# When Explanations Betray Backdoors: Black-Box Auditing for Language Model Classifiers

Anonymous ACL submission

## Abstract

Explanation-bearing language model classifiers are used for moderation, routing, topic triage, and low-resource annotation. We study trigger-unseen, black-box, per-input auditing when the defender has only clean calibration data but can ask the classifier for a label plus a short rationale or quoted evidence. We introduce Groundedness Drift, a lightweight score measuring whether the answer summary remains grounded in the input. Under common non-adaptive OpenBackdoor-style attacks, Groundedness Drift strongly reduces residual target ASR at a fixed clean-FPR budget and outperforms ONION, BBCaL, CoS-style, and prompted-reasoning baselines in most families. We then evaluate Unsupported Groundedness, a multi-probe escalation for explanation-camouflage stress cases. Unsupported Groundedness improves signals but does not close the adaptive gap. The paper is a careful evaluation of when explanation grounding is useful and where support or causal relevance remains unresolved.

## 1 Introduction

Backdoor attacks on text classifiers are a standard NLP security threat: clean inputs behave normally, while trigger-bearing inputs map to an attacker target (Qi et al., 2021b; Cui et al., 2022; Qi et al., 2021a). Instruction-tuned LMs (Language Models) extend this setting because they are often used as classifier-like labelers or routers while returning rationales or evidence for review (Xu et al., 2024; Ge et al., 2025). This is not meant to replace small supervised classifiers for stable high-volume tasks. It is relevant for cold-start, rule-driven, explanation-bearing, or LLM-assisted workflows such as moderation, topic triage, retrieval routing, stance annotation, and data labeling (Brown et al., 2020; Gilardi et al., 2023; Wei et al., 2022; Kojima et al., 2022).

The interface changes the defense problem. A hard-label classifier exposes a label. An LM classifier can expose a short rationale, policy-grounding sentence, or exact evidence quote. We ask whether this explanation channel can be used for trigger-unseen per-input auditing when the defender has only a small clean calibration set and black-box query access.

Our contributions are: (i) a clean-calibrated, trigger-unseen audit setting for explanation-bearing LM classifiers; (ii) Groundedness Drift, a one-query grounding score that detects when a target label no longer matches input-grounded rationale evidence; (iii) a systematic comparison against ONION, BBCaL, CoS-style, and prompted-reasoning baselines at matched clean-FPR (False Positive Rate) budgets; and (iv) Unsupported Groundedness, a multi-probe stress diagnostic showing that support/document consistency helps under explanation camouflage but does not solve adaptive attacks.

## 2 Related Work

Textual backdoors include lexical, syntactic, and style triggers, with OpenBackdoor providing a unified benchmark (Qi et al., 2021b; Cui et al., 2022). ONION detects anomalous tokens by perplexity change (Qi et al., 2021a); BBCaL uses black-box hard-label counterfactual consistency (Hu et al., 2024); CoS audits reasoning-output consistency (Li et al., 2025); BAIT, STAR, and Trigger in the Haystack are adjacent target-inversion, reasoning-time, or model-level scanners (Shen et al., 2025; Park et al., 2026; Bullwinkel et al., 2026). Our setting differs by combining per-input triage, clean calibration, trigger-unseen attacks, and explanation/evidence outputs from LM classifiers.

We also compare to prompted-reasoning detector variants built from Reasoning, ZS-CoT, and Auto-CoT prompts (Wei et al., 2022; Kojima et al.,

2022; Zhang et al., 2023): the prompt elicits a rationale, then the same consistency judge used for CoS assigns a rejection score. This tests whether generic reasoning elicitation is enough, versus explicit input-grounding. Furthermore, recent work distinguishes plausible explanations from faithful evidence (Lei et al., 2016; Jacovi and Goldberg, 2020; DeYoung et al., 2020); recent backdoor work similarly studies explanation behavior (Ge et al., 2025). These motivate us to study the rationale behavior. Finally, our fixed-FPR risk diagnostics follow selective prediction and detection-style operating-point evaluation (Geifman and El-Yaniv, 2017; Hendrycks and Gimpel, 2017; Liang et al., 2018).

## 3 Threat Model and Task

**Notation.** We write $x$ for an input text, $\mathcal{Y}$ for the finite label set, $y_t$ for the attack target label, and $f_\theta$ for the black-box LM classifier. When $f_\theta$ appears with a raw text argument in a metric or probe, it denotes the label predicted by the model under the corresponding classification prompt. The defender's clean calibration set is $\mathcal{C} = \{(x_i, y_i)\}_{i=1}^n$, while $\tilde{y}_i$ denotes the label predicted by $f_\theta$ on calibration input $x_i$ under the relevant audit prompt. The notation $\mathbb{I}[\cdot]$ is the indicator function: it equals 1 when the condition inside brackets is true and 0 otherwise. For a finite set of calibration scores $S$, $Q_{1-\alpha}(S)$ denotes the empirical $(1-\alpha)$ quantile used to target clean-FPR budget $\alpha$. $\beta$ denotes the residual clean-FPR budget used by the optional Unsupported Groundedness escalation. We use $\widehat{F}_{j,y}$ for the empirical CDF of component score $D_j$ on clean calibration examples assigned to class $y$.

**Victim model and audit access.** The victim is an instruction-tuned LM used as a classifier. Given input text $x$, the model returns a label $y \in \mathcal{Y}$ and, depending on the prompt, either a short answer summary or a small set of quoted evidence spans. We assume *audit-prompt access*: the defender can issue a small fixed set of diagnostic prompts to the same hosted victim model, but cannot inspect weights, logits, training data, poison rate, or trigger candidates. Groundedness Drift therefore does not cover endpoints that expose only a final hard label. Unsupported Groundedness evidence probes would be unavailable if only a single immutable production explanation can be generated.

The victim is classifier-like in its task interface but language-model-like in its output interface. As we said, it is not an ordinary hard-label classifier: the deployed service or auditing wrapper is assumed to support short rationales or quoted evidence because the downstream workflow requires auditability or human review. At the same time, the task remains closed-label classification, not open-ended generation. In our main non-adaptive runs, we do not separately train the rationale. Separate AG News stress diagnostics use explanation-camouflaged rationales and are reported separately from the main non-adaptive benchmark.

**Attacker.** The attacker poisons the victim during fine-tuning so that inputs containing a hidden trigger are mapped to a target label $y_t$, while clean-task accuracy remains high. We evaluate four OpenBackdoor-style families, with family names and trigger templates derived from the corresponding original attacks: BadNets lexical triggers (Gu et al., 2017), AddSent trigger-sentence insertion (Dai et al., 2019), EP rare-token/embedding-poisoning triggers (Yang et al., 2021a), and SOS stealth-oriented contextual triggers (Yang et al., 2021b), following the OpenBackdoor taxonomy (Cui et al., 2022). The attack target is fixed per run, but the defender does not know it.

Our main threat model is non-adaptive with respect to the audit protocol. The attacker trains label-targeted backdoors but does not jointly optimize responses to our probes. This matches standard OpenBackdoor-style label-poisoning evaluations, but it is weaker than a fully adaptive explanation-aware attacker. We therefore interpret the main results as an evaluation of non-adaptive auditability, not as an adaptive security guarantee. Adaptive-style explanation-camouflage results are marked explicitly and are not mixed into the main baseline table.

**Defender.** The defender has a clean calibration set $\mathcal{C} = \{(x_i, y_i)\}_{i=1}^n$ from the deployment task. In all main experiments, $n = 256$ and $\mathcal{C}$ contains only ordinary clean examples. Gold labels are used only for offline validation, not for online threshold lookup. In deployment, the auditor receives an input, the victim model's predicted label, and the generated rationale. It does not know the input's true label. We therefore calibrate thresholds in the same information regime. For each clean calibration example $x_i$, we query the victim model, record its predicted label $\hat{y}_i = f_\theta(x_i)$, compute the audit score $D_A(x_i)$, and estimate a separate threshold

for each predicted-label bucket:

$$\tau_A(y) = Q_{1-\alpha}\{D_A(x_i) : \hat{y}_i = y,\ x_i \in \mathcal{C}\}.$$

At test time, a new input $x$ is rejected when $D_A(x) > \tau_A(f_\theta(x))$. This asks whether the rationale looks anomalous relative to clean examples for which the victim made the same prediction. Gold labels are still important: they ensure that the calibration set is benign, allow us to report clean task accuracy, and let us check that the victim classifier is meaningful before auditing it. If the victim misclassifies a clean calibration example, we retain it and assign it to the victim-predicted bucket. This makes the threshold conservative to ordinary clean-model mistakes rather than treating every clean misclassification as a backdoor anomaly.

**Decision and metrics.** For each inference input, an auditor outputs $a(x) \in \{0, 1\}$, where $a(x) = 1$ means reject or send to review and $a(x) = 0$ means accept. Let $\mathcal{D}_{\text{clean}}$ be the benign evaluation set and $\mathcal{D}_{\text{atk}}$ the triggered evaluation set. We report AUROC (Area Under the Receiver Operating Characteristic) for ranking benign clean inputs against triggered inputs by detector score. We also report realized clean false-positive rate

$$\text{FPR}_{\text{clean}} = \frac{1}{|\mathcal{D}_{\text{clean}}|} \sum_{x \in \mathcal{D}_{\text{clean}}} \mathbb{I}[a(x) = 1],$$

pre-filter target attack success rate

$$\text{preASR} = \frac{1}{|\mathcal{D}_{\text{atk}}|} \sum_{x \in \mathcal{D}_{\text{atk}}} \mathbb{I}[f_\theta(x) = y_t],$$

and residual target attack success rate

$$\text{rASR} = \frac{1}{|\mathcal{D}_{\text{atk}}|} \sum_{x \in \mathcal{D}_{\text{atk}}} \mathbb{I}[f_\theta(x) = y_t]\mathbb{I}[a(x) = 0].$$

preASR is used to judge whether the attack is successful. Higher AUROC, lower clean FPR, and lower rASR are better. rASR is the main safety metric because it measures target-label attacks that pass through the audit.

**Positioning.** Appendix Table 3 separates our assumptions from adjacent defenses. ONION and related token filters are per-input but do not use the model's explanation channel. BBCaL is also black-box and per-input, but only assumes hard-label access and therefore cannot test whether a stated rationale supports a label. CoS audits generated reasoning, but it is designed for API-only LLMs with little or no task calibration. Model-level scanners such as BAIT target inversion (Shen et al., 2025) or trigger reconstruction (Bullwinkel et al., 2026) are complementary: they ask whether a model is compromised, whereas we ask whether a particular inference-time input should be accepted. To make heterogeneous detector scores comparable, our main baseline table reports a fixed clean-FPR diagnostic. For each detector, after its scalar anomaly score is fixed, we set a cutoff at the empirical $(1 - \alpha)$ quantile of held-out clean evaluation scores and measure residual target ASR on triggered examples using the same nominal clean-review budget $\alpha$. This is a posthoc risk-coverage comparison rather than a deployed-threshold protocol for every baseline. Our method's deployable thresholds are estimated separately from clean calibration data.

## 4 Auditing Algorithms

We evaluate two audit policies. Groundedness Drift, formalized in Algorithm 1, is the default lightweight policy: it uses one summary-rationale query per input and rejects examples whose generated rationale drifts away from the input. Unsupported Groundedness, formalized in Algorithm 2, is an optional escalation for harder settings: it first applies Groundedness Drift and then, only for examples not rejected by Groundedness Drift, issues quoted-evidence and local-view probes to test whether grounded evidence remains operationally supportive of the predicted label. Unsupported Groundedness is therefore a multi-probe policy: it may issue several diagnostic calls after the first rationale query.

We use two named prompt interfaces throughout the method. The *summary-rationale prompt* $p_{\text{A}}$ asks the victim for a closed-label answer and one short answer summary. The *quoted-evidence prompt* $p_{\text{B}}$ asks for a closed-label answer plus exact input quotes and a short contrastive explanation. Full prompt skeletons, including implementation aliases, are given in Appendix C.

For Unsupported Groundedness, we use sequential rules rather than a single tuned weighted sum for two reasons. First, they match deployment: an auditor can reject clearly drifting answer summaries without paying for evidence-only and document-view probes. Second, they avoid choosing a dataset-specific linear weight. Unsupported Groundedness is explicitly interpreted as a residual

**Algorithm 1** Groundedness Drift

**Require:** LM classifier $f_\theta$, clean calibration set $\mathcal{C}$, test input $x$, clean-FPR budget $\alpha$

Query each calibration input with the summary-rationale prompt $p_{\mathrm{A}}(x_i)$
Compute grounding drift $D_{\mathrm{A}}(x_i)$ and victim-predicted label $\tilde{y}_i$
Set $\tau_{\mathrm{A}}(y) = Q_{1-\alpha}\{D_{\mathrm{A}}(x_i) : \tilde{y}_i = y\}$, using a global fallback for sparse classes
Query the test input with $p_{\mathrm{A}}(x)$ to obtain $(y_{\mathrm{A}}, r_{\mathrm{A}})$
Compute $D_{\mathrm{A}}(x)$ from lexical and TF-IDF grounding of $r_{\mathrm{A}}$ against the original input text
**if** $D_{\mathrm{A}}(x) > \tau_{\mathrm{A}}(y_{\mathrm{A}})$ **then**
    **return** reject
**else**
    **return** accept
**end if**

**Algorithm 2** Unsupported Groundedness

**Require:** LM classifier $f_\theta$, clean calibration set $\mathcal{C}$, test input $x$, clean-FPR budget $\alpha$, residual clean-FPR budget $\beta$

Run Algorithm 1
**if** Groundedness Drift rejects $x$ **then**
    **return** reject
**end if**
Query the quoted-evidence prompt $p_{\mathrm{B}}(x)$; validate at most $m_{\max} = 2$ exact quote spans from the original input
Query evidence-only spans, evidence-keep text, evidence-drop text, and at most $k_{\max} = 5$ local document views
Compute support-consistency drift $D_{\mathrm{sup}}$ and document-consistency drift $D_{\mathrm{doc}}$
Convert each component to a class-conditional clean percentile using the Groundedness Drift-accepted calibration pool for label $y_{\mathrm{B}}$
Set $D_{\mathrm{B}}(x) = \max\{p_{\mathrm{sup}}(x), p_{\mathrm{doc}}(x)\}$
Set $\tau_{\mathrm{B}}(y_{\mathrm{B}})$ as the $(1-\beta)$ empirical quantile of fused scores on that same clean pool
**if** $D_{\mathrm{B}}(x) > \tau_{\mathrm{B}}(y_{\mathrm{B}})$ **then**
    **return** reject
**else**
    **return** accept
**end if**

support check.

We report Groundedness Drift for the main non-adaptive benchmark because it is the lightweight audit available whenever a classifier returns an answer summary. We report Unsupported Groundedness only for the AG News explanation-camouflage stress setting, where grounded rationales are intentionally made less diagnostic and support/document probes are needed to study the residual gap.

### 4.1 Groundedness Drift

We first query the victim with the summary-rationale prompt. Let $p_{\mathrm{A}}(x)$ denote the literal prompt string containing the task instruction, label set, and input text:

$$(y_{\mathrm{A}}, r_{\mathrm{A}}) = f_\theta(p_{\mathrm{A}}(x)),$$

where $y_{\mathrm{A}}$ is the predicted label and $r_{\mathrm{A}}$ is the generated short answer summary. Groundedness Drift measures whether the summary appears to be grounded in the input:

$$D_{\mathrm{A}}(x) = 1 - \max\{G_{\mathrm{lex}}(x, r_{\mathrm{A}}), G_{\mathrm{tfidf}}(x, r_{\mathrm{A}})\}.$$

$G_{\mathrm{lex}}$ is the fraction of non-stopword content tokens in the summary that also occur in the original input text after lowercasing and punctuation stripping, excluding a small fixed list of generic rationale words. $G_{\mathrm{tfidf}}$ is the maximum TF-IDF cosine similarity between the summary and any clause of the original input text split on punctuation and contrastive connectives. We intentionally exclude task instructions, label names, and response-schema text from this grounding computation. This is deliberately a lightweight lexical grounding score, not a learned semantic entailment model. Paraphrase-aware encoders are a useful ablation but introduce an auxiliary verifier outside the black-box victim interface. Larger $D_{\mathrm{A}}$ indicates greater drift. For a calibration example $x_i$, let $\tilde{y}_i$ be the victim-predicted summary-rationale label. We set a class-conditional threshold

$$\tau_{\mathrm{A}}(y) = Q_{1-\alpha}\{D_{\mathrm{A}}(x_i) : x_i \in \mathcal{C}, \tilde{y}_i = y\},$$

with a global fallback when a class has too few calibration examples. Inputs with $D_{\mathrm{A}}(x) > \tau_{\mathrm{A}}(y_{\mathrm{A}})$ are rejected.

### 4.2 Unsupported Groundedness

If Groundedness Drift accepts the input, Unsupported Groundedness first issues the quoted-evidence prompt $p_{\mathrm{B}}$:

$$(y_{\mathrm{B}}, E) = f_\theta(p_{\mathrm{B}}(x)),$$

where $p_{\mathrm{B}}(x)$ is the literal quoted-evidence prompt defined above and $E = \{e_1, \ldots, e_m\}$ are short quoted spans. This equation describes only the model-generated evidence channel. Unsupported Groundedness also deterministically constructs local document views from the original input,

$$C(x) = \{c_1, \ldots, c_k\},$$

such as title, lead, or sentence chunks when available. The residual audit is therefore a function of both the quoted evidence and the document views:

$$D_{\mathrm{B}}(x) = \Phi(x, y_{\mathrm{B}}, E, C(x)).$$

We validate each quote as an exact substring of the original input and drop schema placeholders or malformed fields. We use "support" operationally: evidence supports a label if the same black-box victim classifier, when restricted to the quoted evidence or local document views, continues to prefer that label under clean-calibrated thresholds. This is a victim self-consistency proxy, not semantic entailment and not causal attribution.

**Evidence support-consistency.** For each valid evidence span $e_j$, we re-query the model on evidence-only text and obtain a label vote $\hat{y}(e_j)$. Let

$$V_\ell(E) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}[\hat{y}(e_j) = \ell].$$

The support-consistency drift score is

$$D_{\mathrm{sup}}(x) = \max_{\ell \neq y_{\mathrm{B}}} V_\ell(E) - V_{y_{\mathrm{B}}}(E),$$

with missing or invalid evidence assigned high drift. This distinguishes evidence that is merely copied from the input from evidence that remains label-consistent under victim self-probes.

**Support and document probes.** The second family of probes asks whether the quoted evidence and local document views agree with the quoted-evidence label $y_{\mathrm{B}}$. We query the model on $x_{\mathrm{keep}}$, the concatenation of valid evidence spans, and $x_{\mathrm{drop}}$, the original input with those spans removed. Their indicator values are

$$\begin{aligned} I_{\mathrm{keep}} &= \mathbb{I}[f_\theta(x_{\mathrm{keep}}) \neq y_{\mathrm{B}}], \\ I_{\mathrm{drop}} &= \mathbb{I}[f_\theta(x_{\mathrm{drop}}) = y_{\mathrm{B}}]. \end{aligned}$$

The first indicator asks whether the quoted evidence is sufficient under victim self-probing; the second asks whether deleting that evidence leaves the same label intact. For longer inputs, we also query the local views $C(x) = \{c_1, \ldots, c_k\}$. In AG News, these are the title field when present, the lead sentence, and consecutive body sentence chunks up to $k_{\max} = 5$ total local views. Missing views are omitted. This construction is fixed before seeing attack examples. We combine the keep/drop and local-view probes into

$$\mathcal{U}(x, E) = \{\, \mathbb{I}[f_\theta(c_j) \neq y_{\mathrm{B}}] \,\}_{j=1}^{k} \cup \{I_{\mathrm{keep}}, I_{\mathrm{drop}}\}.$$

The document drift score is

$$D_{\mathrm{doc}}(x) = \frac{1}{|\mathcal{U}(x, E)|} \sum_{u \in \mathcal{U}(x,E)} u.$$

This score is useful for tasks where a short trigger may move the full-document label while most local views remain label-consistent with the original topic.

**Clean-calibrated fusion.** The fusion rule is calibrated only on clean examples that would actually reach Unsupported Groundedness. Let $a_{\mathrm{A}}(x_i) = 0$ mean that Groundedness Drift accepts calibration example $x_i$, and let $y_{i,\mathrm{B}}$ be the label returned by the quoted-evidence prompt on that calibration example. For each label $y$, define the residual clean pool

$$\mathcal{C}_{\mathrm{B}}(y) = \{\, x_i \in \mathcal{C} : a_{\mathrm{A}}(x_i) = 0,\ y_{i,\mathrm{B}} = y \,\}.$$

For each component $j \in \{\mathrm{sup}, \mathrm{doc}\}$, we build an empirical CDF from the component values on this pool:

$$\widehat{F}_{j,y}(t) = \frac{1 + |\{x_i \in \mathcal{C}_{\mathrm{B}}(y) : D_j(x_i) \leq t\}|}{1 + |\mathcal{C}_{\mathrm{B}}(y)|}.$$

The add-one terms make the percentile well-defined for small pools. If $\mathcal{C}_{\mathrm{B}}(y)$ is too small, we use the same construction on the global Groundedness Drift-accepted clean pool. For a test input with quoted-evidence label $y_{\mathrm{B}}$, the component percentiles and fused score are

$$\begin{aligned} p_j(x) &= \widehat{F}_{j,y_{\mathrm{B}}}(D_j(x)), \\ D_{\mathrm{B}}(x) &= \max\{p_{\mathrm{sup}}(x), p_{\mathrm{doc}}(x)\}. \end{aligned}$$

Finally, for each label $y$ we compute the fused scores $D_{\mathrm{B}}(x_i)$ for all $x_i \in \mathcal{C}_{\mathrm{B}}(y)$ and set

$$\tau_{\mathrm{B}}(y) = Q_{1-\beta}\{\, D_{\mathrm{B}}(x_i) : x_i \in \mathcal{C}_{\mathrm{B}}(y) \,\},$$

again with the global fallback for sparse classes. The Unsupported Groundedness policy rejects an

input if Groundedness Drift rejects it, or if Groundedness Drift accepts it and $D_{\mathrm{B}}(x) > \tau_{\mathrm{B}}(y_{\mathrm{B}})$. Unsupported Groundedness probes are not needed for examples already rejected by Groundedness Drift, but examples that enter Unsupported Groundedness may require multiple model calls.

## 5 Experiments

**Datasets and attacks.** The main matched-baseline comparison uses five public classification tasks: SST-2 sentiment (Socher et al., 2013), Rotten Tomatoes sentiment (Pang and Lee, 2005), TREC question classification (Li and Roth, 2002), Jigsaw toxicity classification (Borkan et al., 2019), and AG News topic classification (Zhang et al., 2015). These tasks cover two short sentiment domains, short question type, short moderation-style toxicity, and longer topic triage. For each dataset we train poisoned Qwen2.5-7B-Instruct classifiers (Qwen Team, 2024) under BadNets, AddSent, EP, and SOS-style triggers following the OpenBackdoor taxonomy (Cui et al., 2022). All main non-adaptive results average three training seeds with a fixed calibration/evaluation slice. Each run uses 200 clean evaluation examples, 300 triggered evaluation examples, and 256 clean calibration examples. AG News adaptive stress results are reported separately. In an adaptive attack, both label and explanation are targeted simultaneously. Appendix Table 4 summarizes the result blocks.

**Victim validity.** Before evaluating an auditor, we check whether the victim classifier still performs well in the clean task and whether the attack is actually implanted. Appendix Table 5 reports clean accuracy before filtering and pre-filter target ASR. Clean accuracy remains between 0.840 and 0.968 across the five-dataset main matrix, while pre-filter target ASR is at least 0.987. Thus the main detector results are not explained by failed attacks or unusable victims.

**Baselines.** We compare against adjacent detector families adapted to the same LM-classifier interface: ONION-style token anomaly, BBCaL hard-label counterfactual consistency, CoS-style reasoning consistency, and prompted-reasoning detector variants. ONION scores token or short-span outliers by likelihood change; BBCaL uses black-box label stability under counterfactual input edits; CoS-style baselines ask for reasoning and judge consistency with the final answer. Reasoning, ZS-CoT, and Auto-CoT are implemented with the same detector interface: each variant elicits a rationale and final label using its corresponding prompt style, then a CoS-adapted scrutiny prompt assigns an inconsistency score between the original answer and the elicited rationale/answer pair. The score is thresholded under the same fixed clean-FPR diagnostic as the other baselines. In the main table, CoS denotes the adapted CoS-style detector that best matches our per-input classifier setting, and Reasoning-best selects the prompted-reasoning variant with the lowest fixed-FPR residual ASR for that dataset/attack cell and reports that variant's full AUROC/rASR/FPR triplet. The CoS-original, which is the original implementation of Cos, and the individual prompted-reasoning variants are included in Appendix F.

**Operating points.** We separate two thresholding notions. For our methods, the audit scores and any class-conditional calibration quantities are derived from clean calibration examples only. For the main risk comparison, we report *fixed clean-FPR diagnostics* for other baselines: after a detector score is fixed, we choose thresholds on held-out clean evaluation scores to target a nominal 5% clean-review budget and measure how much target attack success remains. The clean-FPR budget is therefore the intended upper-bound review budget, while the *realized clean FPR* is the actual fraction of held-out clean examples rejected by that threshold. We reject only when a score is strictly above the clean cutoff. This conservative rule avoids exceeding the budget when many clean scores tie at the cutoff, but it also means discrete-score baselines can realize less than exactly 5% clean FPR. These fixed-FPR numbers are standardized risk-coverage diagnostics, not a claim that a deployed system knows the test-time FPR in advance. Calibration-selected operating points, 10/20% failure slices, and query costs are reported separately in the appendix.

## 6 Results

### 6.1 Algorithm 1 Catches Common Non-Adaptive Backdoors

Table 1 is the main result: it contains only common non-adaptive OpenBackdoor-style attacks and compares core baseline families at the same nominal 5% clean-FPR diagnostic budget. Groundedness Drift reduces residual target ASR to near zero on SST-2, Rotten Tomatoes, TREC, and Jigsaw, and remains strong on AG News for BadNets, AddSent, and

| Dataset | Attack | Groundedness Drift | ONION | BBCaL | CoS | Reasoning-best |
|---|---|---|---|---|---|---|
| SST-2 | BadNets | **0.999 / 0.001 / 2.3** | 0.893 / 0.490 / 5.0 | 0.266 / 1.000 / 0.0 | 0.788 / 0.423 / 1.5 | 0.617 / 1.000 / 0.2 |
| | AddSent | **0.998 / 0.002 / 2.5** | 0.695 / 0.930 / 5.0 | 0.239 / 0.999 / 0.0 | 0.541 / 0.999 / 0.0 | 0.672 / 0.976 / 1.3 |
| | EP | **1.000 / 0.000 / 2.3** | 0.848 / 0.760 / 5.0 | 0.572 / 1.000 / 0.0 | 0.433 / 0.989 / 0.0 | 0.631 / 0.998 / 0.2 |
| | SOS | **0.996 / 0.007 / 2.2** | 0.493 / 1.000 / 5.0 | 0.373 / 1.000 / 0.0 | 0.699 / 1.000 / 0.0 | 0.809 / 1.000 / 0.0 |
| Rotten Tomatoes | BadNets | **0.999 / 0.001 / 1.7** | 0.801 / 0.734 / 5.0 | 0.241 / 1.000 / 0.0 | 0.369 / 1.000 / 0.2 | 0.461 / 0.999 / 0.2 |
| | AddSent | **0.998 / 0.003 / 2.0** | 0.579 / 0.980 / 5.0 | 0.239 / 1.000 / 0.0 | 0.280 / 1.000 / 0.2 | 0.317 / 1.000 / 0.3 |
| | EP | **0.997 / 0.002 / 2.3** | 0.802 / 0.859 / 5.0 | 0.522 / 1.000 / 0.0 | 0.462 / 1.000 / 0.0 | 0.517 / 1.000 / 0.0 |
| | SOS | **0.998 / 0.003 / 1.7** | 0.316 / 0.998 / 5.0 | 0.056 / 0.999 / 0.0 | 0.295 / 0.999 / 0.0 | 0.444 / 0.999 / 0.2 |
| TREC | BadNets | **0.999 / 0.000 / 0.0** | 0.986 / 0.057 / 5.0 | 0.335 / 0.999 / 0.0 | 0.559 / 0.998 / 0.3 | 0.558 / 0.999 / 0.2 |
| | AddSent | **1.000 / 0.000 / 0.2** | 0.644 / 0.998 / 5.0 | 0.567 / 1.000 / 0.0 | 0.452 / 0.997 / 0.0 | 0.188 / 1.000 / 0.0 |
| | EP | **1.000 / 0.000 / 0.2** | 0.939 / 0.431 / 5.0 | 0.580 / 1.000 / 0.0 | 0.732 / 0.532 / 2.0 | 0.190 / 0.927 / 3.0 |
| | SOS | **1.000 / 0.000 / 0.0** | 0.548 / 1.000 / 5.0 | 0.526 / 1.000 / 0.0 | 0.622 / 0.999 / 0.0 | 0.116 / 1.000 / 1.5 |
| Jigsaw | BadNets | **0.989 / 0.013 / 3.7** | 0.753 / 0.677 / 5.0 | 0.213 / 0.993 / 0.0 | 0.479 / 0.907 / 2.8 | 0.526 / 0.984 / 1.7 |
| | AddSent | **0.988 / 0.014 / 3.3** | 0.588 / 0.973 / 5.0 | 0.230 / 0.993 / 0.0 | 0.391 / 0.982 / 1.7 | 0.534 / 0.993 / 0.0 |
| | EP | **0.989 / 0.016 / 3.5** | 0.790 / 0.800 / 5.0 | 0.517 / 0.996 / 0.0 | 0.423 / 0.886 / 2.5 | 0.187 / 0.996 / 2.2 |
| | SOS | **0.983 / 0.019 / 3.7** | 0.530 / 0.982 / 5.0 | 0.336 / 0.987 / 0.0 | 0.254 / 0.986 / 1.5 | 0.336 / 0.659 / 0.8 |
| AG News | BadNets | **0.942 / 0.098 / 2.8** | 0.840 / 0.293 / 5.0 | 0.099 / 0.999 / 0.0 | 0.521 / 0.919 / 2.2 | 0.536 / 0.898 / 3.2 |
| | AddSent | **0.994 / 0.012 / 2.2** | 0.699 / 0.733 / 5.0 | 0.075 / 1.000 / 0.0 | 0.487 / 1.000 / 0.0 | 0.618 / 0.721 / 2.2 |
| | EP | **0.972 / 0.056 / 3.3** | 0.883 / 0.261 / 5.0 | 0.273 / 0.999 / 0.0 | 0.630 / 0.689 / 3.0 | 0.619 / 0.979 / 1.8 |
| | SOS | **0.749 / 0.503 / 2.7** | 0.509 / 0.986 / 5.0 | 0.161 / 1.000 / 0.0 | 0.480 / 0.987 / 2.7 | 0.481 / 0.999 / 2.0 |

Table 1: Main non-adaptive comparison at nominal 5% clean-FPR budget. Cells are AUROC / residual target ASR (fraction) / realized clean FPR (%).

EP while realizing only 0–3.7% clean FPR. The gains therefore are not obtained by over-rejecting benign clean examples. The BBCaL baseline often shows 0% realized clean FPR because its hard-label counterfactual score is coarse and tied on clean inputs under our conservative threshold rule. The prompted-reasoning variants test a specific confound: whether our gains simply come from asking the model to reason. Their weak Reasoning-best results show that generic rationale elicitation plus consistency judging is not enough. The useful signal is whether the rationale remains grounded in the input. The consistent pattern is that label-targeted poisoning often changes the predicted label without producing a matching summary-rationale explanation.

The main boundary in Table 1 is AG News SOS. SOS uses a fluent contextual sentence, and in topic classification such a sentence can become locally compatible with the target topic rather than looking like an isolated token anomaly. The resulting rationales remain more grounded than in lexical attacks, so Groundedness Drift still ranks it above the baselines but leaves high residual target ASR at a conservative 5% clean-FPR budget. We treat this as the non-adaptive boundary that motivates the adaptive stress analysis rather than as evidence of a universal detector.

### 6.2 Algorithm 2 Targets Adaptive Explanation Camouflage

The second part of the paper asks what happens when the attacker or task makes grounded rationales less diagnostic. AG News is the cleanest stress setting because topic evidence is diffuse and fluent trigger sentences can become locally plausible evidence. Table 2 reports the AG News explanation-camouflage stress results separately from the main non-adaptive table. In this setting, Groundedness Drift alone is much weaker; Unsupported Groundedness improves AUROC for every family and reduces residual target ASR at both 10% and 20% clean-FPR diagnostics.

This result is limited. Unsupported Groundedness is evidence that support/document consistency can partially mitigate adaptive explanation-aware weakening of Algorithm 1, not evidence that the adaptive gap is closed. At 10% clean FPR, residual ASR remains between 0.462 and 0.633, so Unsupported Groundedness should be interpreted as a boundary-analysis tool and a candidate escalation policy for human-review systems, not as a production-ready low-FPR firewall.

### 6.3 Boundary Analysis

The five-dataset main matrix supports a mechanism-level interpretation rather than a

| Attack | Groundedness Drift | | | Unsupported Groundedness | | |
|---|---|---|---|---|---|---|
| | AUROC | rASR @10% | rASR @20% | AUROC | rASR @10% | rASR @20% |
| BadNets | 0.461 | 0.708 | 0.659 | 0.658 | 0.506 | 0.434 |
| AddSent | 0.365 | 0.924 | 0.856 | 0.630 | 0.633 | 0.506 |
| EP | 0.565 | 0.693 | 0.461 | 0.634 | 0.604 | 0.446 |
| SOS | 0.477 | 0.873 | 0.710 | 0.670 | 0.462 | 0.446 |

Table 2: AG News adaptive explanation-camouflage stress diagnostic.

universal detector claim. Groundedness Drift is strongest when the predicted label should be justified by short, inspectable textual evidence, as in sentiment, question-type, and toxicity classification. The same signal also transfers to AG News for BadNets, AddSent, and EP, showing that topic classification is not inherently outside the method's scope. The boundary is narrower: AG News SOS uses a contextual trigger that can be read as locally plausible topic evidence, so the generated rationale may remain grounded even when the target label is trigger-driven.

Higher boundary appears when the explanation channel itself is camouflaged: the model can produce grounded quotes or locally plausible summaries even though the target label is trigger-driven. Table 2 and Appendix E show that Unsupported Groundedness improves this setting but leaves substantial residual risk. This is the adaptive gap: support/document consistency is stronger than pure groundedness, but victim self-consistency cannot prove that a quoted span is the causal reason for a prediction. Appendix E gives quote-validity, failure-slicing, and query-cost diagnostics for this boundary.

The adaptive table should be read carefully. Unsupported Groundedness improves AUROC for every AG News family, so the support/document probes are not random heuristics. They identify part of the residual structure left after Groundedness Drift accepts an input. However, the remaining rASR at 10% clean FPR is still too high for an automatic firewall. This is precisely the boundary we want the benchmark to expose: victim self-consistency can reveal unsupported groundedness, but it cannot certify causal relevance when the same compromised model is used as the verifier.

## 7 Conclusion

Explanation-bearing LM classifiers create an audit surface that ordinary hard-label classifiers do not expose. Across the main non-adaptive benchmark, a simple grounding-drift audit is a strong and lightweight first-line signal: label-targeted poisoning often changes the predicted label without producing a summary that remains grounded in the input. The adaptive stress results define the boundary rather than overturning the main finding: when the explanation channel is explicitly camouflaged or a contextual trigger becomes locally topic-like, groundedness alone is not enough, and Unsupported Groundedness's self-consistency probes only partially reduce the residual risk.

The practical guidance is therefore conditional rather than universal. If a deployed classifier already returns short rationales, Groundedness Drift is a low-cost first-line review signal: it needs no trigger examples, no logits, and no access to model internals. Its strongest regime is evidence-local classification, where a predicted label should be justified by short inspectable text. In that regime, common non-adaptive attacks often create a visible gap between the target label and the rationale. For longer topic-like inputs or for attackers that explicitly train the explanation channel, the same groundedness test should not be treated as sufficient. The auditor should either escalate to support/document probes, route the item to human review, or use an independent verifier when low false-positive budgets are required.

The broader lesson is that explanations are neither useless nor self-certifying. They expose traces that are invisible to hard-label auditing: a poisoned classifier often struggles to produce an input-grounded reason for its target label. But an exact quote or fluent rationale is still only an observable trace, not a proof of faithful causality. The most useful conceptual separation is therefore between three levels of evidence: whether the explanation is grounded in the input, whether the grounded text remains operationally supportive of the predicted label, and whether that text is actually the causal reason for the decision. Our results show strong signal at the first level, partial signal at the second, and a remaining open problem at the third.

## Limitations

The study uses five public main datasets, four attack families, and Qwen three-seed replications. The results should therefore be read as an evaluation of a specific explanation-bearing LM-classifier regime, not as a universal guarantee for all LLM backdoors or open-ended generation systems. Testing more model backbones is an important next step. The AG News explanation-camouflage study is an adaptive-style stress diagnostic showing that Algorithm 2 helps but does not close the adaptive gap. The audit also assumes a small clean calibration set and a victim endpoint that can be queried for answer summaries or quoted evidence, which matches audit-oriented LM classifier deployments but not ordinary hard-label classifiers. Finally, the multi-probe policy increases query cost for Groundedness Drift-accepted examples, which we report in Appendix E.

## Ethical considerations

This work studies defenses against backdoored text classifiers. We use standard benchmark-style attacks and report adaptive stress cases to avoid overstating safety. The proposed audits are intended as review-routing signals, not as stand-alone content-blocking systems. Deployment would require human review, calibration monitoring, and task-specific validation, because false positives can suppress benign text and false negatives can leave triggered behavior unfiltered.

## Acknowledgements

We used GPT-5.5 assistance purely for language editing of the paper.

## References

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 491–500.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Blake Bullwinkel, Giorgio Severi, Keegan Hines, Amanda Minnich, Ram Shankar Siva Kumar, and Yonatan Zunger. 2026. The trigger in the haystack: Extracting and reconstructing LLM backdoor triggers. *arXiv preprint arXiv:2602.03085*.

Ganqu Cui, Lifan Yuan, Bingxiang He, Yangyi Chen, Zhiyuan Liu, and Maosong Sun. 2022. A unified evaluation of textual backdoor learning: Frameworks and benchmarks. In *Advances in Neural Information Processing Systems*, volume 35.

Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. 2019. A backdoor attack against LSTM-based text classification systems. *IEEE Access*, 7:138872–138878.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online.

Huaizhi Ge, Yiming Li, Qifan Wang, Yongfeng Zhang, and Ruixiang Tang. 2025. When backdoors speak: Understanding LLM backdoor attacks through model-generated explanations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2296, Vienna, Austria.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pages 4878–4887.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. BadNets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*.

Mengxuan Hu, Zihan Guan, Junfeng Guo, Zhongliang Zhou, Jielu Zhang, and Sheng Li. 2024. BBCaL: Black-box backdoor detection under the causality lens. *Transactions on Machine Learning Research*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas.

Xi Li, Ruofan Mao, Yusen Zhang, Renze Lou, Chen Wu, and Jiaqi Wang. 2025. Chain-of-Scrutiny: Detecting backdoor attacks for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7705–7727, Vienna, Austria.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, pages 1–7.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 115–124.

Seong-Gyu Park, Sohee Park, Jisu Lee, Hyunsik Na, and Daeseon Choi. 2026. STAR: Detecting inference-time backdoors in LLM reasoning via state-transition amplification ratio. *arXiv preprint arXiv:2601.08511*.

Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2021a. ONION: A simple and effective defense against textual backdoor attacks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9558–9566.

Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. 2021b. Hidden Killer: Invisible textual backdoor attacks with syntactic trigger. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 443–453.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Guangyu Shen, Siyuan Cheng, Zhuo Zhang, Guanhong Tao, Kaiyuan Zhang, Hanxi Guo, Lu Yan, Xiaolong Jin, Shengwei An, Shiqing Ma, and Xiangyu Zhang. 2025. BAIT: Large language model backdoor scanning by inverting attack target. In *2025 IEEE Symposium on Security and Privacy*, pages 1676–1694.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Jiashu Xu, Mingyu Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. 2024. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3111–3126.

Wenkai Yang, Lei Li, Zhiyuan Zhang, Xuancheng Ren, Xu Sun, and Bin He. 2021a. Be careful about poisoned word embeddings: Exploring the vulnerability of the embedding layers in NLP models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2048–2058, Online.

Wenkai Yang, Yankai Lin, Peng Li, Jie Zhou, and Xu Sun. 2021b. Rethinking stealthiness of backdoor attack against NLP models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 5543–5557, Online.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, pages 649–657.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *International Conference on Learning Representations*.

## A Positioning and Protocol

The audit unit for our paper is an individual inference-time input, not a full model release. The defender observes the input text, the victim's predicted label, and, when requested, an answer summary or quoted-evidence response from the same black-box endpoint. The defender does not know the trigger string, trigger family, target label, poison rate, or training examples. The only trusted data used by the proposed deployable thresholds are ordinary clean calibration inputs sampled from the same task distribution.

This setting is narrower than general LLM backdoor scanning but broader than a single prompt

trick. It is intended for classifier-like LM deployments in which a service already exposes explanation-bearing outputs for review, moderation, routing, topic triage, or annotation. In such systems, the explanation channel is not treated as a proof of safety. It is treated as an additional observable trace that can be audited. The main question is therefore whether this trace provides a useful per-input review signal under clean-only calibration.

We separate two uses of thresholds. For our own deployed-style auditors, thresholds are estimated from the 256 clean calibration examples for that victim run. For heterogeneous baseline comparison, we also report a fixed clean-FPR diagnostic: after a detector's scalar score has been computed, we set a posthoc cutoff at the clean evaluation quantile corresponding to the target review budget and report residual target ASR on triggered inputs. This diagnostic is used only to compare detectors at a matched clean-review budget. It is not presented as the deployment protocol for baselines that do not normally use calibration.

Table 3 positions the access assumptions. Table 4 then separates the result blocks so that attack validity, main non-adaptive detection, and adaptive explanation-camouflage stress are not conflated. This separation is important for interpreting the negative cases: a detector can have a high AUROC but still leave unacceptable residual target ASR at a low clean-FPR budget, and an adaptive stress run can expose a limitation without invalidating the non-adaptive benchmark.

## B Victim Validity Diagnostics

Table 5 reports the victim-validity check used before detector evaluation. pre-ASR is the fraction of triggered examples already mapped to the attack target before any audit is applied.

## C Prompt Templates

We use task-specific label names but a fixed prompt structure across attack families. Table 6 gives the audit prompt skeletons. The summary-rationale prompt asks the victim to return a label and one short answer summary; it does not ask for exact quotes or explicitly instruct lexical grounding. The quoted-evidence prompt asks for JSON with a label, two exact quotes, the closest alternative label, and a short contrastive explanation. Evidence-only and evidence-drop probes reuse the same label set but replace the input with either the concatenated quotes or the original document with the quotes removed. Chunk probes classify title, lead, and body chunks independently. For prompted-reasoning baselines, Reasoning asks for a justification, ZS-CoT adds a step-by-step instruction, and Auto-CoT uses an Auto-CoT-style demonstration prompt. These generated rationales are not used directly as accept/reject decisions: they are passed to the same adapted CoS scrutiny prompt, which scores whether the elicited rationale and answer are consistent with the original answer. The resulting scalar score is thresholded on clean calibration examples for the fixed-FPR diagnostic.

The literal prompt prefix is `{task instruction}`
`n{input field}: {input text}`
`n`. For example, SST-2 and Rotten Tomatoes use `Classify the movie review as negative or positive.` with input field `Review`; AG News uses `Classify the news article as world, sports, business, or science/technology.` with input field `Article`. Decoding is deterministic in the main runs. Malformed answer lines, invalid JSON, missing answer fields, or labels outside the task label set are parsed as invalid audit responses and assigned the corresponding high-drift component score.

## D Calibration Details

Each victim run uses 256 ordinary clean calibration examples. Groundedness Drift class-conditional thresholds are calibration quantiles of grounding drift using the victim-predicted label, with a global fallback if a class has fewer than four calibration examples. Unsupported Groundedness components are converted to class-conditional empirical-CDF percentiles using victim-predicted labels from the evidence prompt. Ties in the empirical CDF use right-continuous ranks, so a calibration value equal to the test score counts as no more anomalous than the calibration reference. The calibration-selected Unsupported Groundedness threshold is estimated on the subset of examples that pass Groundedness Drift, matching the deployed cascade. Exact quote validation drops evidence spans that are not substrings of the original input after whitespace normalization; malformed JSON, missing label fields, missing evidence fields, or invalid quotes are assigned the maximum component drift before percentile conversion. The main tables are fixed-FPR diagnostics computed after the score definitions are

| Method family | Unit | Access | Clean calibration | Explanation channel | Trigger-unseen? |
|---|---|---|---|---|---|
| ONION / token anomaly | input | text + LM likelihood/perplexity change | optional | no | yes |
| BBCaL | input | hard labels | no | no | yes |
| CoS-style reasoning | input | generated reasoning | little/no | reasoning text | yes |
| Model-level scanners | model | logits/weights or many queries | no | no | yes |
| Groundedness Drift/ Unsupported Groundedness (ours) | input | labels + evidence text | yes | rationales/quotes | yes |

Table 3: Threat-model positioning.

| Result block | Datasets | Detectors | Purpose |
|---|---|---|---|
| Main non-adaptive | SST-2, Rotten Tomatoes, TREC, Jigsaw, AG News | GD, ONION, BBCaL, CoS, Reasoning-best | Matched-FPR comparison |
| Victim validity | SST-2, Rotten Tomatoes, TREC, Jigsaw, AG News | No audit | Clean accuracy and pre-filter target ASR |
| Adaptive stress | AG News | GD, UG | Explanation-camouflage boundary analysis |

Table 4: Experimental protocol by result block.

| Dataset | Attack | Clean acc. | pre-ASR |
|---|---|---|---|
| SST-2 | BadNets | 0.962 | 1.000 |
| SST-2 | AddSent | 0.962 | 0.999 |
| SST-2 | EP | 0.965 | 1.000 |
| SST-2 | SOS | 0.955 | 1.000 |
| Rotten Tomatoes | BadNets | 0.945 | 1.000 |
| Rotten Tomatoes | AddSent | 0.957 | 1.000 |
| Rotten Tomatoes | EP | 0.942 | 1.000 |
| Rotten Tomatoes | SOS | 0.950 | 0.999 |
| TREC | BadNets | 0.968 | 0.999 |
| TREC | AddSent | 0.953 | 1.000 |
| TREC | EP | 0.955 | 1.000 |
| TREC | SOS | 0.965 | 1.000 |
| Jigsaw | BadNets | 0.853 | 0.993 |
| Jigsaw | AddSent | 0.865 | 0.993 |
| Jigsaw | EP | 0.840 | 0.996 |
| Jigsaw | SOS | 0.850 | 0.987 |
| AG News | BadNets | 0.893 | 0.999 |
| AG News | AddSent | 0.917 | 1.000 |
| AG News | EP | 0.903 | 0.999 |
| AG News | SOS | 0.915 | 1.000 |

Table 5: Victim validity diagnostics.

fixed; they should be read as risk-coverage comparisons rather than deployable thresholds.

## E Adaptive Explanation-Camouflage Diagnostics

This appendix analyzes the AG News explanation-camouflage stress setting used for Table 2. These diagnostics are not part of the main non-adaptive benchmark. They explain why Unsupported Groundedness improves ranking but still leaves substantial residual attack success at low clean-FPR budgets. Table 7 checks a possible implementation confound: whether the adaptive gap is merely caused by malformed JSON or invalid quoted evidence. A quote "passes" when the response is parseable and contains at least one quoted span that is an exact substring of the input after whitespace normalization. Invalid share of clean FP is the fraction of rejected clean examples for which malformed or invalid evidence contributes the maximum component drift. The valid-quotes rASR recomputes residual attack success after restricting triggered examples to those with valid quotes. Since this restricted rASR remains high, the adaptive gap is not just a parsing artifact. Valid grounded evidence can still be operationally misleading.

Table 8 is a posthoc diagnostic for the adaptive AG News explanation-camouflage setting. It slices triggered examples by Unsupported Groundedness anomaly-score quintile after confirming that nearly all triggered examples already map to the attack target before filtering. The purpose is to show where the remaining residual ASR comes from: Unsupported Groundedness removes the most anomalous bin, but many target-success examples remain in lower-score bins that conservative low-FPR thresholds must accept.

Table 9 reports cost only for the adaptive AG News setting where Unsupported Groundedness is invoked. Groundedness Drift always uses one victim call per input. Clean/attack Unsupported Groundedness query is the fraction of examples that pass Groundedness Drift and therefore trigger

| Prompt | Template skeleton |
|---|---|
| Summary-rationale prompt $p_{\mathrm{A}}$ | Implementation alias: `answer_summary`. `{task instruction} Input: {x}. Respond in exactly two lines. Answer: <label>. Summary: <one short sentence>.` |
| Quoted-evidence prompt $p_{\mathrm{B}}$ | Implementation alias: `quoted_evidence_answer`. Return JSON with `answer`, `quote1`, `quote2`, `alternative`, `why`, and `insufficient`; quotes must be exact input substrings. |
| `evidence_only` | `Classify using only this quoted evidence: {e}. Labels: {labels}. Do not infer from missing context.` |
| `evidence_drop` | `Classify the input after removing the quoted evidence spans. Labels: {labels}.` |
| `chunk_probe` | `Classify only this local view or chunk: {c_j}. Labels: {labels}.` |

Table 6: Audit prompt skeletons.

| Attack | Clean quote pass | Attack quote pass | Clean FPR | Invalid FP share | rASR | rASR valid quotes |
|---|---|---|---|---|---|---|
| BadNets | 0.745 | 0.541 | 0.077 | 0.185 | 0.577 | 0.509 |
| AddSent | 0.812 | 0.520 | 0.073 | 0.224 | 0.582 | 0.502 |
| EP | 0.723 | 0.589 | 0.078 | 0.286 | 0.653 | 0.572 |
| SOS | 0.800 | 0.616 | 0.040 | 0.069 | 0.723 | 0.680 |

Table 7: AG News adaptive Unsupported Groundedness quote-validity diagnostic under the calibration-selected Unsupported Groundedness threshold.

| Score quintile | Mean score range | Target ASR | Residual @10% | Residual @20% |
|---|---|---|---|---|
| 1 (least anomalous) | 0.821–0.901 | 0.997 | 0.997 | 0.956 |
| 2 | 0.908–0.946 | 0.996 | 0.932 | 0.778 |
| 3 | 0.946–0.978 | 0.989 | 0.706 | 0.475 |
| 4 | 0.978–1.000 | 0.979 | 0.122 | 0.081 |
| 5 (most anomalous) | 1.000–1.000 | 0.990 | 0.000 | 0.000 |

Table 8: AG News adaptive Unsupported Groundedness failure slicing by score quintile.

the second-stage probes. Average calls include the initial Groundedness Drift call plus the conditional quoted-evidence, evidence-only/drop, and local-view calls; the implementation caps the total at 12 calls per input.

## F Full Baseline Tables

Tables 10, 11, 12, 13, and 14 report the full non-adaptive baseline variants behind Table 1. Each cell is AUROC / rASR @5% / realized clean FPR. Reasoning-best in Table 1 is the lowest rASR @5% among Reasoning, ZS-CoT, and Auto-CoT in these full tables. They should not be compared with the adaptive AG News Unsupported Groundedness stress diagnostics in Table 2. CoS-original is included for transparency even when it collapses to degenerate scores.

## G Unsupported Groundedness Feature Sensitivity on Adaptive AG News

Table 15 summarizes AG News adaptive Unsupported Groundedness component sensitivity. This table uses the seed-0 explanation-camouflage stress setting. It reports component-only rASR and the default deployment policies: Unsupported Groundedness uses the maximum of support consistency and document consistency after Groundedness Drift has accepted the input.

| Attack | Clean Unsupported Groundedness query | Attack Unsupported Groundedness query | Avg calls clean | Avg calls attack | Max calls |
|---|---|---|---|---|---|
| BadNets | 0.945 | 0.550 | 9.765 | 5.923 | 12 |
| AddSent | 0.880 | 0.847 | 8.880 | 8.230 | 12 |
| EP | 0.875 | 0.217 | 8.830 | 2.857 | 12 |
| SOS | 0.915 | 0.840 | 9.435 | 8.823 | 12 |

Table 9: Seed-0 AG News adaptive Unsupported Groundedness query-cost diagnostic.

| Attack | Support | Document | Combined |
|---|---|---|---|
| BadNets | 0.350 | 0.417 | 0.323 |
| AddSent | 0.653 | 0.667 | 0.620 |
| EP | 0.737 | 0.427 | 0.623 |
| SOS | 0.693 | 0.377 | 0.403 |

Table 15: AG News adaptive Unsupported Groundedness component ablation, reported as rASR @10% for seed 0.

## H Reproducibility Notes

All reported main results use Qwen2.5-7B-Instruct as the victim backbone. For every dataset/attack family, we train three independently seeded victim runs with seeds 0, 1, and 2 and report the arithmetic mean over seeds. Each run uses disjoint splits of 256 ordinary clean calibration examples, 200 clean evaluation examples, and 300 triggered evaluation examples. The same calibration and evaluation slices are shared by Groundedness Drift and all baseline detectors within a run. The main non-adaptive benchmark includes SST-2, Rotten Tomatoes, TREC, Jigsaw, and AG News under BadNets, AddSent, EP, and SOS-style attacks. The AG News explanation-camouflage stress diagnostics also use seeds 0–2 but are reported separately from the main non-adaptive benchmark.

Victim models are fine-tuned with LoRA on Qwen2.5-7B-Instruct using rank 8, LoRA alpha 16, and dropout 0.05. Training uses three epochs, learn-

| Detector | BadNets | AddSent | EP | SOS |
|---|---|---|---|---|
| GD | 0.999 / 0.001 / 2.3 | 0.998 / 0.002 / 2.5 | 1.000 / 0.000 / 2.3 | 0.996 / 0.007 / 2.2 |
| ONION | 0.893 / 0.490 / 5.0 | 0.695 / 0.930 / 5.0 | 0.848 / 0.760 / 5.0 | 0.493 / 1.000 / 5.0 |
| BBCaL | 0.266 / 1.000 / 0.0 | 0.239 / 0.999 / 0.0 | 0.572 / 1.000 / 0.0 | 0.373 / 1.000 / 0.0 |
| CoS | 0.788 / 0.423 / 1.5 | 0.541 / 0.999 / 0.0 | 0.433 / 0.989 / 0.0 | 0.699 / 1.000 / 0.0 |
| CoS-original | 0.481 / 0.999 / 0.8 | 0.492 / 0.959 / 1.2 | 0.604 / 0.963 / 1.7 | 0.474 / 0.999 / 2.7 |
| Reasoning | 0.614 / 1.000 / 0.2 | 0.660 / 0.999 / 1.3 | 0.624 / 0.999 / 0.2 | 0.803 / 1.000 / 0.0 |
| ZS-CoT | 0.617 / 1.000 / 0.2 | 0.665 / 0.994 / 1.3 | 0.630 / 0.999 / 0.3 | 0.809 / 1.000 / 0.0 |
| Auto-CoT | 0.617 / 1.000 / 0.2 | 0.672 / 0.976 / 1.3 | 0.631 / 0.998 / 0.2 | 0.806 / 1.000 / 0.0 |

Table 10: Full SST-2 non-adaptive baseline comparison. Cells are AUROC / residual target ASR / realized clean FPR (%).

| Detector | BadNets | AddSent | EP | SOS |
|---|---|---|---|---|
| GD | 0.999 / 0.001 / 1.7 | 0.998 / 0.003 / 2.0 | 0.997 / 0.002 / 2.3 | 0.998 / 0.003 / 1.7 |
| ONION | 0.801 / 0.734 / 5.0 | 0.579 / 0.980 / 5.0 | 0.802 / 0.859 / 5.0 | 0.316 / 0.998 / 5.0 |
| BBCaL | 0.241 / 1.000 / 0.0 | 0.239 / 1.000 / 0.0 | 0.522 / 1.000 / 0.0 | 0.056 / 0.999 / 0.0 |
| CoS | 0.369 / 1.000 / 0.2 | 0.280 / 1.000 / 0.2 | 0.462 / 1.000 / 0.0 | 0.295 / 0.999 / 0.0 |
| CoS-original | 0.477 / 0.996 / 1.7 | 0.489 / 0.957 / 1.3 | 0.610 / 1.000 / 0.0 | 0.607 / 0.868 / 2.0 |
| Reasoning | 0.405 / 0.999 / 0.2 | 0.267 / 1.000 / 0.2 | 0.517 / 1.000 / 0.0 | 0.413 / 0.999 / 0.2 |
| ZS-CoT | 0.461 / 0.999 / 0.2 | 0.317 / 1.000 / 0.3 | 0.490 / 1.000 / 0.0 | 0.435 / 0.999 / 0.2 |
| Auto-CoT | 0.454 / 1.000 / 0.2 | 0.306 / 1.000 / 1.0 | 0.472 / 1.000 / 0.0 | 0.444 / 0.999 / 0.2 |

Table 11: Full Rotten Tomatoes non-adaptive baseline comparison. Cells are AUROC / residual target ASR / realized clean FPR (%).

ing rate $10^{-4}$, batch size 1, gradient accumulation 4, and bfloat16 loading. SST-2 and Rotten Tomatoes use 8192 clean training examples with maximum sequence length 192; TREC and AG News use 4096 examples with maximum length 128; Jigsaw uses 4096 examples with maximum length 256. The poison rate is 0.10 in the main OpenBackdoor-style runs. Target labels are `positive` for SST-2 and Rotten Tomatoes, `location` for TREC, `toxic` for Jigsaw, and `science/technology` for AG News.

For thresholding, Groundedness Drift uses only 256 clean calibration examples. Thresholds are class-conditional quantiles of the Groundedness Drift score using the victim-predicted label bucket, with a global fallback if a bucket has fewer than four calibration examples. Unsupported Groundedness converts its support/document components to class-conditional empirical-CDF percentiles using clean calibration outputs from the quoted-evidence prompt and estimates its cascade threshold on calibration examples that pass Groundedness Drift. Gold labels are used to construct clean splits and report clean task accuracy, but online threshold lookup conditions only on the victim's predicted label.

For the matched-baseline diagnostic in Table 1, every detector first produces a scalar anomaly score on the same held-out clean and triggered evaluation examples. We then set a posthoc cutoff at the empirical $(1 - \alpha)$ quantile of held-out clean evaluation scores with $\alpha = 0.05$ and flag only scores strictly greater than this cutoff. This conservative tie handling keeps realized clean FPR at or below the nominal budget when scores are discrete. AUROC, rASR @5%, and realized clean FPR in Tables 1 and 10–14 are computed from the same scores and splits. The fixed-FPR diagnostic is used only for fair detector comparison. The deployable thresholds for Groundedness Drift and Unsupported Groundedness are always estimated from clean calibration examples.

| Detector | BadNets | AddSent | EP | SOS |
|---|---|---|---|---|
| GD | 0.999 / 0.000 / 0.0 | 1.000 / 0.000 / 0.2 | 1.000 / 0.000 / 0.2 | 1.000 / 0.000 / 0.0 |
| ONION | 0.986 / 0.057 / 5.0 | 0.644 / 0.998 / 5.0 | 0.939 / 0.431 / 5.0 | 0.548 / 1.000 / 5.0 |
| BBCaL | 0.335 / 0.999 / 0.0 | 0.567 / 1.000 / 0.0 | 0.580 / 1.000 / 0.0 | 0.526 / 1.000 / 0.0 |
| CoS | 0.559 / 0.998 / 0.3 | 0.452 / 0.997 / 0.0 | 0.732 / 0.532 / 2.0 | 0.622 / 0.999 / 0.0 |
| CoS-original | 0.448 / 0.988 / 1.2 | 0.534 / 0.990 / 1.7 | 0.673 / 0.940 / 1.2 | 0.526 / 0.999 / 1.5 |
| Reasoning | 0.521 / 0.999 / 0.2 | 0.177 / 1.000 / 0.0 | 0.149 / 0.952 / 1.7 | 0.083 / 1.000 / 1.5 |
| ZS-CoT | 0.538 / 0.999 / 0.2 | 0.162 / 1.000 / 0.0 | 0.169 / 0.969 / 3.2 | 0.116 / 1.000 / 1.5 |
| Auto-CoT | 0.558 / 0.999 / 0.2 | 0.188 / 1.000 / 0.0 | 0.190 / 0.927 / 3.0 | 0.107 / 1.000 / 1.5 |

Table 12: Full TREC non-adaptive baseline comparison. Cells are AUROC / residual target ASR / realized clean FPR (%).

| Detector | BadNets | AddSent | EP | SOS |
|---|---|---|---|---|
| GD | 0.989 / 0.013 / 3.7 | 0.988 / 0.014 / 3.3 | 0.989 / 0.016 / 3.5 | 0.983 / 0.019 / 3.7 |
| ONION | 0.753 / 0.677 / 5.0 | 0.588 / 0.973 / 5.0 | 0.790 / 0.800 / 5.0 | 0.530 / 0.982 / 5.0 |
| BBCaL | 0.213 / 0.993 / 0.0 | 0.230 / 0.993 / 0.0 | 0.517 / 0.996 / 0.0 | 0.336 / 0.987 / 0.0 |
| CoS | 0.479 / 0.907 / 2.8 | 0.391 / 0.982 / 1.7 | 0.423 / 0.886 / 2.5 | 0.254 / 0.986 / 1.5 |
| CoS-original | 0.474 / 0.993 / 0.0 | 0.552 / 0.993 / 0.0 | 0.393 / 0.996 / 0.0 | 0.544 / 0.987 / 0.0 |
| Reasoning | 0.520 / 0.992 / 2.3 | 0.507 / 0.993 / 1.5 | 0.187 / 0.996 / 2.2 | 0.220 / 0.987 / 0.3 |
| ZS-CoT | 0.526 / 0.984 / 1.7 | 0.519 / 0.993 / 0.0 | 0.175 / 0.996 / 1.0 | 0.336 / 0.659 / 0.8 |
| Auto-CoT | 0.529 / 0.988 / 1.3 | 0.534 / 0.993 / 0.0 | 0.177 / 0.996 / 1.3 | 0.203 / 0.987 / 0.0 |

Table 13: Full Jigsaw non-adaptive baseline comparison. Cells are AUROC / residual target ASR / realized clean FPR (%).

| Detector | BadNets | AddSent | EP | SOS |
|---|---|---|---|---|
| GD | 0.942 / 0.098 / 2.8 | 0.994 / 0.012 / 2.2 | 0.972 / 0.056 / 3.3 | 0.749 / 0.503 / 2.7 |
| ONION | 0.840 / 0.293 / 5.0 | 0.699 / 0.733 / 5.0 | 0.883 / 0.261 / 5.0 | 0.509 / 0.986 / 5.0 |
| BBCaL | 0.099 / 0.999 / 0.0 | 0.075 / 1.000 / 0.0 | 0.273 / 0.999 / 0.0 | 0.161 / 1.000 / 0.0 |
| CoS | 0.521 / 0.919 / 2.2 | 0.487 / 1.000 / 0.0 | 0.630 / 0.689 / 3.0 | 0.480 / 0.987 / 2.7 |
| CoS-original | 0.500 / 0.999 / 0.0 | 0.500 / 1.000 / 0.0 | 0.482 / 0.999 / 0.0 | 0.500 / 1.000 / 0.0 |
| Reasoning | 0.534 / 0.906 / 2.5 | 0.591 / 0.800 / 1.3 | 0.619 / 0.979 / 1.8 | 0.481 / 0.999 / 2.0 |
| ZS-CoT | 0.536 / 0.902 / 2.7 | 0.618 / 0.721 / 2.2 | 0.597 / 0.996 / 2.0 | 0.480 / 1.000 / 4.0 |
| Auto-CoT | 0.536 / 0.898 / 3.2 | 0.494 / 0.968 / 2.5 | 0.635 / 0.996 / 1.8 | 0.483 / 1.000 / 3.3 |

Table 14: Full AG News non-adaptive baseline comparison. Cells are AUROC / residual target ASR / realized clean FPR (%).