Explainable Attribution-Grounded Prompt Refinement for LLM Classifiers

Anonymous ACL submission

Abstract

Prompt engineering for LLM classifiers is typically informal or delegated to black-box optimization, leaving practitioners with little visibility into which input features actually drive predictions, where the model fails, or why a given prompt edit helps or hurts. We present a two-tier attribution workflow that makes LLM classification explainable at the prompt level: errors are decomposed into feature-level contributions, those contributions are translated into targeted prompt edits, and an evidence chain from error pattern to attribution to edit is preserved end-to-end. A cheap proxy SHAP analysis on a TF-IDF + Logistic Regression surrogate surfaces corpus-wide lexical failure patterns at millisecond cost; a direct model-level SHAP analysis on the target LLM (Qwen3.5-27B) using next-token logits validates those findings on a targeted slice. We instantiate the workflow on episodic versus semantic memory classification as a representative case study because its surface cues (past tense, first-person wording) systematically mislead LLMs. Three disambiguation rules derived from the proxy diagnosis yield significant gains on an expanded 200-row evaluation set (macro-F1: $0.736 \rightarrow 0.835$; McNemar p=0.027), outperforming OPRO-style (0.782) and ProTeGi-style (0.762) automatic prompt optimization baselines. The same rules transfer positively to Llama 3.1-8B ($+0.034$) and Ministral-3-14B ($+0.074$) on a 2,169-row held-out set, suggesting the proxy exposes a broadly shared distributional bias rather than a model-specific artifact. The central contribution is methodological: attribution-grounded refinement makes prompt engineering for LLM classifiers auditable by linking error patterns, feature attributions, and prompt edits in a reproducible chain.

1 Introduction

Prompt-based LLM classifiers are now standard for sentence-level NLP, yet the workflow for improving them remains largely opaque. Practitioners adjust wording until accuracy looks acceptable, or delegate the search to black-box prompt optimization (Pryzant et al., 2023; Yang et al., 2024), with little visibility into which input features actually drive a model's predictions, where the prompt systematically fails, or why a given edit helps one class while hurting another. This is at odds with the broader goal of explainable AI for classification (Arrieta et al., 2020; Zhao et al., 2024): when a deployed prompt regresses, the failure mode is rarely auditable, and the gap between black-box accuracy and white-box understanding leaves prompt engineering as a craft rather than a method.

SHAP-based attribution (Shapley, 1953; Lundberg and Lee, 2017) offers a systematic alternative: decompose errors into feature-level contributions, identify which cues drive failures, and translate those findings into targeted prompt edits. Applied at full fidelity to large LLMs this is prohibitively expensive, but a tiered design makes the workflow tractable by using a cheap surrogate for corpus-wide diagnosis and the target LLM itself for validation on a small slice. This paper develops that idea into a complete pipeline (§4): proxy SHAP on a linear surrogate surfaces broad lexical failure patterns at millisecond cost, those findings are translated into targeted prompt edits, and direct model-level SHAP validates the edits on the target LLM. Each edit traces to a specific attribution finding, so regressions are interpretable and the entire chain, from error pattern to attribution to edit, is auditable.

We instantiate the pipeline on episodic vs. semantic memory classification as a representative case study. The task asks a model to decide whether a sentence states durable knowledge (semantic; general facts, traits, habits) or narrates a specific lived event (episodic) (Tulving, 1972, 1985), and is increasingly used in conversational

memory systems (Park et al., 2023), personal assistants, and long-term dialogue agents. It is operationally subtle: temporally anchored facts ("My grandfather was born in 1930") remain semantic, while scene-rich language can be episodic without explicit recall markers, and models that lean on shallow cues such as past tense, time adverbs, and first-person pronouns systematically confuse these cases, consistent with the shortcut-learning literature (McCoy et al., 2019; Geirhos et al., 2020). The combination of a structurally simple binary label space and a strong, well-documented distributional shortcut makes the task a useful testbed: the failure mode is real and high-stakes, and the attribution signal is rich enough to drive concrete prompt edits.

Our contributions are:

1. A two-tier SHAP attribution pipeline (proxy + direct) for auditable prompt refinement of LLM classifiers, preserving an evidence chain from error patterns to edits (§4);
2. An instantiation on episodic vs. semantic memory classification, with significant gains over a minimal baseline (+0.099 mF1; p=0.027) and over OPRO and ProTeGi comparators (§5);
3. Cross-model transfer to three additional model families and robustness analyses (rule ablations, label-surface sensitivity, per-corpus breakdowns), indicating the discovered bias is shared across models rather than model-specific (§7);
4. A proxy–direct agreement analysis showing 80% sentence-level agreement on a diagnostic slice, supporting the proxy's practical utility as a cheap diagnostic surrogate for the target LLM (§6).

2 Related Work

Memory in NLP. The semantic–episodic distinction (Tulving, 1972, 1985) has gained NLP relevance through conversational memory systems (Park et al., 2023) and LLM episodic memory benchmarks (Huet et al., 2025). Reflective memory architectures for LLM agents (Shinn et al., 2023; Tan et al., 2025) and historical-character episodic memory systems (Gonzalez and DiPaola, 2025) further motivate auditable prompt design in memory-sensitive applications.

Token/prompt-level Shapley methods. Our work builds on SHAP (Lundberg and Lee, 2017; Shapley, 1953) and relates to LIME (Ribeiro et al., 2016) and the broader XAI literature (Arrieta et al., 2020; Zhao et al., 2024). TokenSHAP (Horovicz and Goldshmidt, 2024) provides a model-agnostic token-level Shapley estimator using output-text similarity as a value function; TokenShapley (Xiao et al., 2025) offers a similar framework with semantic-embedding value functions and distinct computational trade-offs. Mohammadi (2024) apply partition-based Shapley values to prompt fields to diagnose framing effects and "token noise." Our proxy tier uses a linear surrogate for speed (exact SHAP at near-zero cost), while our direct tier uses next-token logits restricted to the label channel for tighter classification-specific attribution. The methods are complementary: field-level Shapley diagnoses how prompt structure affects behavior under a fixed prompt, whereas our workflow diagnoses which input features cause errors to guide prompt revision.

Attribution for prompt control. FrugalPrompt (Raiyan et al., 2025) compresses prompts by retaining high-attribution tokens; CAGE (Walker and Ewetz, 2025) builds attribution graphs for generation faithfulness. These target efficient or faithful behavior under a fixed prompt, whereas we target prompt revision. Integrated Gradients (Sundararajan et al., 2017) and related path methods offer alternative attribution families for the direct-validation tier when gradients are available. SHAP-guided generation has been shown to outperform naïve prompting in adjacent settings (Martinez et al., 2025).

Prompt engineering and APO. Instruction tuning (Ouyang et al., 2022), chain-of-thought (Wei et al., 2022), and verbal reinforcement (Shinn et al., 2023) advance prompt-based methods, while shortcut-learning (McCoy et al., 2019; Geirhos et al., 2020) and prompt-sensitivity studies (Sclar et al., 2024; He et al., 2024) show when models rely on surface cues. OPRO (Yang et al., 2024) and ProTeGi (Pryzant et al., 2023) optimise prompts via scored search; DSPy/MIPRO (Khattab et al., 2023; Opsahl-Ong et al., 2024) and EvoPrompt (Guo et al., 2024) extend this with programmatic compilation and evolutionary search. Our method differs in kind: the evidence chain from attribution to edit is preserved, making

prompt changes auditable rather than opaque.

3 Task and Data

Task definition. We study binary sentence-level classification: Semantic (general knowledge, stable facts, traits, or relationships presented as known and not grounded in a specific episode) versus Episodic (recollection of a specific personally experienced event, typically with temporal, spatial, or emotional detail). The key distinction is functional: whether a sentence states knowledge or narrates a lived episode (Tulving, 1972). Ambiguous cases default to Semantic when episodic evidence is absent.

Corpora. The data pool draws from five open corpora selected for heterogeneity: ASQ (Fu et al., 2019) (advice-seeking narratives), Buckeye (Pitt et al., 2007) (conversational speech), Hippocorpus (Sap et al., 2020) (recollected/imagined narratives), Persona-Chat (Zhang et al., 2018) (persona-grounded dialogue), and TIFU (Kim et al., 2019) (first-person Reddit narratives). The classified corpus splits will be released for transparency and future research.

Splits. All rows were manually revised against a written operational guide. The benchmark uses six frozen splits, summarized below; full eligibility rules and leakage safeguards are in Appendix A.

- **Train** (n=142)**:** Used only for proxy training and as the APO optimization pool; never for evaluation. Built by a deterministic 60/20/20 corpus×label-stratified split of 238 curated balanced examples (seed 42), with rounding performed within each stratum.
- **Dev** (n=48)**:** The 20% dev partition from the same stratified split; used for prompt iteration and APO selection.
- **Frozen eval** (n=48)**:** The remaining 20% partition of the curated balanced pool, strictly disjoint from train/dev; serves as the historical benchmark and is embedded wholesale into the expanded eval.
- **Expanded eval** (n=200)**:** Headline evaluation set (Table 1). Built by embedding the 48-row frozen eval and sampling 152 additional rows from a separately adjudicated full corpus pool (n=2,311) via a deterministic corpus×label-stratified rule (target 20 per stratum, seed 42); disjoint from train/dev.

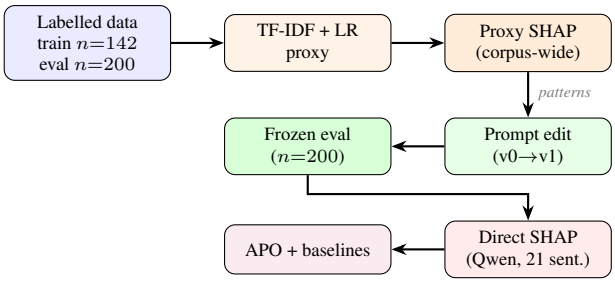

Figure 1: Attribution-grounded prompt refinement pipeline. Proxy SHAP surfaces feature-level failure modes (ms/sent); direct SHAP validates on the target LLM (∼41 s/sent, targeted slice).

- **Extended hold-out** (n=400)**:** Strict superset of the 200-row expanded eval plus 200 additional rows sampled under the same eligibility rules (seed 4242); used for prompting-strategy comparisons.
- **Leakage-free held-out** (n=2,169)**:** Full adjudicated corpus pool minus the 142 train rows; used for rule ablations, label-surface sensitivity, cross-model transfer, and the cross-corpus splits in §7.

Table 1: Expanded evaluation set (n=200): corpus×label strata. Persona-Chat Episodic was capped at 17 (pool exhausted); shortfall redistributed to TIFU Episodic.

Corpus	Semantic	Episodic
ASQ	20	20
Buckeye	20	20
Hippocorpus	20	20
Persona-Chat	20	17
TIFU	20	23
Total	**100**	**100**

4 Method

Figure 1 overviews the pipeline. The workflow is tiered: cheap proxy SHAP drives the iteration loop; standard inference validates at scale; direct SHAP is reserved for targeted high-fidelity checks.

4.1 Two Tiers of SHAP Attribution

Proxy SHAP (corpus-wide). We train a TF-IDF + Logistic Regression model on the 142-row training split and explain it with `shap.LinearExplainer` (Lundberg and Lee, 2017). For the linear model, Shapley values are exact: $\phi_i = w_i(x_i - \mathbb{E}[x_i])$, where w_i is the logistic

regression coefficient and x_i the TF-IDF feature value. This makes corpus-wide attribution essentially free, enabling iterative diagnosis. The trade-off is that the proxy sees only n-gram counts and cannot distinguish context (e.g., "I was" receives the same weight whether in "I was a teacher" or "I was walking to the store").

Direct SHAP (per-sentence, target LLM). For each sentence in a 21-sentence diagnostic slice, we apply `shap.PartitionExplainer` on Qwen/Qwen3.5-27B.[1] The prediction function is a two-way softmax over next-token logits for the label tokens `Semantic` and `Episodic`:

$$f_{\text{direct}}(s) = \frac{\exp(\ell_{\texttt{Sem}})}{\exp(\ell_{\texttt{Sem}}) + \exp(\ell_{\texttt{Epi}})}, \qquad (1)$$

where $\ell_{\texttt{Sem}}$ and $\ell_{\texttt{Epi}}$ are the next-token logits at the label position with the prompt and sentence inserted. Attribution units are whitespace-delimited tokens; the reference state is the prompt with sentence tokens masked. Direct SHAP costs $\sim$41 s per sentence, making it impractical for full-corpus use but valuable for targeted validation.

This two-tier design is related to but distinct from recent token-level Shapley methods for LLMs. TokenSHAP (Horovicz and Goldshmidt, 2024) and TokenShapley (Xiao et al., 2025) provide model-agnostic token-level Shapley estimators using different value functions (e.g., output-text similarity, semantic embeddings). Our direct-SHAP tier differs in using the model's own next-token logits restricted to the label channel, which provides a tighter attribution signal for classification tasks but requires logit access. Mohammadi (2024) apply partition-based Shapley values to prompt fields to diagnose "token noise" and framing effects; our proxy tier instead uses a linear surrogate for speed, with the direct tier validating on the target model.

Why both? The proxy is the diagnostic engine: fast enough for iterative use, it surfaces broad lexical patterns causing errors. Direct SHAP is the validation layer: it tests whether proxy findings reflect target-model behavior and reveals where the proxy breaks down (Appendix F, Table 11).

4.2 Diagnosing the Failure Mode

Under a minimal v0 prompt (task definition + output format only), Qwen produces 50 errors on the n=200 set, of which 48 are false-Episodic. The model systematically over-predicts Episodic.

Proxy SHAP on the error slice reveals two opposing feature groups (Table 2): Episodic-driving features (`i`, `i was`, `went`) share a pattern of first-person past-tense narrative markers; Semantic-driving features (`is`, `are`, `don't`, `always`) are present-tense copulas and habitual markers.

Table 2: Top proxy SHAP features by direction on the error slice (LR coefficients from the 142-row training split).

Episodic-driving		**Semantic-driving**	
Feature	Coef.	Feature	Coef.
`i`	+0.62	`is`	−0.51
`i was`	+0.45	`have`	−0.41
`went`	+0.35	`are`	−0.35
`so i`	+0.32	`don't`	−0.29
`i got`	+0.31	`always`	−0.22

The semantic gap. The critical finding is not the global feature weights but their realized influence on the error slice. We define the semantic gap for feature j:

$$\text{gap}_j = 1 - \frac{|\bar{\phi}_j^{\text{err}}|}{|w_j|}, \qquad (2)$$

where $\bar{\phi}_j^{\text{err}}$ is the mean SHAP value on the error slice and w_j the global coefficient. All key Semantic-driving features have gap > 0.92: they are present in the misclassified sentences but contributing less than 8% of their potential Semantic influence. The model's narrative-leaning prior, plausibly from pretraining on autobiographical text, overpowers these counter-cues under the minimal prompt. This interpretation is consistent with findings on distributional shortcuts (McCoy et al., 2019; Geirhos et al., 2020) and prompt sensitivity (Sclar et al., 2024; Zhao et al., 2021).

4.3 From Diagnosis to Prompt Intervention

Each proxy SHAP finding maps to a specific prompt edit. The v1 prompt adds three disambiguation rules to v0:

1. **Rule 1** (targets: `i`, `i was`): stable first-person states/traits are Semantic ("I am a teacher", "I don't trust people easily").
2. **Rule 2** (targets: `went`, `so i`): specific-moment anchoring signals Episodic ("I went to the store yesterday").

[1] 4-bit quantized; RTX 4090-class GPU. All local runs use this deployment.

3. **Rule 3** (targets: `is`, `don't`, `always`): present-tense habitual verbs prefer Semantic unless a specific event is explicit.

These rules are not ad hoc: each directly counters an attribution-identified failure mode. The v1 prompt also adds "habitual patterns" to the Semantic definition, naming a category v0 left implicit. Full prompt text is in Appendix B, and a formal-procedure statement of the workflow (Algorithm 1) is given in Appendix G.

4.4 Comparators: APO Baselines

We implement OPRO-style (Yang et al., 2024) and ProTeGi-style (Pryzant et al., 2023) comparators. Both use Qwen as the local optimizer (temperature 0.7) on a 110-row optimization pool (55 per class from the training split) with patience-2 early stopping (OPRO: 5 rounds; ProTeGi: 5 steps, beam width 2). We additionally rescore saved predictions with an independent judge (gpt-4.1-mini) to mitigate same-model evaluation circularity. These should be read as controlled comparisons under fixed budgets, not as upper bounds on the broader APO design space, which includes stronger variants such as DSPy/MIPRO (Khattab et al., 2023; Opsahl-Ong et al., 2024) and EvoPrompt (Guo et al., 2024). Full configuration details (optimizer sampling, search space, selection metric, seeds) are in Appendix D.

5 Results

5.1 Main Benchmark

Table 3 reports results on the expanded n=200 evaluation set.

Table 3: Results on the expanded evaluation set (n=200). Bootstrap 95% CIs on macro-F1; McNemar p compares Qwen v0 vs. v1.

Model	Method	Acc.	mF1 [95% CI]
Qwen	v0 (baseline)	.750	.736 [.670, .796]
Qwen	**SHAP v1**	**.835**	**.835 [.780, .885]**
Qwen	OPRO	.785	.782
Qwen	ProTeGi	.770	.762
o1	v0	.785	.782 [.724, .838]
o1	SHAP v1	.795	.795 [.736, .850]
TF-IDF+LR	(anchor)	.685	.677 [.609, .742]

The SHAP-guided v1 prompt significantly improves Qwen (McNemar p=0.027; ΔmF1 = +0.099; 95% CI [+0.030, +0.172]). Both APO methods improve over v0 but do not surpass v1. Independent gpt-4.1-mini rescoring preserves the ordering (OPRO: 0.740; ProTeGi: 0.684).

Confusion matrix shift. The improvement is mechanistically interpretable: v1 sharply reduces Episodic over-prediction (48 $\rightarrow$ 13 false-Episodic), with a smaller increase in the opposite error type (2 $\rightarrow$ 20 false-Semantic), resulting a more balanced classifier and higher macro-F1. Further iterations would likely improve this balance still. Table 4 quantifies the class-balance effect.

Table 4: Class-bias metrics on n=200. $\Delta R = R_{\text{Epi}} - R_{\text{Sem}}$; $\beta = R_{\text{Epi}}/R_{\text{Sem}}$.

Method	ΔR	β	Bias
v0	+0.46	1.89	Strong Epi
SHAP v1	**−0.07**	**0.92**	**Balanced**
OPRO	−0.25	0.73	Mod. Sem
ProTeGi	−0.36	0.62	Strong Sem

The SHAP-guided method achieves the most balanced class recall (β=0.92, near unity). Both APO methods over-correct into Semantic bias (β=0.73 and 0.62), consistent with lacking a diagnostic signal about which class each edit harms. This asymmetry illustrates a structural advantage of attribution-guided refinement: because the semantic-gap analysis identifies which distributional cues cause errors, the resulting edits target the specific imbalance rather than blindly maximizing aggregate accuracy.

Per-class precision/recall. On the n=200 set, v0 achieves high Episodic recall (0.980) but poor Semantic recall (0.520). The SHAP-guided v1 rebalances to Semantic P/R/F1 = 0.813/0.870/0.841 and Episodic 0.860/0.800/0.829. OPRO and ProTeGi over-correct: OPRO pushes Semantic recall to 0.910 but drops Episodic recall to 0.660; ProTeGi reaches 0.950/0.590.

5.2 Error Analysis

Table 5 illustrates representative cases from the 48-row frozen set, grouped by v1 outcome.

The fixes share a clear pattern: v0 over-fired Episodic on sentences using first-person wording to express stable states or habitual reflections, exactly the mechanism surfaced by proxy SHAP. The regressions reveal the complementary risk: v1's stable-state rules can overcorrect on genuinely Episodic sentences that happen to contain

Table 5: Error analysis: representative cases from the evaluation sets. Fixes and regressions are drawn from the frozen n=48 set; the persistent error is from the expanded n=200 set. Gold labels from final refresh.

Sentence (abridged)	Gold	v0/v1	Type
Sometimes I think about the bad stuff I've said or done.	Sem	E→S	Fix
I love working with computers.	Sem	E→S	Fix
When I do decide to go see my mom . . . I feel so left out.	Epi	E→S	Regr.
. . . the dean would tell students to rise and President Bacow would declare them graduated.	Epi	E→S	Regr.
I know it cost $3,000 when my ex bought it 4 years ago . . .	Sem	E→E	Pers.

Semantic-leaning surface cues ("don't trust", "feel so left out"). The persistent errors expose boundary cases at the limits of prompt-level intervention: "i was a cop" reads as a biographical fact to both prompts despite its Episodic gold label.

5.3 Prompting-Strategy Baselines

On the n=400 extended hold-out (Table 6), the ranking is consistent: methods incorporating SHAP-derived rules match or approach v1 (few-shot + SHAP: 0.810; self-consistency: 0.827), while methods without them trail by 2–5 points (zero-shot v0: 0.775; CoT: 0.793; DistilBERT: 0.799). Self-consistency (k=5, temperature 0.7) produces identical predictions to greedy v1, indicating that Qwen's errors are stable misclassifications rather than sampling noise.

Table 6: Prompting baselines on n=400 hold-out. McNemar p vs. SHAP v1. DistilBERT fine-tuned on 142 rows.

Method	mF1 [95% CI]	p vs v1
Zero-shot v0	.775 [.732, .815]	.076
SHAP v1 (ref.)	**.827 [.790, .862]**	N/A
Few-shot (k=5)	.805 [.765, .842]	.268
Few-shot + SHAP	.810 [.770, .847]	.392
CoT	.793 [.753, .832]	.118
CoT + SHAP	.810 [.770, .847]	.248
Self-consist. (k=5)	.827 [.790, .862]	1.00
DistilBERT (66M)	.799 [.758, .837]	.215

6 Proxy–Direct SHAP Agreement

On a 21-sentence diagnostic slice, direct SHAP shows 80% sentence-level and 74% token-level agreement with the proxy (Cohen's κ=0.46 from 23 token comparisons; cluster-bootstrap 95% CI $[0.56, 0.90]$ for tokens, $[0.60, 1.00]$ for sentences). Wide CIs reflect the small sample; these are descriptive rather than population-level estimates.

The qualitative pattern is informative. For the sentence "I love working with computers" (gold: Semantic), the proxy assigns a small Episodic contribution to "I" (dataset prior on first-person pronouns), while direct SHAP assigns Semantic-leaning attribution to every token; Qwen reads the full sentence coherently as a preference statement. This reveals where the proxy breaks down: context-sensitive cases where the LLM overrides its distributional prior are precisely the cases where direct validation changes the interpretation.

Figure 2 compares proxy and direct attributions on the same sentence. In the proxy view, `I` still carries a small Episodic contribution (dataset prior on first-person pronouns), while `love`, `working`, `computers` lean Semantic. In the direct view, Qwen assigns Semantic-leaning attribution to every token including `I`, reading the full string as a preference statement. This disagreement localises where the proxy breaks down: a proxy-only reading might suggest first-person wording is Episodic evidence in this row, while direct SHAP shows the model coherently reads the sentence as Semantic. The practical impact is that context-sensitive cases (first-person preference or trait statements) are precisely where direct validation is most informative.

LOO perturbation on o1. Since o1 lacks logit access, we approximate token sensitivity via leave-one-out perturbation (LOO) on the same 21-sentence slice. Each of 308 masked-token variants is submitted to o1; only 16 (5.2%) cause a label flip, indicating high stability under single-token perturbation. Of the 16 flips, 75% agree in direction with Qwen direct SHAP, suggesting shared sensitivity to the most discriminative tokens across model families.

7 Robustness

Per-corpus breakdown. Table 7 shows v1 gains are positive on 4/5 corpora; Persona-Chat is flat (0.865 → 0.863), consistent with near-ceiling v0 performance. The largest gains are on corpora richest in autobiographical first-person content (ASQ: +0.136; TIFU: +0.155), supporting the distributional-prior correction interpretation.

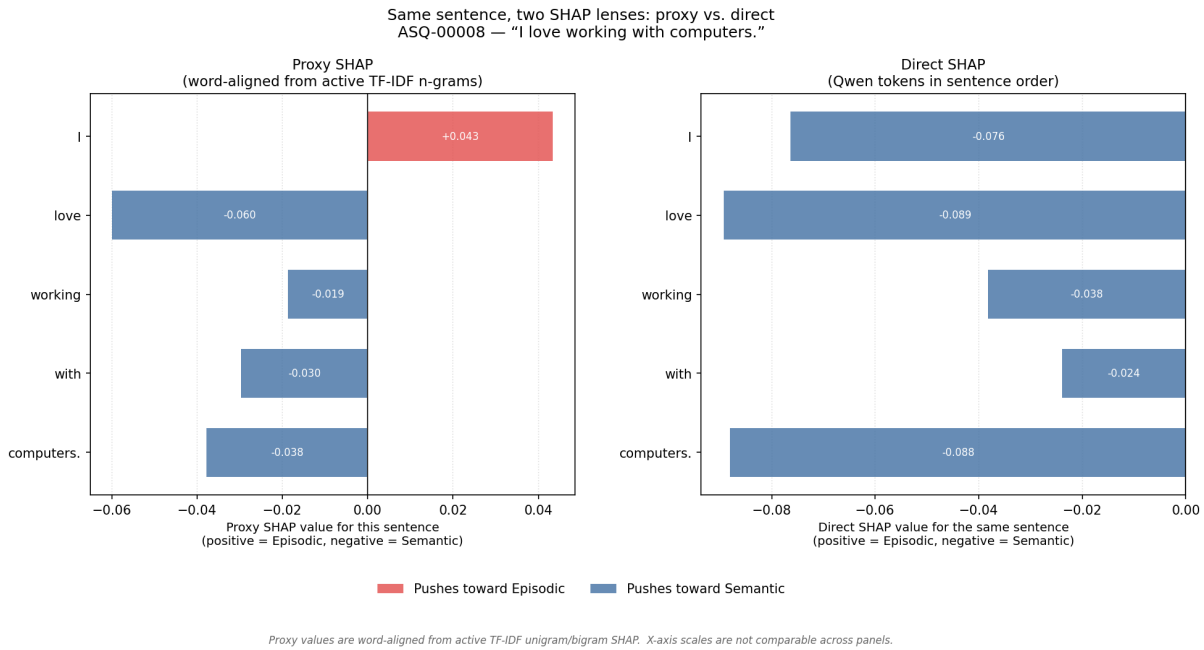

Figure 2: Same-sentence attribution comparison for "I love working with computers." Left: proxy (TF-IDF+LR SHAP values distributed over words). Right: direct (Qwen token-level SHAP). Red = Episodic, blue = Semantic. Scales are not comparable across panels.

Table 7: Per-corpus macro-F1 on n=200 for Qwen v0 vs. v1. Bootstrap 95% CIs (10,000 resamples).

Corpus	v0 mF1	v1 mF1	Δ
ASQ	.581 [.41, .73]	.716 [.56, .85]	+.136
Buckeye	.733 [.58, .87]	.899 [.79, .98]	+.166
Hippo.	.733 [.58, .87]	.774 [.64, .90]	+.040
P-Chat	.865 [.74, .97]	.863 [.74, .97]	−.001
TIFU	.751 [.60, .88]	.907 [.81, .98]	+.155

Rule ablations. On the n=2,169 leakage-free held-out set (proxy-train rows excluded), removing any single rule causes $\leq$0.003 mF1 loss, confirming a distributed effect: the three rules address a shared bias cluster, not independent failure modes.

Label-surface sensitivity. Replacing `Semantic/Episodic` with `Factual/Experiential` or `General/Personal` changes mF1 by at most 0.9 pp, arguing against label-token prior confounds.

Surrogate stability. Bootstrap resampling (1,000 draws of the 142-row training split) shows that all key Semantic-driving features (`is`, `are`, `have`, `don't`, `always`) have 95% CIs excluding zero; only `was` (mixed episodic/semantic role) crosses zero.

Cross-model transfer. On the same n=2,169 held-out, frozen v0/v1 prompts yield consistent gains across three families beyond Qwen: Llama 3.1-8B (+0.034 mF1), Ministral-3-14B (+0.074), and o1 (+0.013 on n=200). The v1 rules were derived solely from Qwen errors via the attribution loop; none of these three transfer families contributed any signal during prompt design, so their positive transfer supports the view that proxy SHAP exposed a broadly shared failure mode rather than a Qwen-specific artifact.

Full-corpus robustness. On the full 2,311-sentence corpus pool, v1 generalizes for Qwen (mF1: 0.764 → 0.828, $\Delta = +0.064$), while o1 remains within noise of its strong baseline (0.839 → 0.830, $\Delta = -0.009$). This asymmetry reflects the headroom pattern: o1's high baseline leaves less room for prompt-level improvement.

Cross-corpus transfer. Holding out entire corpora (Split A: train ASQ/Buckeye/Hippocorpus, test Persona-Chat/TIFU; Split B: reversed), v1 transfers positively for Qwen (Split A: +0.079; B: +0.042) but not for o1 (−0.010; −0.019), whose strong baseline leaves little headroom for prompt-level gains. Post-hoc analysis of 102 o1 regressions shows 60.8% involve first-person habitual constructions, the same patterns v1 targets, suggesting overcorrection when the model already in-

ternalizes these patterns.

8 Discussion

The distributional mechanism. The core finding is mechanistic: LLMs pretrained on vast autobiographical corpora acquire a strong prior that first-person past-tense language signals event recollection. Under the minimal v0 prompt, this prior causes systematic over-prediction of Episodic. Proxy SHAP makes the mechanism visible by decomposing errors into feature-level contributions: Semantic-driving features (`is`, `don't`, `always`) are present in the misclassified sentences but their influence is overridden by salient Episodic cues (`i was`, `went`). The v1 rules do not retrain the model; they shift its effective decision boundary by explicitly telling it which surface cues should not be treated as sufficient evidence for Episodic. The confusion-matrix shift confirms this: false-Episodic errors drop from 48 to 13, directly reflecting the diagnosed mechanism.

SHAP vs. APO. Both APO comparators improve over v0, confirming that prompt optimization extracts gains on this task. The gap between APO and SHAP-guided refinement reflects complementary information sources: APO searches the prompt space through scored proposals but has no mechanism to diagnose why a prompt fails or which class is being harmed by each edit. This helps explain the over-correction pattern in Table 4: without a diagnostic signal, OPRO and ProTeGi push accuracy upward by flipping Semantic recall at the expense of Episodic, producing a new imbalance rather than correcting the original one. The SHAP method targets the specific distributional asymmetry, achieving a more balanced outcome. Its additional advantage is auditability: each edit traces to a specific attribution finding, making it possible to predict where the prompt might fail and to debug regressions when they appear.

When transfer fails. Cross-corpus results show that v1 transfers positively for Qwen but not for o1, whose strong v0 baseline leaves little headroom. The o1 regression analysis reveals a concrete mechanism: 60.8% of o1's regressions involve first-person habitual constructions, the same patterns v1 targets. o1 already maintains a stronger prior toward Semantic for these patterns; the v1 rules amplify this to the point of over-correction. This suggests that model-conditional prompt tuning may be needed: rules that improve a weaker model's boundary may require softening for a stronger one.

Tiered cost structure. The workflow is naturally tiered: proxy SHAP at ms cost drives iteration; standard inference at $\sim$0.5 s/sentence validates at scale; direct SHAP at $\sim$41 s/sentence is reserved for targeted validation. This cost structure is a feature, not a limitation: the proxy need not be perfectly faithful to be useful, as long as direct SHAP can check the cases where faithfulness matters most.

9 Conclusion

We presented an attribution-grounded prompt refinement workflow that makes LLM classification explainable at the prompt level, instantiated on episodic vs. semantic memory classification as a case study. Starting from a minimal prompt, proxy SHAP on a linear surrogate diagnoses the distributional mechanism behind classification failures; three targeted disambiguation rules derived from the diagnosis significantly improve Qwen/Qwen3.5-27B (ΔmF1 $=$ $+0.099$; p=0.027), outperforming OPRO and ProTeGi comparators while maintaining class balance. Direct model-level SHAP validates the proxy signal on the target LLM, and the same rules transfer to Llama 3.1 and Ministral-3 without re-running attribution. More broadly, the paper argues for treating prompt engineering for LLM classifiers as an auditable process: the evidence chain from error pattern to attribution to edit to regression analysis makes prompt changes inspectable, reproducible, and debuggable. Semantic–episodic classification serves as the testbed; the deeper result is methodological, and the same pipeline applies to any classification task where a model's default feature weighting conflicts with the task definition.

Limitations

Our evaluation is limited to English first-person narrative across five corpora; generalization to other languages or genres has not been established. The annotation policy defaults ambiguous cases to Semantic, which may under-detect episodic content in indirect or culturally variable narratives. Direct SHAP was run on Qwen only (o1 uses LOO approximation capturing single-token marginal effects only). The direct SHAP

implementation uses whitespace tokenization (not model subwords), first-label-token logits (an approximation for multi-token labels), and a fixed reference state (prompt with masked sentence); alternative configurations could change fine-grained attribution values. A sensitivity analysis over mask token choice, subword alignment, and reference baselines remains future work, as does substituting other attribution families (e.g., Integrated Gradients, perturbation-based methods) at the direct-validation tier. Proxy–direct agreement was measured on a small diagnostic slice (n=21) with wide CIs; larger-sample and multi-model validation is needed. APO comparators ran under limited budgets (3 rounds/steps before early stopping); stronger search could narrow the gap. The proxy depends on a 142-example training set; while bootstrap analysis shows stable feature directions, the small corpus limits generality of learned lexical patterns. The DistilBERT comparison is not scale-fair (66M vs. 27B parameters). The benchmark pool co-evolved with iterative modeling; circularity safeguards are described in §3 and Appendix A, but residual risk remains.

Ethics Statement

We separate benchmark prompts from adjudication prompts and document dataset drift. Because gold labels reflect operational policy, we recommend human oversight in any deployment affecting user memories. Multi-annotator adjudication on fresh samples would further harden the benchmark.

Acknowledgments

[Redacted for review.]

References

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, and 1 others. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

Liye Fu, Jonathan P Chang, and Cristian Danescu-Niculescu-Mizil. 2019. Asking the right question: Inferring advice-seeking intentions from personal narratives. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 528–541.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Rafael Arias Gonzalez and Steve DiPaola. 2025. Cognitively-inspired episodic memory architectures for accurate and efficient character ai. *arXiv preprint arXiv:2511.10652*.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2024. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *International Conference on Learning Representations*, volume 2024, pages 34133–34156.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does prompt formatting have any impact on llm performance? *arXiv preprint arXiv:2411.10541*.

Miriam Horovicz and Roni Goldshmidt. 2024. Tokenshap: Interpreting large language models with monte carlo shapley value estimation. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 1–8.

Alexis Huet, Zied Houidi, and Dario Rossi. 2025. Episodic memories generation and evaluation benchmark for large language models. In *International Conference on Learning Representations*, volume 2025, pages 51409–51457.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, and 1 others. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of reddit posts with multi-level memory networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2519–2531.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Paula Joy B Martinez, Jose Marie Antonio Miñoza, and Sebastian C Ibañez. 2025. An interpretability-guided framework for responsible synthetic data generation in emotional text. *arXiv preprint arXiv:2511.16132*.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3428–3448.

Behnam Mohammadi. 2024. Explaining large language models decisions using shapley values. *arXiv preprint arXiv:2404.01332*.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing instructions and demonstrations for multi-stage language model programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9340–9366.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Mark A Pitt, Laura Dilley, Keith Johnson, Scott Kiesling, William Raymond, Elizabeth Hume, and Eric Fosler-Lussier. 2007. Buckeye corpus of conversational speech (2nd release). *Columbus, OH: Department of Psychology, Ohio State University*, pages 265–270.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 7957–7968.

Syed Rifat Raiyan, Md Farhan Ishmam, Abdullah Al Imran, and Mohammad Ali Moni. 2025. Frugalprompt: Reducing contextual overhead in large language models via token attribution. *arXiv preprint arXiv:2510.16439*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Maarten Sap, Eric Horvitz, Yejin Choi, Noah A Smith, and James Pennebaker. 2020. Recollection versus imagination: Exploring human memory and cognition via neural language models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1970–1978.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. 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, pages 25055–25083.

Lloyd S. Shapley. 1953. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker, editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, and 1 others. 2025. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8416–8439.

Endel Tulving. 1972. Episodic and semantic memory. In Endel Tulving and Wayne Donaldson, editors, *Organization of Memory*, pages 381–403. Academic Press, New York.

Endel Tulving. 1985. *Elements of Episodic Memory*, revised ed. edition. Oxford Psychology Series. Oxford University Press.

Chase Walker and Rickard Ewetz. 2025. Explaining the reasoning of large language models using attribution graphs. *arXiv preprint arXiv:2512.15663*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yingtai Xiao, Yuqing Zhu, Sirat Samyoun, Wanrong Zhang, Jiachen T Wang, and Jian Du. 2025. Tokenshapley: Token level context attribution with shapley value. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3882–3894.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. In *International Conference on Learning Representations*, volume 2024, pages 12028–12068.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

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

A Data Curation Details

Benchmark lineage. We used GPT-4 to generate preliminary sentence labels as a pre-annotation step to accelerate curation. The pre-labels were then independently reviewed and revised by two human annotators against a written operational guide grounded in Tulving (1972, 1985); disagreements or uncertain cases were discussed and resolved per the guide, and the corpus was consolidated through repeated cycles of prompt design, model comparison, and SHAP analysis.

Eligibility flags. All evaluation sets (n=200, n=400, n=2,169) are drawn from a single adjudicated pool filtered by two flags: Phase 6 `corpus_eligible` (length and content suitability) and Phase 0B `triage_ok` (no fragmentary or out-of-policy rows). The n=400 hold-out adds 200 rows under the same flags with seed 4242.

Circularity safeguards. After all manifests were frozen, no evaluation sentence was relabeled in response to v1 behavior. The leakage-free n=2,169 slice (proxy-train rows excluded from the full pool) provides complementary evidence with provably no overlap with the proxy training data.

B Prompt Text

Prompt v0 (minimal).

```
You are classifying a sentence as either
Semantic or Episodic.

Semantic = general knowledge, stable facts,
preferences, traits, or relationships not tied
to one specific lived event.
Episodic = a specific personally experienced
event, interaction, or scene.

Reply with exactly one label:
Semantic
or
Episodic

Sentence: {{SENTENCE}}
```

Prompt v1 (SHAP-guided). Three disambiguation rules (bold) added:

```
You are classifying a sentence as either
Semantic or Episodic.

Semantic = general knowledge, stable facts,
preferences, traits, habitual patterns, or
relationships not tied to one specific lived
event.
Episodic = a specific personally experienced
event, interaction, or scene.

Key disambiguation:
- Sentences about stable states, traits, or
ongoing situations are Semantic even if
written in first person.
- Sentences anchored to a specific moment,
occasion, or sequence of events are
Episodic.
- If a sentence uses present-tense verbs
to describe a recurring pattern or stable
fact, prefer Semantic unless a specific
event is explicit.

Reply with exactly one label:
Semantic
or
Episodic

Sentence: {{SENTENCE}}
```

C Full Robustness Table

Table 8: Robustness on the n=2,169 held-out corpus. Panels: A baselines, B ablations (Δ from full v1), C label sensitivity, D cross-family.

	Condition	mF1	95% CI	Δ
A	v0	.766	[.748, .783]	N/A
	v1	.828	[.812, .843]	+.062
B	v1−R1	.825	[.809, .841]	−.003
	v1−R2	.828	[.812, .844]	+.001
	v1−R3	.825	[.809, .841]	−.002
C	Factual/Exp.	.831	[.815, .847]	+.003
	General/Pers.	.822	[.806, .838]	−.006
D	Llama v0	.628	[.606, .649]	N/A
	Llama v1	.662	[.640, .682]	+.034
	Ministral v0	.671	[.651, .691]	N/A
	Ministral v1	.745	[.726, .763]	+.074

D APO Comparator Details

Both APO protocols use Qwen/Qwen3.5-27B (4-bit local) as optimizer and scorer.

Table 9: APO configuration.

Setting	Value
Optimizer sampling	temp. 0.7, top-p 0.9
Scoring	greedy, max 20 tokens
Optimization pool	110 rows (55/55)
Dev slice	20 rows (10/10)
Selection metric	dev_acc $-0.25\times$ parse_fail
OPRO	5 rounds, patience 2
ProTeGi	5 steps, beam 2, patience 2
Seed	42

OPRO early-stopped after round 3 (best: round 1, dev_acc $= 0.90$). ProTeGi early-stopped after step 3 (best: step 1, dev_acc $= 0.85$). Both APO-generated prompts independently discovered some distinctions targeted by v1 (habit vs. event, summary vs. scene), but arrived at them through scored search rather than attribution evidence.

E Full-Corpus Pool Composition

Table 10: Full corpus pool (n=2,311) by corpus and label.

Corpus	Sem.	Epi.	Total
ASQ	303	194	497
Buckeye	350	24	374
Hippocorpus	101	394	495
Persona-Chat	445	14	459
TIFU	67	419	486
Total	**1,266**	**1,045**	**2,311**

F SHAP Tier Comparison

Table 11: Proxy vs. direct SHAP comparison.

	Proxy	**Direct**
Model	TF-IDF + LR	Qwen 3.5-27B
Unit	n-gram (global)	token (per-sent.)
Context	No	Yes
Cost	∼ms	∼41 s
Scale	Full corpus	Targeted slice

G Formal Procedure

Algorithm 1 formalizes the pipeline described in §4.

Algorithm 1 Attribution-grounded prompt refinement

Require: Labelled train $\mathcal{D}_{\text{tr}}$, eval $\mathcal{D}_{\text{ev}}$, LLM $\mathcal{M}$, seed prompt p_0
Ensure: Refined prompt p^*, evidence trail $\mathcal{E}$
Stage 1: Baseline & error characterisation
$\hat{\mathbf{y}}_0 \leftarrow \mathcal{M}(p_0, \mathcal{D}_{\text{ev}})$; $\mathcal{C}_0 \leftarrow \textsc{ConfMat}(\hat{\mathbf{y}}_0, \mathbf{y}_{\text{gold}})$
$\mathcal{S}_{\text{err}} \leftarrow \{(x_i, y_i) : \hat{y}_{0,i} \neq y_i\}$
Stage 2: Proxy SHAP diagnosis
Train f_{proxy} on $\mathcal{D}_{\text{tr}}$ (TF-IDF + LR)
$\boldsymbol{\phi} \leftarrow \textsc{LinearShap}(f_{\text{proxy}}, \mathcal{S}_{\text{err}})$
Identify semantic-gap features $\mathcal{G}$: active in errors but $|\bar{\phi}| \ll |w|$
Stage 3: Prompt intervention
for each feature group g in $\{\mathcal{F}_{+}^{\text{err}}, \mathcal{G}\}$ **do**
  Draft disambiguation rule r_g targeting pattern of g
end for
$p_1 \leftarrow p_0 \cup \{r_g\}$
Stage 4: Evaluation
$\hat{\mathbf{y}}_1 \leftarrow \mathcal{M}(p_1, \mathcal{D}_{\text{ev}})$; McNemar test, bootstrap CI
Stage 5: Direct SHAP validation (targeted)
Select diagnostic slice $\mathcal{S}_{\text{diag}} \subset \mathcal{D}_{\text{ev}}$
for $x_i \in \mathcal{S}_{\text{diag}}$ **do** $\boldsymbol{\phi}_i^{\text{direct}} \leftarrow \textsc{PartShap}(\mathcal{M}, p^*, x_i)$
end for
Compare $\boldsymbol{\phi}^{\text{proxy}}$ vs. $\boldsymbol{\phi}^{\text{direct}}$