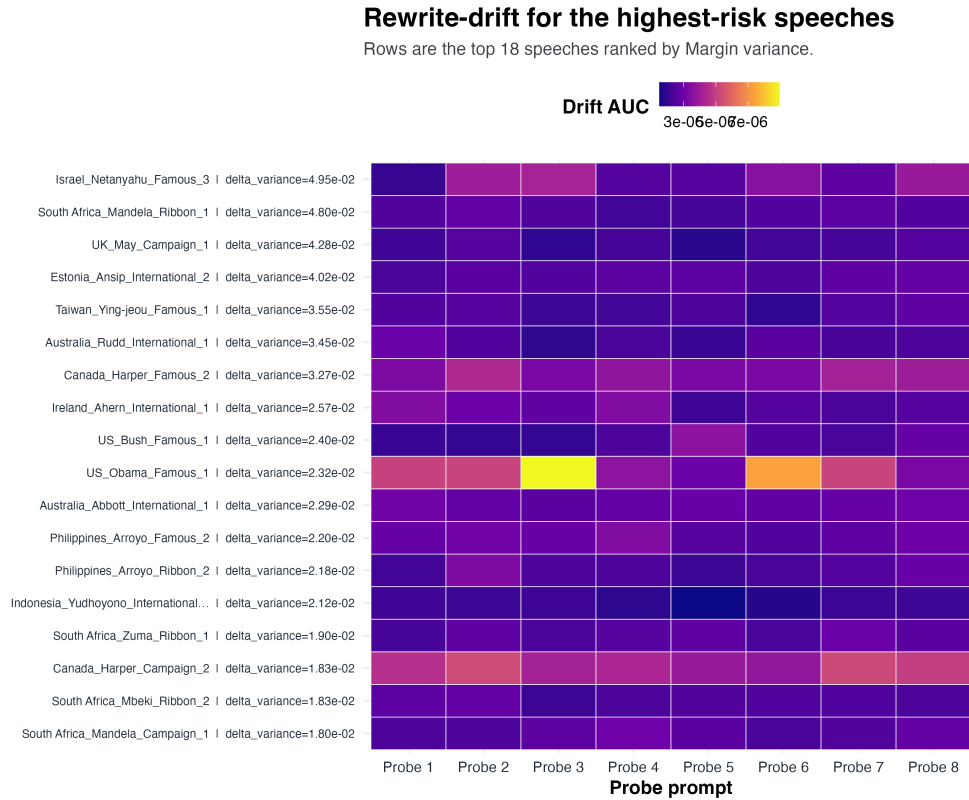


Figure 8: rewrite-drift signatures for the highest-risk speeches ranked by prompt sensitivity. Rows are speeches and columns are rewrite prompt comparisons.

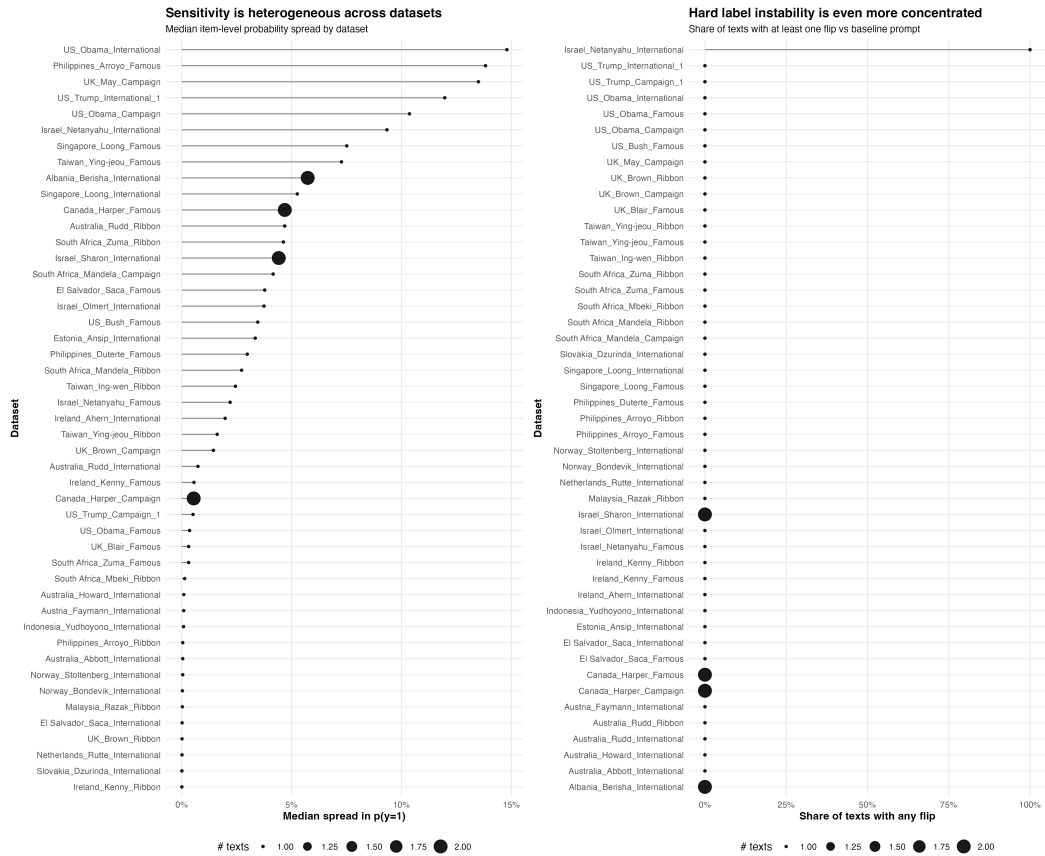


Figure 9: Exploratory text-level heterogeneity across speeches.