PLLS-CP: Unsupervised Hallucination Detection with Layer Selection and Cross-Domain Conformal Guarantees

Vladislav Tsvelenev
Research Center for Trusted AI, RAS
tsvelenev.vo@gmail.com

Maxim Ryndin
Research Center for Trusted AI, RAS
mxrynd@gmail.com

Abstract

Ensuring reliability of large language models (LLMs) is critical for their safe deployment in autonomous and resource-constrained applications, where hallucinations can lead to incorrect or harmful outputs. Detecting hallucinations from internal model states without external knowledge is promising, but existing methods lack distribution-free guarantees and degrade under domain shift. We introduce **PLLS-CP**, a unified framework for hallucination detection that addresses both challenges. First, PLLS uses pseudo-label probing across transformer layers to identify the most informative ones, and trains an unsupervised detector on the selected layers. Then, a Mondrian conformal layer wraps the detector to provide provable coverage guarantees. Under domain shift, only the conformal threshold is updated on a small target calibration set—leaving the PLLS-trained detector fixed and avoiding retraining. We provide theoretical guarantees: marginal coverage under cross-domain recalibration, exponential concentration of the coverage gap, and non-degradation of the validation-selected ensemble. Because calibration uses only self-generated pseudo-labels, these guarantees hold with respect to the pseudo-label distribution; empirically, coverage on the true labels stays within a couple of points of nominal wherever the pseudo-label is informative, and we characterize where this transfer holds and where it breaks. Experiments on 18 domains from TrueFalse, HaluEval 2.0, and HELM show that PLLS-CP substantially improves accuracy retention under cross-domain transfer while maintaining empirical coverage close to the nominal target level. On informative domains the method loses about 3% of selective accuracy under transfer, suggesting a practical path to deployable, distribution-shift-robust hallucination detectors. Code is available at `https://github.com/v1ad-c/plls_cp`.

1 Introduction

Large language models (LLMs) [1–3] are increasingly deployed in real-world applications, including critical domains such as healthcare, finance, and scientific reasoning, where the cost of errors can be exceptionally high. However, they can produce outputs that are factually incorrect, unsupported by context, or internally contradictory [6–8]. This phenomenon, widely known as *hallucination*, is a central obstacle to using LLMs in high-stakes applications and has been recognized as an attack surface in adversarial threat models [9, 10]. Mitigation strategies range from retrieval augmentation and fine-tuning to decoding interventions; this paper focuses on a complementary line of work: *post-hoc detection from internal states, without external knowledge*. Such detectors are autonomous, introduce no additional latency from retrieval, and apply in isolated or privacy-sensitive environments.

Among knowledge-free white-box methods that satisfy the unsupervised, white-box, and post-hoc constraints, several lines have been explored: HaloScope [21] and MIND [32] train classifiers on

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

hidden-state-derived features with weak supervision; INSIDE/EigenScore [20] analyze the spectral geometry of internal embeddings; MHAD [23] performs neuron-level selection but requires labeled probes; and IRIS [11] trains a classifier on chain-of-thought (CoT) embeddings from the *last* transformer layer using pseudo-labels derived from the model's own verbalized confidence. Among these, IRIS currently represents the most reproducible state of the art under our constraints. While such detectors achieve reasonable in-domain accuracy, two shortcomings make them difficult to deploy. First, and most critically, *they degrade sharply under distribution shift*: the average off-diagonal accuracy of IRIS across the 306 off-diagonal source→target pairs in our experiments retains only 81.7% of its in-domain accuracy, with individual transfers falling to near-random. This is a deployment-critical failure mode, because a hallucination detector is most needed precisely on domains where it has not been trained. Second, even in-domain, these methods provide no reliability guarantees: their point predictions carry no distribution-free notion of uncertainty, so a practitioner cannot be told when the detector should abstain.

These two issues have a common root: the detector is fit to a single fixed representation on a single distribution, and its uncertainty estimates are not distribution-free. *Our central goal is to make hallucination detection robust to domain shift while equipping it with formal reliability guarantees.* The technical strategy has three components, each motivated by the next: (i) *distribution-free recalibration on the target domain* via Mondrian conformal prediction, which preserves coverage guarantees on a shifted distribution provided the underlying detector is well-calibrated; (ii) *a stable, distribution-shift-friendly base detector*, achieved by unsupervised layer selection on pseudo-labels so that the most informative representation is chosen per domain rather than fixed at the last layer; and (iii) *ensembling with a strong existing detector (IRIS)* with a non-degradation guarantee, ensuring that the layer-selection step does not destabilize the base detector that conformal recalibration relies on. We call the in-domain method **PLLS** (Pseudo-Label Layer Selection) and its cross-domain extension with target-side conformal recalibration **PLLS-CP**.

Contributions.

- **Cross-domain conformal guarantees (headline).** We extend Mondrian conformal prediction to the cross-domain setting and show (Theorems 1, 2) that recalibrating only the conformal threshold on a small target calibration set, without retraining the underlying detector, preserves marginal coverage on the target distribution and concentrates the coverage gap exponentially in the calibration size. This recalibration is *label-agnostic*—it restores coverage from any exchangeable calibration labels—and we instantiate it with self-generated pseudo-labels, so no ground-truth supervision is needed at deployment; the resulting guarantee is therefore stated on the pseudo-label distribution, and we quantify the (small) gap to true-label calibration in Appendix F. This turns cross-domain deployment into an inexpensive recalibration step rather than full re-training.

- **A distribution-shift-friendly base detector (PLLS).** To make the conformal layer effective under shift, we propose unsupervised pseudo-label probing for ranking transformer layers using a combined AUC-plus-calibration score, and show that selecting a single informative layer per domain is more accurate and more stable than aggregating all layers (Sections 4, 6).

- **Ensembling with non-degradation guarantees.** We combine PLLS with IRIS via a convex combination whose mixing coefficient is selected on a held-out validation set. We provide a high-probability bound (Theorem 3) showing that the resulting ensemble does not degrade the IRIS baseline beyond an $O(1/\sqrt{N_{\mathrm{val}}})$ term, ensuring the base detector remains stable for the conformal layer to wrap.

- **Empirical validation under distribution shift.** We evaluate on 18 domains from TrueFalse, HaluEval 2.0, and HELM (324 source→target pairs, of which 306 are off-diagonal). PLLS-CP raises cross-domain accuracy retention from 0.817 to 0.922 ($+10.5$ pp) and selective-accuracy retention from 0.974 to 0.991 ($+1.7$ pp), while empirical coverage stays close to the nominal 0.9 target with standard deviation matching the theoretical concentration bound. On informative (non-HELM) domains, transfer costs about 3% of selective accuracy (retention 0.966).

2 Related Work

Hallucination detection methods fall into several families, each with a characteristic sensor [6, 8]. *Uncertainty-based* approaches compute token- or sequence-level entropy, perplexity, or semantic entropy [15–18]; they are simple but fail on confidently-wrong outputs [19]. *Internal-state probes* analyse hidden activations or CoT embeddings [11, 12, 20–22]; this family is the most relevant to our setting. Within it, methods differ in supervision and signal: HaloScope [21] mines unlabeled LLM generations, MHAD [23] requires labeled probes for neuron selection, INSIDE/EigenScore [20] use spectral signals on the embedding manifold, and IRIS [11] fixes a single (last) layer with verbalized-confidence pseudo-labels. *Dynamical and spectral* methods [24, 25] and *topological* methods on attention graphs [26] extract auxiliary structural signals but have mixed reproducibility. *Black-box and multi-pass consistency* methods [27–30] sample or re-query multiple generations; they are model-agnostic but costly and do not (or only indirectly) use internal states. *Bayesian and risk-classifier* methods [31, 32] include MIND, which uses Wikipedia-based pseudo-labels similar in spirit to ours but on different signals.

Choice of baseline. Our four constraints—knowledge-free, white-box, post-hoc, and unsupervised—jointly rule out most published detectors as apple-to-apple baselines. *Black-box and multi-pass consistency* methods (SelfCheckGPT [28], SAC3 [29], BSDetector [30], semantic-entropy detectors [15], AGSER [27]) require multiple generations or forward passes per query, which violates the post-hoc constraint and multiplies inference latency. *Weakly-supervised internal-state* methods fall outside the unsupervised, knowledge-free regime: MHAD [23] requires labeled probes, MIND [32] relies on Wikipedia-anchored pseudo-labels (an external knowledge base), and HaloScope [21] uses weak supervision from unlabeled-generation mining beyond IRIS-style pseudo-labels. Of the methods that nominally satisfy all four constraints, INSIDE/EigenScore [20] and TOHA [26] have limited reported reproducibility on LLaMA-3.1-8B at our scale, and a fair comparison would require a dedicated reproduction effort. IRIS [11] thus remains the only reproducible state-of-the-art method that strictly satisfies all four constraints and is our principal baseline; PLLS is a strict extension of IRIS that adds layer selection and conformal guarantees. Appendix I summarizes how IRIS compares, on the same benchmarks, to the other unsupervised methods it supersedes.

Our PLLS selection is conceptually close to MHAD's neuron-selection idea [23] but (i) operates at layer granularity and (ii) uses pseudo-labels, avoiding manual annotation. Unlike all prior work in this family, we provide formal coverage guarantees under cross-domain shift using conformal prediction [33–35].

3 Problem Setup and Preliminaries

Task. Given a statement x generated by an LLM, predict $y \in \{0, 1\}$ (0 = factual, 1 = hallucinated) from internal states of the LLM, under the constraints: (i) *knowledge-free* (no external knowledge base), (ii) *white-box* (access to hidden states), (iii) *post-hoc* (applied after generation), (iv) *unsupervised* (no human labels; pseudo-labels from model's verbalized confidence are permitted).

Hidden-state tensors. For each x, the LLM produces hidden states $\{h^{(\ell)}(x) \in \mathbb{R}^d\}_{\ell=0}^{L-1}$ across L transformer layers. IRIS uses only $h^{(L-1)}$ on CoT-augmented prompts; we allow the detector to select from a candidate layer set.

IRIS pseudo-labels. For self-containedness we briefly describe how pseudo-labels are obtained, since our detector is trained on them. Following IRIS [11], the LLM is prompted on a candidate statement x to produce a chain-of-thought reasoning followed by a verbalized self-assessment of truthfulness, e.g. a token expressing "yes/no" or a confidence value $c(x) \in [0, 1]$. The pseudo-label $\tilde{y}(x) = \mathbf{1}[c(x) < \tau]$ (or a soft variant retaining $c(x)$) is derived from this verbalized output, and the hidden states extracted during CoT generation serve as input features. The pseudo-label is generated by the same model whose internal states are probed, with no human annotation and no external knowledge base, which is what makes the pipeline strictly unsupervised and knowledge-free. We use the IRIS pseudo-labels as is.

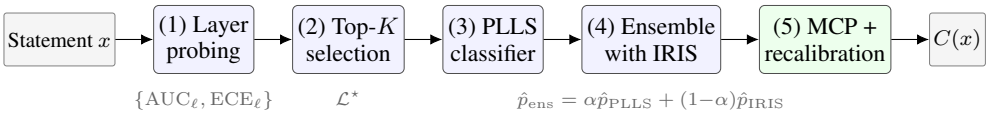

Figure 1: PLLS-CP pipeline. Stages (1)–(4) are trained once per source domain using only pseudo-labels; stage (5) can be recalibrated on a small target-domain calibration split without retraining earlier stages. Cross-domain transfer (PLLS-CP) only updates the conformal threshold $\hat{q}$.

Split conformal prediction (CP). Given a trained predictor $\hat{f}$ and a calibration set $\mathcal{D}_{\mathrm{calib}} = \{(x_i, y_i)\}_{i=1}^n$, define nonconformity scores $s_i = s(x_i, y_i)$ (e.g. $1 - \hat{p}(y_i \mid x_i)$). The conformal threshold at level ε is

$$\hat{q} = \mathrm{Quantile}\Big(\{s_i\}_{i=1}^n,\ \tfrac{\lceil (n+1)(1-\varepsilon)\rceil}{n}\Big),$$

and the prediction set is $C(x) = \{y : s(x, y) \le \hat{q}\}$. Under exchangeability of calibration and test data, split CP guarantees $\mathbb{P}[y_{n+1} \in C(x_{n+1})] \ge 1 - \varepsilon$ [33].

Mondrian conformal prediction (MCP). MCP computes a *per-class* threshold $\hat{q}_y$ on examples with true label y, yielding class-conditional coverage $\mathbb{P}[y_{n+1} \in C(x_{n+1}) \mid y_{n+1} = y] \ge 1 - \varepsilon$ for each y [35]. This matters here because label priors and error costs are asymmetric.

Leakage-safe split. We partition each domain into four disjoint subsets: *train* (probing, classifier training), *val* (ensemble weight α), *calib* (conformal threshold estimation), *test* (final evaluation). Thresholds are computed only on *calib*.

4 Method: PLLS and PLLS-CP

4.1 Overview

The pipeline has five stages: (1) layer probing on pseudo-labels, (2) layer selection by a combined score, (3) an attention-based classifier on the selected layer(s), (4) ensembling with IRIS, (5) Mondrian conformal prediction. Stages (1)–(5) form **PLLS**; its cross-domain variant with recalibration (Section 4.6) is **PLLS-CP**. Figure 1 illustrates the full flow.

4.2 PLLS: Pseudo-Label Layer Selection

Let $\mathcal{L}$ be a candidate set of layers (we use 8 approximately equally spaced layers in practice; the exact choice is described in Section 6). For each $\ell \in \mathcal{L}$ we train a lightweight logistic-regression probe on $\{h^{(\ell)}(x_i), \tilde{y}_i\}_{i \in \mathrm{train}}$, where $\tilde{y}_i$ is the IRIS pseudo-label. Two statistics are recorded on the validation split: the probe's ROC-AUC (auc_ℓ) and its expected calibration error (ECE_ℓ). *Both statistics are computed against the IRIS pseudo-labels $\tilde{y}$, never against ground-truth labels; ground-truth labels are not consulted at any point during selection, classifier training, or ensembling, preserving the unsupervised status of the pipeline.* The combined score is

$$\mathrm{score}_\ell \ = \ \lambda \cdot \mathrm{auc}_\ell \ + \ (1 - \lambda) \cdot (1 - \mathrm{ECE}_\ell), \tag{1}$$

with $\lambda \in [0, 1]$ controlling the trade-off between discrimination and calibration. Higher λ emphasizes the probe's separability; lower λ penalizes miscalibration so that the downstream conformal layer receives well-behaved scores. We use $\lambda = 0.7$ throughout, selected via a coarse search over the grid $\lambda \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ on the train/validation splits of three randomly sampled domains. Crucially, the calibration and test splits of all 18 evaluation domains were strictly held out during this hyperparameter selection and are used only for the final evaluation reported in Section 6. An ablation against $\lambda = 1$ (AUC-only) is reported in Section 6.5 and confirms that the calibration term is needed to avoid regressions on harder domains. The top-K layers by score_ℓ are then selected. Ablations (Section 6.5) identify $K = 1$ as the stability-optimal choice.

4.3 Attention-based classifier

The classifier consumes hidden states from the K selected layers and outputs $\hat{p}_{\mathrm{PLLS}}(y \mid x)$. When $K > 1$, embeddings are projected into a shared low-dimensional space and aggregated via self-

attention; for $K = 1$ the classifier reduces to an MLP applied to the selected layer. Training uses the same IRIS pseudo-labels, preserving the unsupervised setting.

4.4 Ensemble with IRIS

Final in-domain scores combine PLLS with IRIS:

$$\hat{p}_{\text{ens}}(y \mid x; \alpha) = \alpha \cdot \hat{p}_{\text{PLLS}}(y \mid x) + (1 - \alpha) \cdot \hat{p}_{\text{IRIS}}(y \mid x), \qquad \alpha \in \{0, 0.1, \ldots, 1.0\}. \tag{2}$$

The optimal $\hat{\alpha}$ is chosen on *val* by minimizing empirical loss. Because $0 \in \{0, 0.1, \ldots, 1.0\}$, the ensemble can always fall back to IRIS; Theorem 3 formalizes this non-degradation guarantee.

4.5 Mondrian conformal layer

We apply MCP at level $\varepsilon = 0.1$ on $\hat{p}_{\text{ens}}$, with nonconformity scores $s(x, y) = 1 - \hat{p}_{\text{ens}}(y \mid x)$. The prediction set $C(x)$ may contain one class (confident prediction), both classes (abstention), or be empty. Abstention is a feature, not a bug: it lets the detector admit ignorance on hard examples.

4.6 PLLS-CP: cross-domain recalibration

When deploying to a new domain $\mathcal{D}_{\text{target}}$, PLLS-CP *does not* retrain the detector. It keeps $(\hat{f}_{\text{PLLS}}, \hat{\alpha})$ frozen and recomputes only the conformal threshold $\hat{q}$ on a small target calibration set (typically $n \in [100, 200]$, pseudo-labeled). Theorem 1 shows that this suffices to guarantee coverage on the target distribution; Theorem 2 quantifies the coverage-gap concentration. Note that because calibration relies on pseudo-labels, the theoretical guarantee formally bounds coverage *with respect to the pseudo-label distribution*, not the true hallucination labels; a closed-form bound transferring pseudo-label coverage to true-label coverage is left as future work. Empirically, pseudo-label calibration recovers almost all of the true-label coverage attained under true-label calibration—within one to two percentage points on informative domains, with no shortfall on HELM (Appendix F); we characterize this dependence, together with its breakdown under systematic pseudo-label inversion, in Appendices F and G.

5 Theoretical Analysis

We provide three formal results. Theorems 1 and 2 justify cross-domain recalibration; Theorem 3 justifies ensembling. Full proofs are provided in Appendix A.

Theorem 1 (Cross-domain coverage under recalibration)**.** *Let $s\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be any nonconformity score whose parameters are fixed before observing target data (e.g. trained on a source domain). Let $Z_{\text{cal}} = \{z_1, \ldots, z_n\}$ with $z_i = (x_i, y_i)$ and a test point z_{n+1} be* exchangeable *draws from the target distribution P_{target}. For $\varepsilon \in (0, 1)$, define*

$$\hat{q} = \inf\Big\{ q \in \mathbb{R} : \tfrac{1}{n} \sum_{i=1}^{n} \mathbf{1}[s(x_i, y_i) \le q] \ge \tfrac{\lceil (n+1)(1-\varepsilon) \rceil}{n} \Big\},$$

and $C(x_{n+1}) = \{y : s(x_{n+1}, y) \le \hat{q}\}$. Then $\mathbb{P}[y_{n+1} \in C(x_{n+1})] \ge 1 - \varepsilon$.

Proof sketch. Since s is fixed prior to target data, the scalar scores $S_i = s(x_i, y_i)$ inherit exchangeability from the data. The rank of S_{n+1} among $S_1, \ldots, S_{n+1}$ is therefore uniform on $\{1, \ldots, n+1\}$, so $\mathbb{P}[S_{n+1} \le \hat{q}] \ge \lceil (n+1)(1-\varepsilon) \rceil / (n+1) \ge 1 - \varepsilon$. Full proof in Appendix A.

The practical import: *any* detector (including our ensemble $\hat{f}_{\text{PLLS}}$) trained on a source domain can be deployed on a target domain with a marginal coverage guarantee, provided exchangeability holds *within* the target calibration+test draw—which is automatic for a random split of target data.

Theorem 2 (Coverage gap concentration)**.** *Under the setting of Theorem 1, with $k = \lceil (n+1)(1-\varepsilon) \rceil$, the conditional coverage $P_{\text{cond}} = \mathbb{P}[Y \in C(X) \mid \mathcal{D}_{\text{calib}}]$ satisfies $P_{\text{cond}} \sim \text{Beta}(k, n - k + 1)$ [34, 36]. Consequently, for any $\delta > 1/(n+1)$,*

$$\mathbb{P}\big[\, |P_{\text{cond}} - (1 - \varepsilon)| \ge \delta \,\big] \;\le\; 2 \exp\!\left(-2n \Big(\delta - \tfrac{1}{n+1} \Big)^2 \right).$$

Proof sketch. By the Beta–Binomial duality, for integer k the tail $\mathbb{P}[P_{\text{cond}} \geq p]$ equals the Binomial tail $\mathbb{P}[\text{Bin}(n, p) \leq k - 1]$, to which Hoeffding's inequality applies legitimately (a sum of n independent Bernoulli trials). Centering the resulting bound on $1 - \varepsilon$ via $|\mathbb{E}[P_{\text{cond}}] - (1 - \varepsilon)| \leq \frac{1}{n+1}$ yields the stated inequality, with the same constant. Full proof in Appendix A.

At $n = 200$, $\delta = 0.1$ gives $1/(n + 1) \approx 0.005$, and Theorem 2 predicts that large deviations are highly improbable (e.g., $\leq 2e^{-3.61} \approx 0.054$ probability that coverage deviates by more than 10% from the target). Empirically, we observe a tight concentration around the target coverage with a standard deviation of only 0.035 and a maximum gap of 0.071 across the 306 off-diagonal transfer pairs (Section 6.4), strictly satisfying the theoretical bounds.

Theorem 3 (Ensemble non-degradation)**.** *Let $L(p, y) \in [0, M]$ be a bounded loss. Let $p_{\text{ens}}(\alpha) = \alpha p_{\text{PLLS}} + (1 - \alpha) p_{\text{IRIS}}$ with α selected from a finite grid $\mathcal{A} \subset [0, 1]$ of size $|\mathcal{A}|$ containing 0, by $\hat{\alpha} = \arg\min_{\alpha \in \mathcal{A}} \hat{L}_{\text{val}}(\alpha)$ on a validation set of size N_{val}. Then with probability at least $1 - \gamma$:*

$$L_{\text{ens}}(\hat{\alpha}) \;\leq\; L_{\text{IRIS}} + 2M\sqrt{\tfrac{\ln(2|\mathcal{A}|/\gamma)}{2N_{\text{val}}}}.$$

Proof sketch. $0 \in \mathcal{A}$ implies $\hat{L}_{\text{val}}(\hat{\alpha}) \leq \hat{L}_{\text{val}}(0)$. A uniform Hoeffding+union-bound argument over $\mathcal{A}$ controls both $|L_{\text{ens}}(\hat{\alpha}) - \hat{L}_{\text{val}}(\hat{\alpha})|$ and $|\hat{L}_{\text{val}}(0) - L_{\text{IRIS}}|$ by $M\sqrt{\ln(2|\mathcal{A}|/\gamma)/(2N_{\text{val}})}$; summing gives the bound. Full proof in Appendix A.

With $\mathcal{A} = \{0, 0.1, \ldots, 1.0\}$ (so $|\mathcal{A}| = 11$) and $M = 1$ (clipped cross-entropy or Brier score), the bound shrinks as $O(1/\sqrt{N_{\text{val}}})$. This is the formal reason our ensemble can only fail to improve on IRIS by an amount that vanishes with validation size.

6 Experiments

6.1 Setup

Base model and benchmarks. Our primary experiments use LLaMA-3.1-8B-Instruct [3], which has 32 transformer layers; the candidate layer set is fixed to $\mathcal{L} = \{0, 4, 8, 12, 16, 20, 24, 28\}$, i.e. 8 approximately equally spaced layers. We additionally replicate the full pipeline on a second model family (Qwen3-8B [4], with Qwen2.5-32B-Instruct [5] as a larger reader; Appendix D). We evaluate on three benchmarks covering 18 domains: **TrueFalse** [12] (7 topical domains: *animals, cities, companies, elements, facts, generated, inventions*), **HaluEval 2.0** [14]—the domain-structured extension of HaluEval [13], hereafter abbreviated *HaluEval* in tables and figures—(5 domains: *Bio-Medical, Education, Finance, Open-Domain, Science*), and **HELM** (Hallucination detection Evaluation for multiple LLMs) [32] (6 model-specific domains: *falcon40b, gptj7b, llamabase7b, llamachat7b, llamachat13b, opt7b*). CoT embeddings are extracted following the IRIS protocol.

Protocol. Each domain is split train/val/calib/test. Pseudo-labels for probing and classifier training are generated by IRIS's verbalized-uncertainty procedure (Section 3), and *all training-time signals—including layer-selection scores, ensemble weight selection on val, and conformal threshold calibration on calib—rely exclusively on these pseudo-labels.* Ground-truth labels are used only to compute the final test-time metrics reported below, which is consistent with the standard unsupervised-learning evaluation protocol. Conformal target coverage is $1 - \varepsilon = 0.9$. For cross-domain experiments, the detector is trained and validated on the source domain, then calibrated and tested on the target.

In-domain and cross-domain results are produced by two separate evaluation protocols and are not intended to be read across tables. Table 1 evaluates each domain's own held-out test fold under a four-way $60/10/10/20$ train/val/calib/test split. Table 2 evaluates cross-domain transfer under the deployment protocol of Section 4.6: for every target domain, 100 stratified points form the calibration set and the remainder is the test fold, so its diagonal is the in-domain cell *of that protocol* and is the denominator of the retention ratios reported there. Diagonal and off-diagonal entries within Table 2 always come from the same detectors, the same seed, and the same folds.

Metrics. We report metrics standard in the conformal-prediction [34] and selective-classification [38, 39] literatures. *Point-level accuracy* measures how often the argmax prediction

Table 1: In-domain accuracy aggregated over 18 evaluation domains. PLLS improves on the IRIS baseline while maintaining stability, with non-worse behaviour on $16/18$ domains. PLLS acc. is the accuracy of the PLLS–IRIS ensemble (column Ens of Table 6).

Domain group	IRIS acc.	**PLLS** acc. (ours)	# domains $\geq$ IRIS	# domains $>$ IRIS	Mean Δ acc.
All 18 domains	0.739	**0.768**	16	10	+0.029
Informative (12)	0.784	**0.815**	11	6	+0.031
HELM (6)	0.648	**0.673**	5	4	+0.025

Table 2: Cross-domain transfer. Retention is off-diagonal / diagonal; higher is better. IRIS and PLLS-CP are evaluated with the *same* IRIS instance in both columns (single seed, single run). PLLS-CP gains $+10.5$ pp in accuracy-retention and $+1.7$ pp in selective-accuracy-retention over IRIS.

	Accuracy		Selective accuracy	
	IRIS	**PLLS-CP**	IRIS	**PLLS-CP**
Diagonal mean	0.742	**0.768**	0.723	**0.746**
Off-diagonal mean	0.606	**0.708**	0.704	**0.739**
Retention	0.817	**0.922**	0.974	**0.991**

matches the ground-truth label, independent of abstention. *Coverage* is the fraction of test examples whose true label lies in $C(x)$, nominally $1 - \varepsilon = 0.9$; the *coverage gap* is its signed deviation from 0.9; *abstention rate* is the fraction of test examples for which $C(x)$ contains both classes; *selective accuracy* is the accuracy on examples where the detector commits. Point accuracy alone hides the safety/utility trade-off of abstention, while coverage alone is trivially achieved by abstaining everywhere, so we report all four.

6.2 In-domain results

Table 1 aggregates in-domain results across 18 domains. PLLS is non-worse than IRIS on 16 of 18 domains, strictly improves on 10, and shows two small regressions (-0.004 on *true_false:companies*, where IRIS is already at 0.921, and -0.009 on *helm:opt7b*, where IRIS is at 0.684). This results in a mean improvement of $+0.029$ accuracy across domains. The gain is concentrated on weaker domains: $+0.157$ on Bio-Medical (from 0.404 to 0.560), $+0.120$ on Science (from 0.505 to 0.625), and $+0.086$ on falcon40b (from 0.581 to 0.667). Full per-domain tables are in Appendix B.

6.3 Cross-domain transfer

We evaluate cross-domain transfer across all $18 \times 18 = 324$ source$\rightarrow$target pairs. Table 2 reports diagonal (in-domain), off-diagonal (transferred), and retention (= off-diag/diag) for accuracy and selective accuracy.

Accuracy-retention rises from 0.817 to 0.922 ($+10.5$ pp).[1] Crucially, this improvement is driven by an absolute gain in off-diagonal accuracy ($+0.102$, from 0.606 to 0.708), not by changes in the diagonal denominator: the diagonal mean changes by only $+0.026$. A block-level breakdown (Table 3) reveals that the largest gains are on *cross-family* transfers, particularly HaluEval$\rightarrow$TrueFalse ($+0.240$) and HELM$\rightarrow$TrueFalse ($+0.214$). Figure 2 visualises this per-pair: blue cells (PLLS-CP wins) dominate all off-diagonal blocks, while red cells concentrate on HELM- and HaluEval-as-target transfers where both methods are near chance. Out of the 306 off-diagonal pairs, PLLS-CP strictly improves over IRIS on 225 (74%); the remaining 81 are ties (10) or regressions (71), the latter concentrated on HaluEval and HELM targets (29% of cells each, vs. 13% for TrueFalse targets), with 23 regressions exceeding -0.05.

[1] The adapt-mode split used for Table 2 evaluates each diagonal cell on the full target domain, including that domain's own training fold, so the retention denominator is mildly optimistic. A leakage-free re-run that evaluates every cell only on held-out 20% folds (from which all detectors are excluded) gives $0.832 \rightarrow 0.931$ accuracy retention and $0.921 \rightarrow 0.984$ selective-accuracy retention—a $+9.9$ pp gain, close to the $+10.5$ pp of Table 2—so the conclusions are unaffected.

Table 3: Accuracy by source→target family block. Largest gains on cross-family transfers.

Block	IRIS	**PLLS-CP**	Δ
TrueFalse → TrueFalse	0.870	0.883	+0.013
TrueFalse → HaluEval	0.577	0.600	+0.024
TrueFalse → HELM	0.597	0.606	+0.009
HaluEval → TrueFalse	0.571	0.812	**+0.240**
HaluEval → HaluEval	0.496	0.623	+0.127
HaluEval → HELM	0.511	0.605	+0.094
HELM → TrueFalse	0.659	0.873	**+0.214**
HELM → HaluEval	0.535	0.648	+0.114
HELM → HELM	0.543	0.626	+0.084

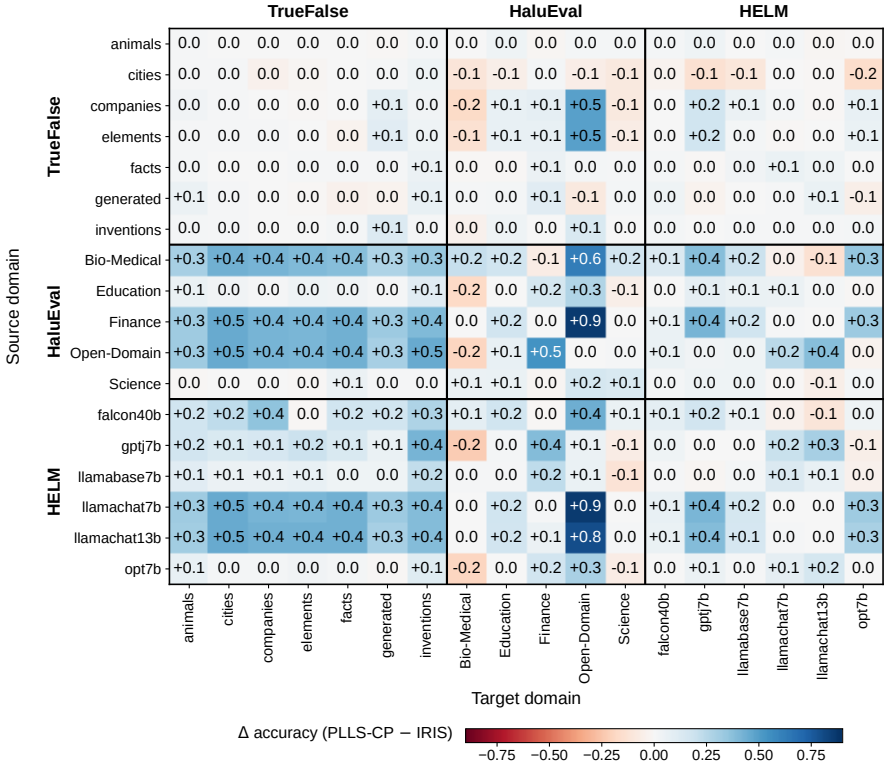

Figure 2: Accuracy delta $\Delta_{ij} = \mathrm{Acc}_{ij}^{\mathrm{PLLS\text{-}CP}} - \mathrm{Acc}_{ij}^{\mathrm{IRIS}}$. Blue favors PLLS-CP, red favors IRIS. Major gains concentrate in off-diagonal transfers (e.g., HaluEval→TrueFalse $+0.240$, HELM→TrueFalse $+0.214$). Regressions cluster on HaluEval and HELM targets where both methods operate near chance.

6.4 Reliability under conformal prediction

We next examine whether gains survive when the detector is asked to either answer correctly or abstain. Table 4 compares IRIS + MCP against PLLS-CP on the 18 in-domain evaluations. Both methods attain empirical coverage close to the nominal 0.90 target and are comparable on this metric—IRIS + MCP's 0.933 is in fact marginally closer to 0.90 than PLLS-CP's 0.938, and their mean absolute coverage gaps are similar (0.041 vs. 0.042). Coverage is not where PLLS-CP's advantage lies: it reduces abstention by 0.006 and raises selective accuracy by 0.013, with the larger reliability benefit appearing in cross-domain selective-accuracy retention (Table 2). Per-domain, the largest selective-accuracy gains are again on HaluEval (Science $+0.221$, Bio-Medical $+0.076$); HELM domains remain hard.

Empirical verification of Theorem 2. Across the 306 off-diagonal cross-domain transfers, IRIS-based CP has coverage-gap mean $+0.007$ and standard deviation 0.035; 75% of pairs have gap $\leq +0.032$. The transfer protocol calibrates on $n = 100$ target points, at which the Theorem 2 bound

Table 4: Reliability metrics averaged across 18 in-domain evaluations. Coverage is near-nominal and comparable across methods; PLLS-CP attains lower abstention and higher selective accuracy. We also report the mean absolute coverage gap (MAD gap) across domains. Selective accuracy of IRIS + MCP is averaged over 17 domains, since on helm:falcon40b it abstains on every test example.

	coverage	$MAD(gap)$	abstention	sel. acc.
IRIS + MCP	0.933	0.041	0.533	0.800
PLLS-CP	0.938	0.042	**0.527**	**0.813**

Table 5: Ablations. "Combined score" uses Eq. 1; "AUC only" drops calibration term. "Multi-layer agg." aggregates all 8 candidate layers without selection. "$K = 1$" selects the single best layer.

Variant	non-worse / improved / regressed	mean Δ acc
Multi-layer aggregation (no selection, all 8 layers)	13/8/5	+0.031
PLLS, $K = 4$, combined score	13/8/5	+0.026
PLLS, $K = 1$, AUC-only score	13/9/5	+0.030
PLLS, $K = 1$, combined score (ours)	**16/10/2**	+0.029

is loose (probability ≤ 0.394 that coverage deviates by more than 10% from target); nonetheless the empirical maximum gap across all 306 pairs is only $+0.071$ (largest absolute deviation 0.094), i.e. *no* pair deviates by more than 10%, and the concentration tightens rapidly with n (at $n = 200$ the bound is 0.054). A finite-sample study directly matching the Beta–Binomial prediction as n varies is reported in Appendix H. Cross-domain recalibration is thus practically reliable, not merely worst-case.

6.5 Ablations

We isolate the contribution of each pipeline component. Table 5 summarizes the four most informative ablations; full results are in Appendix C.

Three takeaways: (i) naive multi-layer aggregation yields the highest *mean* Δ but fails on 5 domains, making it unreliable; (ii) larger K increases gains on some domains but brings noise-induced regressions; (iii) the calibration term in the selection score (Eq. 1) is what keeps the profile stable—dropping it leaves the mean gain unchanged ($+0.030$ vs. $+0.029$) but raises regressions from 2 to 5. The $K = 1$ combined-score configuration delivers the most stable non-worse profile ($16/18$). A component decomposition (IRIS alone, PLLS alone, and the ensemble) in and across domains is reported in Appendix D.

6.6 Scope of applicability: HELM as a weak-pseudo-label regime

On HELM domains, IRIS itself—previously SOTA among knowledge-free white-box methods—is weaker than on the other two benchmarks: mean accuracy 0.648 (Table 1), with per-domain AUCs on the conformal test folds ranging from 0.41 to 0.72. PLLS still improves on it: it is non-worse on $5/6$ HELM domains and strictly better on $4/6$, with a mean gain of $+0.025$ (vs. $+0.031$ on informative domains), and the validation-selected ensemble uses PLLS ($\hat{\alpha} > 0$) on 5 of 6 HELM domains.

Why the HELM bottleneck is the pseudo-label, not the representation. HELM differs from TrueFalse and HaluEval in how its statements are produced. HaluEval and TrueFalse contain statements judged by the *same* LLM whose internal states we probe (LLaMA-3.1-8B-Instruct), whereas HELM contains outputs generated by *other* models (falcon40b, gptj7b, opt7b, llamabase7b, llamachat7b, llamachat13b). It is tempting to attribute the HELM drop to an *absence* of hidden-state signal for third-party text; a closer analysis shows otherwise. To separate the two bottlenecks—representation and pseudo-label—we compare, per benchmark, the agreement of the verbalized-confidence pseudo-label with ground truth against a *diagnostic ceiling*: a probe from the same family PLLS uses (a single-layer linear probe) but trained on ground-truth labels, used only to characterize the available signal and never as part of the unsupervised method. Their difference, $\text{headroom} = \text{ceiling} - \text{AUC}(\text{verb} \rightarrow \text{GT})$, measures how much a better pseudo-label could still recover with the *same* read-out (Appendix E, Table 17). The three benchmarks separate cleanly:

- **TrueFalse** (headroom $\approx +0.07$, both terms high): the regime the method is designed for.
- **HaluEval** (headroom $\approx +0.04$): the pseudo-label is already close to what this read-out can extract, so improving the pseudo-label would recover little—the remaining limitation lies in the representation/read-out rather than in the label. This accounts for the plateau in HaluEval selective accuracies (0.64–0.89).
- **HELM** (headroom $\approx +0.19$, with a *weak* pseudo-label): the signal is present at ceiling AUC ≈ 0.79, but the verbalized-confidence pseudo-label is only weakly informative (AUC 0.56–0.63 across all six domains), so the binding constraint is the *pseudo-label*.

A larger reader (Qwen2.5-32B) produces a better pseudo-label on HELM (verb$\rightarrow$GT AUC 0.67) and the detector improves accordingly (AUC ≈ 0.67 vs. ≈ 0.60 for LLaMA-3.1-8B; Appendix D), consistent with pseudo-label quality being the binding constraint.

We therefore interpret HELM as a regime where the self-generated pseudo-label is weak. Unsupervised CoT-embedding detectors are most reliable on hallucinations of the same model whose hidden states they read; on third-party outputs the verbalized-confidence pseudo-label is only weakly informative, and PLLS-CP still improves on IRIS there, but with a smaller reliability margin; approaching the ceiling would need a stronger labeling signal. The conformal layer reflects this honestly: abstention rates on HELM reach 0.71–0.97, which is itself a useful deployment signal.

7 Discussion

PLLS ports MHAD's selection idea [23] to the unsupervised regime (layers instead of neurons, pseudo-labels instead of human annotation), making selection available in the IRIS-compatible setting. Operationally, Theorem 1 decouples the detector from the calibrator: updating $\hat{q}$ on 100–200 pseudo-labeled target examples restores coverage guarantees with no fine-tuning, which is cheap enough to perform on arrival at any new domain.

8 Conclusion

We addressed deploying hallucination detectors under distribution shift, where existing knowledge-free white-box methods lose nearly a fifth of their accuracy. PLLS-CP treats cross-domain deployment as recalibration rather than retraining: only the conformal threshold is updated on a small target calibration set. PLLS provides a distribution-shift-friendly base detector via pseudo-label layer selection, with a non-degradation guarantee against IRIS. Three theorems give marginal coverage under cross-domain recalibration, exponential concentration of the coverage gap, and ensemble non-degradation. On 18 domains and 324 transfer pairs, PLLS-CP raises accuracy retention from 0.817 to 0.922, with coverage standard deviation (0.035) consistent with the theoretical bound.

Limitations and future work. We evaluate on two model families (LLaMA-3.1-8B-Instruct and Qwen3-8B, with Qwen2.5-32B-Instruct as a larger reader; Appendix D); broader generalization to further families remains to be verified, though the framework is model-agnostic. The reliability guarantee is formal only with respect to the pseudo-label distribution: true-label coverage tracks pseudo-label quality, holding while the pseudo-label is not systematically inverted—a boundary we characterize empirically (Appendix G) but do not certify in closed form. HELM domains, where the self-generated pseudo-label is weak (verb$\rightarrow$GT AUC ≈ 0.60), show that gains are bounded by pseudo-label quality. We leave a closed-form pseudo-to-true-label transfer bound, and joint learning of layer selection and conformal score, as future work. All results are single-run.

Acknowledgments and Disclosure of Funding

Funding. This research was supported by the Russian Science Foundation grant No. 25-11-90002, `https://rscf.ru/project/25-11-90002/`.

Competing interests. The authors declare no competing interests.

References

[1] Achiam, J., Adler, S., Agarwal, S., et al. (2023). GPT-4 technical report. *arXiv:2303.08774*.

[2] Touvron, H., Martin, L., Stone, K., et al. (2023). LLaMA 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*.

[3] Dubey, A., Jauhri, A., Pandey, A., et al. (2024). The LLaMA 3 herd of models. *arXiv:2407.21783*.

[4] Yang, A., Li, A., Yang, B., et al. (2025). Qwen3 technical report. *arXiv:2505.09388*.

[5] Yang, A., Yang, B., Zhang, B., et al. (2024). Qwen2.5 technical report. *arXiv:2412.15115*.

[6] Ji, Z., Lee, N., Frieske, R., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys* 55(12):1–38.

[7] Zhang, Y., Li, Y., Cui, L., et al. (2023). Siren's song in the AI ocean: A survey on hallucination in LLMs. *arXiv:2309.01219*.

[8] Huang, L., Yu, W., Ma, W., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43(2), Article 42. *arXiv:2311.05232*.

[9] MITRE Corporation (2025). MITRE ATLAS: Adversarial Threat Landscape for AI Systems. `https://atlas.mitre.org/`.

[10] Spracklen, J., Wijewickrama, R., Sakib, A.H.M.N., Maiti, A., Viswanath, B., Jadliwala, M. (2025). We have a package for you! A comprehensive analysis of package hallucinations by code-generating LLMs. *34th USENIX Security Symposium*, 3687–3706. *arXiv:2406.10279*.

[11] Srey, P., Wu, X., Luu, A.T. (2025). Unsupervised hallucination detection by inspecting reasoning processes. *EMNLP*, 22117–22129. *arXiv:2509.10004*.

[12] Azaria, A., Mitchell, T. (2023). The internal state of an LLM knows when it's lying. *Findings of EMNLP*, 967–976. *arXiv:2304.13734*.

[13] Li, J., Cheng, X., Zhao, W.X., Nie, J.-Y., Wen, J.-R. (2023). HaluEval: A large-scale hallucination evaluation benchmark for large language models. *EMNLP*, 6449–6464.

[14] Li, J., Chen, J., Ren, R., Cheng, X., Zhao, W.X., Nie, J.-Y., Wen, J.-R. (2024). The dawn after the dark: An empirical study on factuality hallucination in large language models (HaluEval 2.0). *ACL*, 10879–10899. *arXiv:2401.03205*.

[15] Kuhn, L., Gal, Y., Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *ICLR*. *arXiv:2302.09664*.

[16] Fadeeva, E., Rubashevskii, A., Shelmanov, A., et al. (2024). Fact-checking the output of large language models via token-level uncertainty quantification. *Findings of ACL*, 9367–9385.

[17] Ma, H., Pan, J., Liu, J., et al. (2025). Semantic Energy: Detecting LLM hallucination beyond entropy. *arXiv:2508.14496*.

[18] Kossen, J., Han, J., Razzak, M., Schut, L., Malik, S., Gal, Y. (2024). Semantic entropy probes: Robust and cheap hallucination detection in LLMs. *arXiv:2406.15927*.

[19] Tian, K., Mitchell, E., Zhou, A., et al. (2023). Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *EMNLP*, 5433–5442. *arXiv:2305.14975*.

[20] Chen, C., Liu, K., Chen, Z., Gu, Y., Wu, Y., Tao, M., Fu, Z., Ye, J. (2024). INSIDE: LLMs' internal states retain the power of hallucination detection. *ICLR*. *arXiv:2402.03744*.

[21] Du, X., Xiao, C., Li, Y. (2024). HaloScope: Harnessing unlabeled LLM generations for hallucination detection. *NeurIPS*. *arXiv:2409.17504*.

[22] Sriramanan, G., Bharti, S., Sadasivan, V.S., Saha, S., Kattakinda, P., Feizi, S. (2024). LLM-Check: Investigating detection of hallucinations in large language models. *NeurIPS* 37:34188–34216.

[23] Zhang, L., Song, D., Wu, Z., et al. (2025). Detecting hallucination in large language models through deep internal representation analysis. *IJCAI*, 8357–8365.

[24] Li, Q., Geng, J., Chen, Z., et al. (2025). HD-NDEs: Neural differential equations for hallucination detection in LLMs. *ACL*, 6173–6186. *arXiv:2506.00088*.

[25] Li, J., Tu, G., Cheng, S., et al. (2025). LLM hallucination detection: A fast Fourier transform method based on hidden layer temporal signals. *arXiv:2509.13154*.

[26] Bazarova, A., Volodichev, A., Yugay, A., et al. (2026). Hallucination detection in LLMs with topological divergence on attention graphs. *ACL*, 15449–15470. *arXiv:2504.10063*.

[27] Liu, Q., Chen, X., Ding, Y., Song, B., Wang, W., Wu, S., Wang, L. (2025). Attention-guided self-reflection for zero-shot hallucination detection in large language models. *EMNLP*, 21005–21021. *arXiv:2501.09997*.

[28] Manakul, P., Liusie, A., Gales, M.J.F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *EMNLP*, 9004–9017. *arXiv:2303.08896*.

[29] Zhang, J., Li, Z., Das, K., Malin, B., Kumar, S. (2023). SAC3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. *Findings of EMNLP*, 15445–15458. *arXiv:2311.01740*.

[30] Chen, J., Mueller, J. (2024). Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. *ACL*, 5186–5200. *arXiv:2308.16175*.

[31] Gal, Y., Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *ICML*, 1050–1059.

[32] Su, W., Wang, C., Ai, Q., et al. (2024). Unsupervised real-time hallucination detection based on the internal states of large language models. *Findings of ACL*, 14379–14391. *arXiv:2403.06448*.

[33] Vovk, V., Gammerman, A., Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.

[34] Angelopoulos, A.N., Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv:2107.07511*.

[35] Vovk, V., Lindsay, D., Nouretdinov, I., Gammerman, A. (2003). Mondrian confidence machine. Tech. Rep., Royal Holloway, Univ. of London.

[36] Vovk, V. (2012). Conditional validity of inductive conformal predictors. *Asian Conference on Machine Learning (ACML)*, PMLR 25:475–490.

[37] Tibshirani, R.J., Barber, R.F., Candès, E.J., Ramdas, A. (2019). Conformal prediction under covariate shift. *NeurIPS*, 2530–2540.

[38] El-Yaniv, R., Wiener, Y. (2010). On the foundations of noise-free selective classification. *JMLR* 11:1605–1641.

[39] Geifman, Y., El-Yaniv, R. (2017). Selective classification for deep neural networks. *NeurIPS*.

A Full Proofs

A.1 Proof of Theorem 1

Proof. **Step 1 (score determinism).** By assumption, $s(\cdot,\cdot)$ is fixed before any target data is observed; it is therefore deterministic given $Z_{\mathrm{cal}} \cup \{z_{n+1}\}$.

Step 2 (score exchangeability). Define $S_i = s(x_i, y_i)$ for $i \in \{1, \ldots, n+1\}$. The map $(z_1, \ldots, z_{n+1}) \mapsto (S_1, \ldots, S_{n+1})$ is deterministic and symmetric (applied coordinatewise). Since the underlying $(z_1, \ldots, z_{n+1})$ are exchangeable, so are the scores.

Step 3 (uniform rank). For exchangeable $S_1, \ldots, S_{n+1}$, the rank of S_{n+1} (with ties broken uniformly at random) is uniform on $\{1, \ldots, n+1\}$.

Step 4 (threshold as empirical quantile). By construction $\hat{q}$ is the $\lceil (n+1)(1-\varepsilon) \rceil / n$-quantile of the empirical distribution of $S_1, \ldots, S_n$. The event $\{S_{n+1} \le \hat{q}\}$ is equivalent to S_{n+1} being among the $\lceil (n+1)(1-\varepsilon) \rceil$ smallest of all $n+1$ scores.

Step 5 (coverage). By uniform-rank,

$$\mathbb{P}[S_{n+1} \le \hat{q}] \;\ge\; \frac{\lceil (n+1)(1-\varepsilon) \rceil}{n+1} \;\ge\; \frac{(n+1)(1-\varepsilon)}{n+1} \;=\; 1-\varepsilon.$$

Step 6. By definition $y_{n+1} \in C(x_{n+1}) \iff s(x_{n+1}, y_{n+1}) \le \hat{q} \iff S_{n+1} \le \hat{q}$. Substituting yields $\mathbb{P}[y_{n+1} \in C(x_{n+1})] \ge 1 - \varepsilon$. □

A.2 Proof of Theorem 2

Proof. Write $c = 1 - \varepsilon$, $k = \lceil (n+1)(1-\varepsilon) \rceil$, and $\mu = \mathbb{E}[P_{\text{cond}}] = k/(n+1)$.

Step 1 (Beta law of conditional coverage). In split CP with exchangeable calibration and test data, the conditional coverage $P_{\text{cond}} = \mathbb{P}[Y \in C(X) \mid \mathcal{D}_{\text{calib}}]$ equals $F_{n+1}(\hat{q})$, where F_{n+1} is the target CDF of scores. Classical CP results give $P_{\text{cond}} \sim \text{Beta}(k,\, n+1-k)$.

Step 2 (mean location). Since $x \le \lceil x \rceil \le x + 1$,

$$(n+1)(1-\varepsilon) \le k \le (n+1)(1-\varepsilon) + 1 \implies c \le \mu \le c + \tfrac{1}{n+1},$$

and $n\mu = k - \mu$.

Step 3 (Beta–Binomial duality). For *integer* k and all $p \in [0,1]$,

$$\mathbb{P}[P_{\text{cond}} \le p] = I_p(k,\, n+1-k) = \sum_{j=k}^{n} \binom{n}{j} p^j (1-p)^{n-j} = \mathbb{P}[\text{Bin}(n,p) \ge k],$$

where I_p is the regularized incomplete Beta function; hence $\mathbb{P}[P_{\text{cond}} \ge p] = \mathbb{P}[\text{Bin}(n,p) \le k-1]$. (For $\varepsilon \ge 1/(n+1)$ we have $1 \le k \le n$, so the duality applies; this holds for all our settings.)

Step 4 (Binomial Hoeffding — the legitimate application). $\text{Bin}(n,p)$ is a sum of n i.i.d. $\text{Bernoulli}(p)$ variables, so for $s > 0$,

$$\mathbb{P}[\text{Bin}(n,p) \le np - s] \le e^{-2s^2/n}, \qquad \mathbb{P}[\text{Bin}(n,p) \ge np + s] \le e^{-2s^2/n}.$$

Step 5 (upper tail). For $t > 0$ put $p = \mu + t$; then $np = k - \mu + nt$, and

$$\mathbb{P}[P_{\text{cond}} \ge \mu + t] = \mathbb{P}[\text{Bin}(n,p) \le k-1], \quad s = np - (k-1) = nt + (1-\mu) \ge nt,$$

so $\mathbb{P}[P_{\text{cond}} \ge \mu + t] \le e^{-2s^2/n} \le e^{-2nt^2}$.

Step 6 (lower tail). Put $p = \mu - t$; then $np = k - \mu - nt$, and

$$\mathbb{P}[P_{\text{cond}} \le \mu - t] = \mathbb{P}[\text{Bin}(n,p) \ge k], \quad s = k - np = \mu + nt \ge nt,$$

so $\mathbb{P}[P_{\text{cond}} \le \mu - t] \le e^{-2nt^2}$. Summing Steps 5–6, $\mathbb{P}[|P_{\text{cond}} - \mu| \ge t] \le 2e^{-2nt^2}$.

Step 7 (center on $1-\varepsilon$). Since $|\mu - c| \le \frac{1}{n+1}$, for $\delta > \frac{1}{n+1}$ set $t = \delta - \frac{1}{n+1} > 0$; then $\{|P_{\text{cond}} - c| \ge \delta\} \subseteq \{|P_{\text{cond}} - \mu| \ge t\}$, giving

$$\mathbb{P}\big[|P_{\text{cond}} - (1-\varepsilon)| \ge \delta\big] \le 2\exp\Big(-2n\big(\delta - \tfrac{1}{n+1}\big)^2\Big).$$ □

Remark. The bound coincides exactly with the one obtained by applying Hoeffding's inequality directly to P_{cond}; the Beta–Binomial reduction is what makes that application rigorous, since Hoeffding requires a sum of independent bounded variables (here the n Bernoulli trials of the Binomial), not merely a single bounded variable. All downstream numerics ($n = 200$, $\delta = 0.1 \Rightarrow 2e^{-3.61} \approx 0.054$; the empirical std 0.035; the maximum gap 0.071) are therefore unchanged.

A.3 Proof of Theorem 3

Proof. **Step 1.** $0 \in \mathcal{A}$, and $\hat{\alpha}$ minimizes $\hat{L}_{\text{val}}$ over $\mathcal{A}$, so $\hat{L}_{\text{val}}(\hat{\alpha}) \le \hat{L}_{\text{val}}(0)$.

Step 2 (uniform Hoeffding). For each $\alpha \in \mathcal{A}$, $\hat{L}_{\text{val}}(\alpha)$ is a mean of N_{val} bounded terms in $[0, M]$. By Hoeffding, $\mathbb{P}[|L_{\text{ens}}(\alpha) - \hat{L}_{\text{val}}(\alpha)| \ge t] \le 2\exp(-2N_{\text{val}}t^2/M^2)$. Union-bound over $|\mathcal{A}|$ values:

$$\mathbb{P}[\exists \alpha :\ |L_{\text{ens}}(\alpha) - \hat{L}_{\text{val}}(\alpha)| \ge t] \le 2|\mathcal{A}|\exp(-2N_{\text{val}}t^2/M^2).$$

Setting the RHS to γ and solving, with probability $\ge 1 - \gamma$ we have uniformly over $\mathcal{A}$: $|L_{\text{ens}}(\alpha) - \hat{L}_{\text{val}}(\alpha)| \le M\sqrt{\ln(2|\mathcal{A}|/\gamma)/(2N_{\text{val}})}$.

Step 3. Apply the bound at $\alpha = \hat{\alpha}$: $L_{\text{ens}}(\hat{\alpha}) \leq \hat{L}_{\text{val}}(\hat{\alpha}) + M\sqrt{\ln(2|\mathcal{A}|/\gamma)/(2N_{\text{val}})}$.

Step 4. Substitute Step 1: $L_{\text{ens}}(\hat{\alpha}) \leq \hat{L}_{\text{val}}(0) + M\sqrt{\ln(2|\mathcal{A}|/\gamma)/(2N_{\text{val}})}$.

Step 5. Apply the Step-2 bound at $\alpha = 0$: $\hat{L}_{\text{val}}(0) \leq L_{\text{IRIS}} + M\sqrt{\ln(2|\mathcal{A}|/\gamma)/(2N_{\text{val}})}$. Adding gives

$$L_{\text{ens}}(\hat{\alpha}) \leq L_{\text{IRIS}} + 2M\sqrt{\frac{\ln(2|\mathcal{A}|/\gamma)}{2N_{\text{val}}}}. \qquad \square$$

B Per-Domain Tables

B.1 In-domain point metrics

Table 6: Per-domain in-domain accuracy: IRIS vs PLLS vs ensemble. α is the ensemble mixing weight selected on val; "Layer" is the layer selected by PLLS.

Domain	n	IRIS acc	PLLS acc	Ens acc	α	Layer
halueval:Bio-Medical	166	0.404	0.590	0.560	0.9	[20]
halueval:Education	147	0.660	0.680	0.707	0.5	[20]
halueval:Finance	189	0.772	0.772	0.772	0.0	[12]
halueval:Open-Domain	59	0.932	0.932	0.932	0.0	[24]
halueval:Science	200	0.505	0.650	0.625	0.9	[16]
helm:falcon40b	105	0.581	0.648	0.667	0.7	[12]
helm:gptj7b	115	0.730	0.739	0.739	1.0	[24]
helm:llamabase7b	113	0.575	0.637	0.619	0.1	[8]
helm:llamachat13b	96	0.698	0.688	0.698	0.0	[16]
helm:llamachat7b	100	0.620	0.640	0.640	1.0	[20]
helm:opt7b	114	0.684	0.693	0.675	0.6	[20]
true_false:animals	202	0.802	0.837	0.837	0.9	[16]
true_false:cities	292	0.955	0.966	0.966	0.8	[12]
true_false:companies	240	0.921	0.912	0.917	0.5	[28]
true_false:elements	186	0.892	0.871	0.892	0.0	[16]
true_false:facts	123	0.911	0.911	0.911	0.0	[12]
true_false:generated	49	0.755	0.735	0.755	0.1	[24]
true_false:inventions	176	0.898	0.915	0.909	0.5	[28]

B.2 In-domain conformal metrics (MCP, $\varepsilon = 0.1$)

Table 7: Per-domain MCP metrics for PLLS-CP in the in-domain setting.

Domain	coverage	abstention	sel. accuracy
halueval:Bio-Medical	0.916	0.765	0.641
halueval:Education	0.959	0.776	0.818
halueval:Finance	0.937	0.704	0.786
halueval:Open-Domain	0.949	0.559	0.885
halueval:Science	0.945	0.550	0.878
helm:falcon40b	0.943	0.790	0.727
helm:gptj7b	0.965	0.835	0.789
helm:llamabase7b	0.956	0.841	0.722
helm:llamachat13b	0.990	0.969	0.667
helm:llamachat7b	0.910	0.710	0.690
helm:opt7b	0.982	0.947	0.667
true_false:animals	0.871	0.144	0.850
true_false:cities	0.945	0.045	0.989
true_false:companies	0.933	0.175	0.919
true_false:elements	0.925	0.091	0.917
true_false:facts	0.894	0.016	0.909
true_false:generated	0.939	0.531	0.870
true_false:inventions	0.920	0.045	0.917

Table 8: Calibration examples per class. In-domain: the calib fold of the four-way split (Table 7). Transfer: the 100-point target calibration set (Table 2).

Domain	N	In-domain calib: halluc. / factual	Transfer calib (100): halluc. / factual
halueval:Bio-Medical	828	49 / 34	59 / 41
halueval:Education	731	41 / 33	55 / 45
halueval:Finance	944	22 / 73	23 / 77
halueval:Open-Domain	294	28 / 2	94 / 6
halueval:Science	996	49 / 51	49 / 51
helm:falcon40b	521	26 / 27	50 / 50
helm:gptj7b	572	41 / 17	70 / 30
helm:llamabase7b	565	32 / 25	57 / 43
helm:llamachat13b	476	15 / 33	31 / 69
helm:llamachat7b	497	19 / 31	38 / 62
helm:opt7b	566	38 / 19	68 / 32
true_false:animals	1008	50 / 51	50 / 50
true_false:cities	1458	73 / 73	50 / 50
true_false:companies	1200	60 / 61	50 / 50
true_false:elements	930	47 / 47	50 / 50
true_false:facts	613	31 / 31	50 / 50
true_false:generated	245	13 / 12	51 / 49
true_false:inventions	876	41 / 47	47 / 53

Mondrian CP estimates one threshold per class. A class with no calibration examples falls back to the pooled threshold (this never occurs here). With $\varepsilon = 0.1$ the finite-sample level $\lceil (n_y + 1)(1 - \varepsilon) \rceil / n_y$ exceeds 1 when a class has fewer than 9 calibration examples; it is then capped at 1, i.e. the threshold is the largest calibration score of that class. This happens only for the factual class of halueval:Open-Domain (2 in-domain, 6 in the transfer calibration set).

C Full Ablations

C.1 Effect of K (number of selected layers)

Table 9: Effect of K on representative domains (accuracy / AUC of PLLS classifier before ensembling).

Domain	K=2	K=4	K=6	K=8
true_false:facts	0.927 / 0.947	0.943 / 0.984	0.927 / 0.891	0.927 / 0.948
true_false:cities	0.908 / 0.954	0.908 / 0.942	0.908 / 0.894	0.908 / 0.868
halueval:Education	0.463 / 0.634	0.714 / 0.765	0.646 / 0.754	0.680 / 0.750

C.2 Entropy regularization on multi-layer attention aggregator

Table 10: Effect of entropy-regularization weight on accuracy of naive multi-layer attention (prior to PLLS selection).

Domain	w_{ent}=0.1	w_{ent}=0.3	w_{ent}=0.5
true_false:facts	0.943	0.927	0.927
true_false:cities	0.908	0.908	0.908
halueval:Education	0.714	0.605	0.558

C.3 Selection criterion

Table 11: Combined selection score (Eq. 1) vs AUC-only selection. Combined score gives more stable per-domain behaviour.

Domain	Criterion	acc	AUC	Selected layers
true_false:facts	combined	0.943	0.984	[20, 24, 28, 16]
true_false:facts	AUC only	0.927	0.949	[20, 16, 8, 24]
halueval:Education	combined	0.714	0.765	[12, 16, 20, 24]
halueval:Education	AUC only	0.429	0.573	[12, 8, 16, 28]

D Additional Model Family and Component Decomposition

To test that the gains are not specific to LLaMA, we ran the *full* PLLS-CP pipeline on a second family, Qwen3-8B [4], across all 18 domains, under the same protocol as the LLaMA cross-domain run. In-domain, mean accuracy improves from 0.687 (IRIS) to 0.720 (PLLS-CP), $\Delta = +0.032$, a margin comparable to LLaMA. Cross-domain transfer is likewise preserved (Table 12).

Table 12: Cross-domain transfer on Qwen3-8B (18 domains, 324 pairs). Retention is off-diagonal / diagonal.

Qwen3-8B	Diagonal (in-domain)	Off-diagonal (transferred)	Retention
IRIS	0.687	0.607	0.884
PLLS-CP	0.720	0.684	0.951

PLLS-CP is $\geq$ IRIS on both diagonal and off-diagonal accuracy, and the improvement is driven by a genuine gain in transferred accuracy ($0.607 \rightarrow 0.684$), as on LLaMA.

Table 13: Per-domain in-domain accuracy on Qwen3-8B: IRIS vs PLLS vs ensemble (PLLS-CP). α is the ensemble weight selected on val.

Domain	IRIS acc	PLLS acc	Ens acc	α
halueval:Bio-Medical	0.564	0.673	0.661	0.7
halueval:Education	0.552	0.552	0.552	0.0
halueval:Finance	0.733	0.722	0.733	0.0
halueval:Open-Domain	0.729	0.492	0.729	0.0
halueval:Science	0.520	0.672	0.672	1.0
helm:falcon40b	0.500	0.500	0.500	0.0
helm:gptj7b	0.296	0.435	0.443	0.1
helm:llamabase7b	0.451	0.531	0.531	0.2
helm:llamachat13b	0.695	0.695	0.695	0.0
helm:llamachat7b	0.620	0.540	0.620	0.0
helm:opt7b	0.558	0.584	0.584	0.9
true_false:animals	0.765	0.810	0.840	0.8
true_false:cities	0.945	0.945	0.945	0.0
true_false:companies	0.908	0.904	0.908	0.8
true_false:elements	0.866	0.871	0.871	0.4
true_false:facts	0.967	0.984	0.967	0.0
true_false:generated	0.816	0.510	0.816	0.7
true_false:inventions	0.886	0.886	0.886	0.0
Mean	0.687	0.684	0.720	

The ensemble is non-worse than IRIS on all 18 domains and strictly better on 7.

In-domain (Table 6), mean accuracy is 0.739 for IRIS alone, 0.768 for PLLS alone, and 0.768 for the ensemble: the in-domain gain comes from PLLS itself, and the ensemble's role there is non-degradation (Theorem 3).

Component decomposition. To attribute the cross-domain gain, we decompose transferred accuracy and retention into IRIS alone, PLLS alone (no ensemble), and the full ensemble (Table 14).

Table 14: Cross-domain component decomposition (off-diagonal accuracy / retention).

	LLaMA-3.1-8B		Qwen3-8B	
Component	off-diag	retention	off-diag	retention
IRIS alone	0.606	0.817	0.607	0.884
PLLS alone	0.618	0.806	0.644	0.941
PLLS + IRIS (ensemble)	0.708	0.922	0.684	0.951

The decomposition is on the same single-IRIS run as Table 2; since the conformal layer does not change the argmax, the ensemble row equals the PLLS-CP column of Table 2 exactly (0.708 off-diagonal, 0.922 retention on LLaMA). The two families differ in *where* the cross-domain robustness comes from: on LLaMA it is the ensemble step that lifts retention ($0.817 \rightarrow 0.922$; PLLS alone does not, at 0.806), whereas on Qwen PLLS alone already raises it ($0.884 \rightarrow 0.941$). In both cases the combined PLLS-CP is $\geq$ IRIS on transferred accuracy, and the ensemble is not a mere IRIS fallback: the validation-selected weight uses PLLS ($\alpha > 0$) on 13 of 18 domains on LLaMA and 9 of 18 on Qwen.

Larger backbone. We also ran the full pipeline with a larger reader, Qwen2.5-32B-Instruct [5], on all 18 domains (Tables 15–16).

Table 15: In-domain accuracy by benchmark: IRIS vs PLLS-CP, LLaMA-3.1-8B (Table 6) and Qwen2.5-32B-Instruct.

	LLaMA-3.1-8B		Qwen2.5-32B	
Benchmark	IRIS	PLLS-CP	IRIS	PLLS-CP
TrueFalse	0.876	0.884	0.908	0.909
HaluEval	0.655	0.719	0.700	0.726
HELM	0.648	0.673	0.600	0.615

Table 16: Cross-domain transfer on Qwen2.5-32B-Instruct (18 domains, 324 pairs). Retention is off-diagonal / diagonal.

Qwen2.5-32B	Diagonal (in-domain)	Off-diagonal (transferred)	Retention
IRIS	0.748	0.699	0.935
PLLS-CP	0.760	0.735	0.967

On Qwen2.5-32B, PLLS-CP is non-worse than IRIS on all 18 domains (strictly better on 5, mean $+0.013$) and improves transfer (off-diagonal $0.699 \rightarrow 0.735$, retention $0.935 \rightarrow 0.967$). With a larger reader, both the reader's own verbalized pseudo-label and the detector improve on HELM: verb$\rightarrow$GT AUC 0.67 (vs. 0.60 for LLaMA-3.1-8B), detector AUC 0.66–0.68 (IRIS 0.68, PLLS-CP ensemble 0.66; vs. ≈ 0.60 for LLaMA-3.1-8B IRIS), abstention ≈ 0.72. The detector tracks the quality of the pseudo-label, consistent with the pseudo-label being the binding constraint (Section 6.6).

E Pseudo-Label Reliability Diagnostic

To characterize when the unsupervised pseudo-label can be trusted, we compare, per benchmark, how well the verbalized-confidence pseudo-label agrees with ground truth (ROC-AUC of verbalized confidence against ground-truth labels, on the full domain) against a *diagnostic ceiling*: a probe from the same family PLLS uses (a single-layer linear probe) but trained on ground-truth labels (5-fold out-of-fold AUC on the full domain), used only to characterize the available signal and never as part

of the unsupervised pipeline. Their difference, $\text{headroom} = \text{ceiling} - \text{AUC}(\text{verb} \to \text{GT})$, is how much a better pseudo-label could still recover with the same read-out (Table 17).

Table 17: Pseudo-label agreement with ground truth, a same-family ground-truth-trained ceiling, and the resulting headroom. Verb→GT is the mean over domains (range in brackets); the ceiling is a single-layer linear probe. The ceiling is not part of the method; it isolates whether the binding constraint is the pseudo-label or the representation.

Benchmark	mean AUC(verb → GT)	Ceiling AUC	Headroom
TrueFalse	0.89 (0.78–0.96)	0.96	+0.07
HaluEval	0.65 (0.51–0.73)	0.69	+0.04
HELM	0.60 (0.56–0.63)	0.79	+0.19

The three benchmarks separate cleanly (Section 6.6): TrueFalse is the designed regime; on HaluEval the pseudo-label is already near the read-out's ceiling (headroom $+0.04$), so the binding constraint is the representation/read-out, not the label; on HELM the signal is present at ceiling 0.79 but the pseudo-label is only weakly informative, so the constraint is the pseudo-label. Every HELM domain is weakly informative:

Table 18: Per-domain verb→GT AUC on the six HELM domains (all above 0.5 but weak).

	gptj7b	falcon40b	opt7b	llamachat13b	llamabase7b	llamachat7b
AUC(verb → GT)	0.625	0.614	0.592	0.590	0.590	0.563

Two caveats. (i) The ceiling is a single-layer linear probe fit on ground-truth labels; it is a *lower bound* on the representation's content (a stronger read-out could extract more) and, being trained on more data than the main pipeline's probes, an appropriately generous estimate of what this read-out achieves. Headroom is therefore stated relative to the same probe family PLLS uses, not to the information-theoretic limit. (ii) One domain, halueval:Education, has verb→GT 0.690 above its own ceiling 0.654 (headroom -0.035)—the only such case among the 18—indicating that verbalized confidence there carries signal the single-layer linear read-out does not recover, which reinforces caveat (i).

F True-Label vs. Pseudo-Label Calibration

We isolate the cost of calibrating on pseudo-labels rather than true labels. The same frozen detector is recalibrated on (a) verbalized-confidence pseudo-labels and (b) ground-truth labels, and *true-label* coverage is measured in both cases. True-label calibration is the natural upper bound (standard supervised split-CP, near-nominal true-label coverage by construction). Relative to it, pseudo-label calibration recovers almost all of the true-label coverage on informative domains and shows no shortfall on HELM, where the conformal layer abstains heavily (Table 19).

Table 19: Shortfall of pseudo-label-calibrated true-label coverage relative to true-label calibration (the supervised upper bound); target $1 - \varepsilon = 0.90$, soft-pseudo detector, pooled over split and Mondrian.

Domain group	True-label coverage gap vs. true-label calibration (pseudo-calibration)
Informative (TrueFalse + HaluEval)	≈ -1.6 pp (recovers $\sim 98\%$ of GT-calibration coverage)
HELM	$\approx +0.6$ pp (no shortfall)
Overall	≈ -0.9 pp

The cost of not having true labels at calibration time is a small, bounded under-coverage ($\approx$ 1–2 pp on informative domains, none on HELM, where abstention absorbs the weak pseudo-label). In absolute terms, pseudo-label calibration attains a true-label coverage of 0.906 on informative domains, 0.931 on HELM and 0.914 overall (nominal 0.90).

G Pseudo-Label Inversion Stress Test

The guarantee is stated on the pseudo-label distribution, and the conformal layer abstains when the pseudo-label is merely *noisy*. It is natural to ask what happens when the pseudo-label is not noisy but *systematically inverted*. We test this directly by synthetically inverting the calibration pseudo-label at a controlled rate q (the frozen detector is unchanged; true-label coverage is measured on the test fold), aggregated over 3 domains $\times$ 3 seeds (Table 20).

Table 20: Pseudo-label inversion sweep. q is the fraction of calibration pseudo-labels flipped; metrics are averaged over 3 domains $\times$ 3 seeds. The ± 0.22 on selective accuracy at $q = 0.5$ marks the transition, where behaviour is unstable across runs.

q (inversion rate)	abstention	true-coverage	pseudo-coverage	sel. acc. (true)
0.0	0.11	0.92	0.92	0.93
0.2	0.61	0.99	0.92	0.97
0.4	0.78	0.98	0.92	0.91
0.5	0.81	0.95	0.90	0.79 (± 0.22)
0.6	0.76	0.77	0.89	0.06
0.7	0.65	0.67	0.88	0.05
0.8	0.54	0.56	0.90	0.05
0.9	0.38	0.41	0.93	0.05
1.0	0.11	0.19	0.92	0.09

The result is a sharp phase transition. Up to $q \approx 0.5$ (noisy / near-random pseudo-labels), the layer abstains and preserves true coverage. Beyond $q \approx 0.6$ (systematic inversion), it fails *silently*: at $q = 1.0$, pseudo-coverage still reports 0.92 while true coverage collapses to 0.19 and abstention drops to 0.11—confident, incorrect singletons. Two consequences. First, pseudo-coverage stays ≈ 0.90 across the entire sweep even as true coverage collapses, which is a concrete illustration that pseudo-label coverage is not, by itself, evidence of true-label reliability. Second, the abstention-based safety of the conformal layer is an *empirical* property bounded by pseudo-label inversion, not a guarantee: it holds while the pseudo-label is not systematically inverted ($q \lesssim 0.5$). All of our evaluated domains, including HELM (verb-vs-GT AUC 0.56–0.63; Appendix E), sit in the noisy regime *below* this boundary ($q < 0.5$), not in the inversion regime beyond it.

H Calibration Robustness under Exchangeability Violations

Theorem 1 assumes exchangeability of calibration and test data within the target domain. We study how coverage behaves when that assumption is deliberately violated, along class imbalance, calibration mismatch, and covariate shift, plus a finite-sample study. In all cases the detector is trained once per domain under the paper protocol; only the calibration/test folds are re-drawn, so the shift is isolated from any training effect.

This study uses the canonical soft-pseudo-label pipeline (Section 6).

Control (no shift). Under a correctly-specified calibration, coverage is at nominal: split CP 0.916, Mondrian 0.930, weighted CP 0.910. This study uses a separate repeated-fold protocol from the single-split in-domain evaluation of Table 4, so its control coverage is close to, but not identical to, the 0.938 reported there. The control, the $\gamma = 0.5$ row of Table 21, and the in-domain cells of the calibration-mismatch study are independent resamplings of the same setting and agree to within 0.002.

Class imbalance (structural robustness). Robustness to a change in the label prior is structural for Mondrian CP, not incidental. Because MCP calibrates *within each class*, class-conditional coverage implies marginal coverage for any label prior:

$$\mathbb{P}(Y \in C(X)) = \sum_y \mathbb{P}(Y{=}y)\, \mathbb{P}(Y \in C(X) \mid Y{=}y) \geq \sum_y \mathbb{P}(Y{=}y)(1-\varepsilon) = 1-\varepsilon,$$

so a change of $\mathbb{P}(Y)$ between calibration and test cannot break the guarantee. Empirically, over a 5×5 grid of hallucination-rate gaps up to ± 0.60, the fraction of runs violating coverage drops from

0.222 (split CP) to 0.103 (Mondrian); the worst domain improves from 0.896 to 0.930. Several of our real domains are already imbalanced (e.g. halueval:Open-Domain has a hallucination rate of 0.933, halueval:Finance 0.229), so this regime is not purely synthetic.

Calibration mismatch. Keeping the *source* threshold instead of recalibrating on the target costs up to 13 pp of coverage ($0.915 \rightarrow 0.788$ for split CP); target-side recalibration restores nominal coverage. This quantifies the operational value of Theorem 1: the recalibration step the paper proposes is precisely what prevents this failure.

Covariate-shift severity. Sweeping the strength of a covariate shift (representation separation γ from 0.5 to 1.0), coverage degrades nearly monotonically and moderately for both methods: Mondrian loses about 1.5 pp (from 0.929 to 0.914) at full separation (Table 21) and stays above split CP at every γ; we still attribute the Mondrian advantage to class imbalance (above), where it is structural, rather than to covariate shift. If the covariate shift is known, weighted conformal prediction [37] restores coverage (0.935 at $\gamma = 1.0$), at the cost of higher abstention—a standard fallback within the same framework.

Table 21: Coverage under increasing covariate-shift strength γ ($\gamma = 0.5$: no shift; $\gamma = 1.0$: full representation separation).

γ	split CP	Mondrian	weighted CP
0.5 (no shift)	0.916	0.929	0.911
0.6	0.915	0.929	0.909
0.7	0.912	0.927	0.904
0.8	0.913	0.925	0.907
0.9	0.909	0.918	0.909
1.0 (max)	0.907	0.914	0.935

Temporal (record-order proxy). The benchmarks carry no timestamps, so we use record order as a *proxy* for temporal drift: calibrate on the earliest records, test on the latest. Coverage holds on average (split CP 0.904, Mondrian 0.929); with only 18 ordered cells we draw no conclusion about method ordering here. We stress that this is a proxy—record order need not reflect genuine distributional drift over time—and leave a study on truly time-stamped streams to future work.

Finite-sample behaviour. The empirical coverage distribution matches the Beta–Binomial prediction of Theorem 2 closely (Table 22); below $n \approx 100$ the predictor is markedly conservative (discreteness of the quantile), and $n \geq 100$ is required for the empirical distribution to match the prediction. This supports the $n \in [100, 200]$ recommendation of Section 4.6 with direct evidence.

Table 22: Empirical vs. Beta–Binomial-predicted coverage as calibration size n varies (target 0.90).

n_{calib}	coverage (emp.)	coverage (theor.)	sd (emp.)	sd (theor.)
100	0.9102	0.9010	0.0419	0.0436
200	0.9054	0.9005	0.0378	0.0388

I Comparison Against Other Unsupervised Detectors

Our four constraints (Section 3) restrict direct baselines to IRIS. For context on where IRIS itself stands, Table 23 reproduces, as reported by Srey et al. [11] on the same three benchmarks and base model, IRIS's comparison against other unsupervised internal-state and uncertainty methods. IRIS dominates them by a wide margin; several of these competitors are moreover evaluated under *relaxed* constraints (SAR and EigenScore [20] use multiple generations per query; CCS uses contrast pairs; MIND [32] uses Wikipedia-anchored labels) and still underperform IRIS. PLLS is a strict extension of IRIS (Theorem 3), so it inherits this positioning while adding layer selection and cross-domain conformal guarantees.

Table 23: IRIS vs. other unsupervised detectors, as reported by Srey et al. [11] on the same benchmarks and base model. Values are accuracy (%) as reported therein.

Benchmark	EigenScore	SAR	CCS	MIND	**IRIS**
TrueFalse	56.1	59.9	63.2	52.4	**90.4**
HaluEval	43.0	66.0	56.9	50.7	**70.6**
HELM	42.4	58.7	56.8	54.1	**68.4**

J Implementation and Computational Details

Environment. Python 3.12.2; PyTorch, HuggingFace Transformers, NumPy/Pandas, scikit-learn. All experiments were run on a single CPU workstation for the downstream detector; hidden-state extraction from LLaMA-3.1-8B-Instruct, Qwen3-8B and Qwen2.5-32B-Instruct was pre-computed on GPU and cached.

Runtime. A full run on a single domain (training + validation + calibration + test) takes ≈ 19 seconds on CPU. PLLS classifier training dominates the cost; conformal calibration and metric computation are negligible. The 18-domain sweep scales linearly.

Memory. Peak RAM during a single-domain run is ≈ 1.06–1.25 GB, dominated by per-layer hidden-state tensors.

Complexity. Let N be the number of examples, L the number of candidate layers, K the number of selected layers, d the embedding dimension. Layer probing: $\mathcal{O}(LNd)$ per epoch. Attention aggregation on selected layers: $\mathcal{O}(NK^2d)$. Ensembling: $\mathcal{O}(N)$. Conformal calibration: $\mathcal{O}(N \log N)$; inference: $\mathcal{O}(N)$.

Hyperparameters. Attention classifier: projection size 256, 4 attention heads, AdamW with learning rate 5×10^{-4}, reduce-on-plateau scheduler, dropout 0.1. Ensemble grid: $\alpha \in \{0, 0.1, \ldots, 1.0\}$. Conformal level: $\varepsilon = 0.1$. Layer candidate set: $\{0, 4, 8, 12, 16, 20, 24, 28\}$. Selection score weights: $\lambda = 0.7$ on AUC, $1 - \lambda = 0.3$ on $(1 - \mathrm{ECE})$. Selection size: $K = 1$.

Data splits. Train/val/calib/test are disjoint within each domain; splits are fixed with a seed. For cross-domain experiments, the detector uses source train+val, and calib/test of the *target* domain drives calibration and evaluation.

Reproducibility. Source code is available at `https://github.com/v1ad-c/plls_cp`. The repository contains the full pipeline: pseudo-label and hidden-state generation, layer selection, detector training, conformal calibration, and in-domain and cross-domain evaluation, with fixed seeds and a README with instructions to reproduce the main in-domain and cross-domain results.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state four contributions: unsupervised layer selection (Section 4), ensembling with non-degradation guarantee (Theorem 3), cross-domain conformal guarantees (Theorems 1, 2), and empirical validation on 18 domains with 324 transfer pairs (Section 6). All claims are matched by corresponding sections, and the scope is explicitly bounded by the four constraints (knowledge-free, white-box, post-hoc, unsupervised) stated in Section 3.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section 6.6 (HELM domains as a weak-pseudo-label regime of CoT-embedding methods) and in the Limitations and Future Work paragraph at the end of Section 8 (two model families evaluated, generalization to further families remains to be verified, lack of multi-seed retraining due to compute constraints).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All three theorems (Theorems 1, 2, 3) explicitly state their assumptions (exchangeability, fixed nonconformity score before target observation, bounded loss). Proof sketches are given immediately after each theorem, and full step-by-step proofs are provided in Appendix A. References to classical results ([33], [36]; Hoeffding's inequality) are included where used.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 6 (Setup) describes the base model, benchmarks, splits, and protocols. Appendix J (Implementation and Computational Details) provides the environment, runtime, memory, complexity, and the full hyperparameter list. Per-domain results, ablation tables, and the conformal level $\varepsilon = 0.1$ are all explicitly reported. Pseudo-label generation follows the published IRIS protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is available at `https://github.com/v1ad-c/plls_cp` with instructions to reproduce the main in-domain and cross-domain results. All datasets used (TrueFalse, HaluEval, HELM) are publicly available and cited in Section 6.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 3 (Leakage-safe split paragraph) defines the four-way train/val/calib/test partition. Section 4.2 explicitly describes how $\lambda = 0.7$ was selected (on the train/validation splits of three randomly chosen domains, with the calibration and test splits of all 18 domains held out). Section 6 states the conformal level $\varepsilon = 0.1$ and benchmark composition. Appendix J lists all hyperparameters: optimizer schedule, projection size, learning rate, regularization weights, ensemble grid, candidate layer set.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report bootstrapped confidence intervals for individual cells of the cross-domain matrix because each evaluation is performed on the held-out target test split with fixed seed. Reported aggregates (retention, mean improvement) are instead supported by (i) a large evaluation surface ($18 \times 18 = 324$ transfer pairs), (ii) an explicit theoretical concentration bound (Theorem 2) for the most variable metric (coverage gap), and (iii) empirical verification of that bound (Section 6.4, observed standard deviation 0.035 matching the predicted regime). We did not run multi-seed retraining due to compute constraints; this is acknowledged in the limitations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix J reports compute resources: hidden-state extraction was precomputed on GPU and cached; downstream detector training and evaluation run on a single CPU workstation, with ≈ 19 seconds per single-domain run and peak memory ≈ 1.06–1.25 GB. The full 18-domain sweep scales linearly.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses publicly available benchmarks and publicly released language models. No human subjects, no scraped personal data, no deployment of generative models with identified misuse risks. The work concerns reliability and safety of LLM outputs and aligns with the Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Positive impact: improving the reliability of LLMs in high-stakes settings (healthcare, finance, scientific reasoning) via principled abstention rather than overconfident generation; the framework is unsupervised and inexpensive to deploy. Potential negative impact: the existence of a confidence-aware detector may foster over-reliance on automated filtering; users should not assume that uncovered or abstained outputs are safe in every setting. Section 6.6 discusses the HELM regime, where the self-generated pseudo-label is weak and the detector's reliability margin is smaller, which is a relevant deployment caveat.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release new models, datasets, or generative tools that present high misuse risk. We release only an analysis pipeline (detector + conformal layer) trained on top of existing publicly available LLMs.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All used assets are cited in the bibliography: LLaMA-3.1-8B-Instruct [3] under the Llama 3.1 Community License; Qwen3-8B [4] and Qwen2.5-32B-Instruct [5] under the Apache 2.0 license; TrueFalse [12], HaluEval [13, 14], and HELM [32] are cited and used under their respective publicly available terms; IRIS [11] is cited as the baseline method. No licensing terms are violated.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The released repository includes the PLLS-CP code, pseudo-label generation scripts, fixed seeds, and a README with run instructions. No human-subjects data or third-party assets requiring consent are introduced.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects. All labels used are pseudo-labels generated by the LLM itself via verbalized confidence; no human annotators are employed.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The LLMs (LLaMA-3.1-8B-Instruct, and Qwen3-8B and Qwen2.5-32B-Instruct in Appendix D) are the central object of study: their hidden states form the input to the detector, and their verbalized confidence is the source of pseudo-labels. The models and their role are described throughout Section 3 (Hidden-state tensors), Section 4, and Section 6. The use of LLMs for manuscript editing (grammar, phrasing) is separately disclosed via the conference submission form.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.