

Diagnostic-Driven Selective Abstention in Agentic Pipelines: A Case Study on AVeriTeC Dataset

Alberto Moccardi^{1,*}, Rajib Chandra Ghosh¹, Egidia Cirillo¹ and Flora Amato¹

¹University of Naples Federico II, DIETI, Via Claudio 21, 80125, Naples, Italy

Abstract

Agentic LLM pipelines operate under dynamic and noisy conditions, systematically producing unreliable outputs without revealing which module failed. We propose a diagnostic framework that supports selective prediction in uncertain conditions, thus improving both accuracy and modular accountability of AI systems. The abstention mechanism is regulated by per-module nonconformity scores, calibrated via split conformal prediction and combined into an emit-or-escalate routing rule. Experimental results on the AVeriTeC fact-checking benchmark validate the proposed Conformal Selective Abstention (CSA) framework, which attains the highest accuracy on emitted items among the methods compared, raising it from 0.67 to 0.79 (95% CI [0.64, 0.88]) at 69% output coverage. A per-class analysis reveals an intrinsic asymmetry of this conservative approach: CSA escalates majority-class errors but also diverts correct minority-class items, reducing class-coverage macro recall from 0.52 to 0.34. A Mondrian extension partially addresses this asymmetry by exploiting within-stratum structure, and we analyze the resulting trade-off through a deployment-cost lens. All code and data necessary to reproduce our analysis are publicly available.¹

Keywords

AI systems, Large Language Models, Fact Checking, Agentic AI, Software Architecture

1. Introduction

Agentic LLM pipelines compose specialised modules (e.g., decomposer, a retriever, a reader, a verifier) into a single inference path. Each module is stochastic and miscalibrated; their composition inherits every component’s failure modes while obscuring which component is responsible. For high-stakes tasks such as automated fact-checking, where a wrong verdict can be more harmful than no verdict, verification of automated decision pipelines is well established across domains ; in this work we focus on *selective abstention* [1, 2], a calibrated decision system that emits the verdict or escalates to human review.

Conformal prediction [3, 4] provides finite-sample coverage guarantees and has been applied to LLM outputs [5, 6, 7], but never to multi-module agentic pipelines with cross-module attribution. We address this gap on AVeriTeC [8] through three research questions:

RQ1 *What selective signal do per-module nonconformity scores carry, and what does conformal calibration add over simple thresholds?*

RQ2 *Does module attribution escalate genuinely harder items?*

RQ3 *How does selective abstention interact with class imbalance, and can within-stratum recalibration address the asymmetry?*

Our contributions are: (i) an emit-or-escalate framework with module attribution for agentic LLM pipelines; (ii) a per-class analysis exposing an accuracy–recall asymmetry in selective abstention under imbalance; and (iii) a Mondrian extension that partially addresses it.

¹<https://github.com/Al-Moccardi/Conformal-Diagnostic-for-LLM-Pipeline>

WOA 2026: 27th Workshop “From Objects to Agents”, June 2026, Italy

*Corresponding author.

✉ alberto.moccardi@unina.it (A. Moccardi); rajib.chandraghosh@unina.it (R. C. Ghosh); egidia.cirillo@unina.it (E. Cirillo); flora.amato@unina.it (F. Amato)

🆔 0009-0001-6136-9368 (A. Moccardi); 0009-0001-1137-3465 (R. C. Ghosh); 0009-0005-3227-6073 (E. Cirillo); 0000-0002-5128-5558 (F. Amato)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Related Work

Selective classification [1, 2] trades coverage for accuracy by allowing abstention on uncertain items. Conformal prediction [3, 4] provides distribution-free prediction sets, extended by Mondrian calibration [9] to per-stratum guarantees and by adaptive prediction sets [10] to per-class coverage. Recent conformal work targets LLM outputs: multi-choice QA [5], generation stopping criteria [6], nucleus-sampling calibration [11], and self-evaluation as a selective signal [12, 13]. In selective QA, Kamath et al. [14] study abstention under domain shift, and Varshney et al. [15] benchmark selective prediction across tasks. AVeriTeC [8] extends earlier benchmarks [16] with real-world claims and web evidence via retrieval-augmented reasoning [17]. None of the cited works addresses multi-module agentic pipelines with cross-module attribution.

3. Methodology

3.1. Four-module pipeline architecture

We denote the number of pipeline modules by K ; the full pipeline has $K = 4$ and processes a single claim in four sequential stages (Tab. 1). The **decomposer** (M1) takes the raw claim and produces atomic sub-claims that downstream modules can address independently. The **retriever** (M2) issues semantic searches against an evidence corpus and returns a ranked list of passages per sub-claim. The **reader** (M3) synthesises the retrieved evidence into a structured summary. The **verifier** (M4) draws $K_s = 10$ stochastic samples at temperature $T = 0.3$ over the four AVeriTeC verdict classes and emits the modal class as the final prediction. All modules are quantised open-weight models served locally, ensuring reproducible and observable sampling distributions.

Table 1

Pipeline modules and their backbones.

Module	Role	Backbone
M1 Decomposer	Sub-claim extraction	llama3.2:3b
M2 Retriever	Evidence retrieval	nomic-embed-text
M3 Reader	Evidence synthesis	llama3.1:8b
M4 Verifier	Verdict sampling ($K_s = 10$)	qwen2.5:14b

3.2. Per-module nonconformity scores

Each module produces a scalar nonconformity score $s_k(x)$, where higher values signal greater deviation from calibration behaviour. By construction $s_2, s_4 \in [0, 1]$; s_1 and s_3 lie in $[0, 2]$ formally (cosine similarity ranges over $[-1, 1]$) but empirically take values in $[0, 1]$ on our data. Let $\cos(\mathbf{u}, \mathbf{v})$ denote cosine similarity between embedding vectors.

s_1 (**decomposer**). Given predicted sub-questions $\{p_i\}_{i=1}^P$ and gold sub-questions $\{g_j\}_{j=1}^G$ with embeddings from nomic-embed-text:

$$s_1(x) = 1 - \frac{1}{G} \sum_{j=1}^G \max_{i \in \{1, \dots, P\}} \cos(\mathbf{p}_i, \mathbf{g}_j). \quad (1)$$

Higher s_1 means poorer coverage of the reference decomposition.

s_2 (**retriever**). Let $\mathcal{D}^*$ be the set of gold evidence documents and $\text{rank}(d)$ the position of document d in the retrieved top-10 list (∞ if absent):

$$s_2(x) = 1 - \frac{1}{|\mathcal{D}^*|} \sum_{d \in \mathcal{D}^*} \frac{1}{\text{rank}(d)}. \quad (2)$$

This is $1 - \text{MRR}$; $s_2 = 0$ for perfect retrieval, $s_2 = 1$ when no gold document appears.

s_3 (**reader**). Given $N=5$ stochastic reader outputs $\{r_i\}$ generated at $T=0.7$:

$$s_3(x) = 1 - \frac{2}{N(N-1)} \sum_{i < j} \cos(\mathbf{r}_i, \mathbf{r}_j). \quad (3)$$

This semantic self-consistency score is distribution-free and requires no gold answer, preserving exchangeability with the full calibration set. Low s_3 indicates confident, clustered outputs; high s_3 indicates divergent, uncertain outputs.

s_4 (**verifier**). Let $\hat{\pi}$ be the empirical label distribution from $K_s=10$ stochastic verifier samples, with classes sorted by decreasing frequency $\hat{\pi}_{(1)} \geq \dots \geq \hat{\pi}_{(L)}$, and let y be the gold label. The Adaptive Prediction Set (APS) score [18] is:

$$s_4(x) = \sum_{\ell=1}^L \hat{\pi}_{(\ell)} \mathbf{1}[\text{rank of } y \geq \ell] = \sum_{\ell=1}^{r_y} \hat{\pi}_{(\ell)}, \quad (4)$$

where r_y is the rank of the gold label. When all K_s samples agree on a single label, $\hat{\pi} = (1, 0, \dots, 0)$ and $s_4 = 1.0$ regardless of correctness — we term this *APS saturation*. On our calibration set, 53/61 items saturate, rendering s_4 uninformative for the dominant regime and motivating the Mondrian extension (Sec. 3.5).

Scores s_1 (Eq. 1) and s_2 (Eq. 2) require gold annotations at calibration time (standard for split conformal prediction), while s_3 (Eq. 3) is self-supervised. The codebase includes alternative scores (negative log-probability, disagreement-rate, margin) explored during development but not used in the reported configuration.

3.3. Split conformal calibration

Given a calibration set $\mathcal{C}$ of size $|\mathcal{C}| = 61$ and a target risk level $\alpha \in [0, 1]$, we compute the per-module $(1-\alpha)$ -quantile with the standard finite-sample correction:

$$\hat{q}_k(\alpha) = \text{Quantile}_{(1-\alpha)(1+1/|\mathcal{C}|)}(\{s_k(x) : x \in \mathcal{C}\}), \quad (5)$$

delivering the marginal guarantee $\Pr(s_k(X_{\text{test}}) \leq \hat{q}_k(\alpha)) \geq 1 - \alpha$ *per module independently*. Two levels are used: a strict level α_h and a moderate level $\alpha_m = \alpha_h + 0.10$. The strict threshold flags items in the tight tail of a module’s score distribution; the moderate threshold flags items that are merely unusual. The combined emit-or-escalate rule is a function of all four scores and does *not* inherit a single closed-form joint coverage guarantee; its empirical performance is validated on the test split (Sec. 5).

Two of our scores have a point mass at the upper bound: $s_2 = 1$ when no gold document is retrieved (a hard retrieval failure), and $s_4 = 1$ under APS saturation. The marginal coverage guarantee (Eq. 5) holds regardless, but the conformal quantile $\hat{q}_k(\alpha)$ saturates at the upper bound whenever the calibration saturation rate exceeds α . In our calibration data, $s_2 = 1$ on 33% of items and $s_4 = 1$ on 87%, so the strict quantile $\hat{q}_4 = 1$ at every tested $\alpha_h \in [0.20, 0.40]$, and $\hat{q}_2 = 1$ at $\alpha_h \leq 0.35$ (at $\alpha_h = 0.40$, $\hat{q}_2 = 0.875$). The decision rule (Sec. 3.4) flags items with $s_k(x) \geq \hat{q}_k(\alpha)$, so these saturated thresholds remain operational — they reduce to the membership tests “FLAG IF $s_2 = 1$ ” (complete retrieval failure) and “FLAG IF $s_4 = 1$ ” (full APS saturation), the latter firing on most items and providing no discriminative signal. This degeneracy is why the ESCALATE-VERIFIER channel never fires in isolation (Sec. 5.1) and why we introduce the score-stratum Mondrian extension (Sec. 3.5), which conditions on the saturated stratum and re-calibrates a non-saturated score (s_3) within it.

3.4. Emit-or-escalate decision rule

For each test item, the framework counts modules exceeding their strict quantile and modules exceeding their moderate quantile. The item is **escalated** if at least one module exceeds its strict threshold or two or more modules simultaneously exceed their moderate thresholds; otherwise it is **emitted**. Each escalated item is attributed to one of four sources, providing actionable information for the human reviewer:

- **ESCALATE-RETRIEVAL**: M2 alone exceeds its strict quantile — the evidence base is unreliable.
- **ESCALATE-READER**: M3 alone exceeds its strict quantile — the evidence synthesis failed.
- **ESCALATE-VERIFIER**: only M4 flags — the upstream pipeline operated nominally but the verifier’s stochastic samples conflict.
- **ESCALATE-MULTIPLE**: ≥ 2 modules in their moderate tails — the input is anomalous on multiple dimensions.

M1 (decomposer) is empirically not predictive of $K=4$ verdict failure on this benchmark; items where M1 alone triggers are folded into ESCALATE-MULTIPLE.

3.5. Score-stratum Mondrian extension

The marginal scheme treats calibration as one homogeneous population. Because minority classes carry only 7% (NEE) and 8% (Conflicting) of calibration mass, they contribute no separable signal to per-module quantiles. We introduce a Mondrian conformal extension [9] that exploits class-conditional structure *within* the APS-saturated stratum.

Mondrian conformal prediction partitions $\mathcal{C}$ into disjoint strata $\{\mathcal{C}_S\}_{S \in \mathcal{S}}$ via a measurable function $\kappa : \mathcal{X} \rightarrow \mathcal{S}$ observable at test time, and calibrates a separate threshold in each stratum:

$$\hat{q}_S(\alpha) = \text{Quantile}_{(1-\alpha)(1+1/|\mathcal{C}_S|)}(\{s(x) : x \in \mathcal{C}_S\}). \quad (6)$$

For a test point assigned to $S = \kappa(X_{\text{test}})$, the within-stratum guarantee is $\Pr(s(X_{\text{test}}) \leq \hat{q}_S(\alpha)) \geq 1 - \alpha$ under conditional exchangeability.

We stratify on verifier APS saturation by defining

$$\kappa(x) = \mathbf{1}[s_4(x) = 1.0], \quad \mathcal{S} = \{A, B\}, \quad (7)$$

yielding **Stratum A** (APS-saturated; $|\mathcal{C}_A|=53$) and **Stratum B** (unsaturated; $|\mathcal{C}_B|=8$). Although $s_4(x)$ in Eq. 4 depends on the gold label y , the saturation event $s_4(x) = 1$ occurs if and only if the verifier’s K_s samples all coincide on a single class (Sec. 3.2), which is determined by the prediction distribution alone; hence κ is label-independent and the conditional-exchangeability assumption underlying Eq. 6 is preserved. We do not stratify on the predicted class because that would condition on the prediction we are validating; we instead stratify on an upstream observable (APS saturation) that isolates the regime in which s_4 is uninformative. Within Stratum A we re-calibrate a different score (s_3), which side-steps the saturation. Stratum B is too small ($n=8$) for a reliable threshold; the method’s empirical contribution comes from Stratum A.

Within Stratum A, the reader score s_3 (Eq. 3) shows class-conditional structure: NEE items cluster at low $s_3 \approx 0.04$, Conflicting items at moderate $s_3 \approx 0.07\text{--}0.13$, and majority classes spread across the threshold. We compute $\tau_A = \hat{q}_A(\alpha)$ via Eq. 6 on s_3 and divert any EMIT item to **ESCALATE-MONDRIAN-M3** according to:

$$\text{divert}(x) = \mathbf{1}[s_3(x) > \hat{q}_{\kappa(x)}(\alpha)]. \quad (8)$$

The filter (Eq. 8) operates at the routing layer only: the verdict itself is unchanged. M3 is the unique score with both sufficient within-stratum range and class-conditional separation; M2 and M4 saturate at 1.0 on the bulk of items, and M1 is uninformative.

4. Experimental Setting

4.1. Dataset and data splits

We evaluate on AVeriTeC [8], which provides 4,568 real-world claims with verdicts in four classes: Supported, Refuted, Not Enough Evidence (NEE), and Conflicting Evidence/Cherry-picking. From the validation split we select 500 items and create a stratified 61/61 calibration–test partition (random seed

42), the largest equal split preserving class proportions. Both splits share the same class distribution: Refuted 37 (61%), Supported 15 (25%), Conflicting 5 (8%), NEE 4 (7%). The heavy class imbalance is a known property of AVeriTeC and directly motivates the Mondrian extension. The strict α_h sweep is $\{0.20, 0.25, 0.30, 0.35, 0.40\}$ with $\alpha_h = 0.30$ as the reference operating point.

4.2. System configuration and reproducibility

The hardware and software configuration for the experiments is represented by CPU Intel Ultra 9 185H (16C/22T, 2.5 GHz, L3 24 MB) and GPU Intel Arc (NVIDIA RTX 4070). All modules run locally via Ollama; the full pipeline requires ~ 1.5 h GPU time for verifier sampling and ~ 5 s CPU time per post-hoc diagnostic analysis. No cloud APIs are used. All configurations, source code, and results are available at <https://github.com/Al-Moccardi/Conformal-Diagnostic-for-LLM-Pipeline>.

4.3. Evaluation metrics

We report: (i) **verdict accuracy** on emitted items (modal prediction vs. gold label) with Wilson 95% confidence intervals; (ii) **coverage** (fraction of test items emitted); (iii) **class-coverage recall** (= correct in emit/total in class), the standard selective-classification metric [2]; (iv) the **AURC** (Area Under the Risk-Coverage curve); and (v) **would-be accuracy** of escalated items, measuring whether attribution sources isolate genuinely harder items. We use verdict accuracy rather than the joint evidence-and-verdict AVeriTeC score; extending evaluation to the full AVeriTeC score is future work.

5. Results

5.1. Selective abstention performance

Fig. 1 shows per-module score distributions across the α_h sweep.

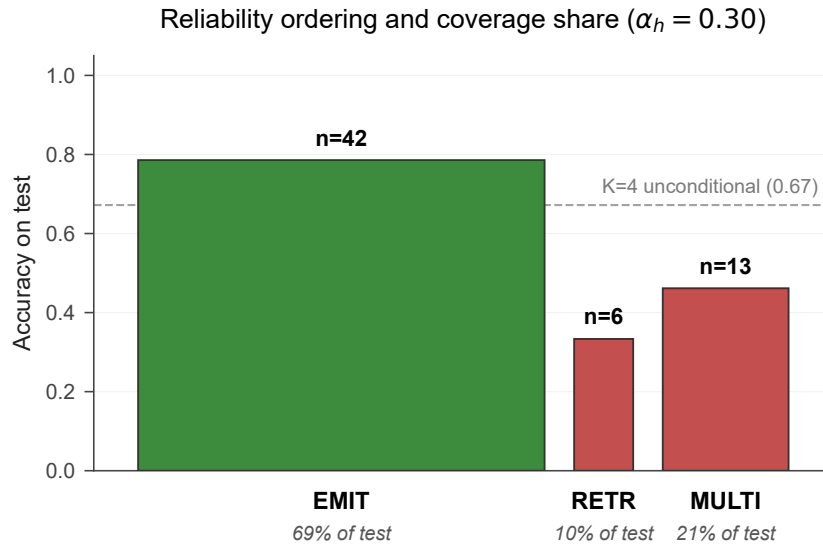


Figure 1: Per-module score distributions, EMIT (green) vs. ESCALATE (red), across α_h levels. M2’s spike at 1.0 identifies retrieval failure; M4’s saturation at 1.0 characterises the stratum exploited by the Mondrian extension.

M2 (retriever) displays a pronounced spike at $s_2 = 1.0$ on escalated items, identifying retrieval failure as the dominant escalation trigger. M4 (verifier) is uniformly concentrated at $s_4 = 1.0$ due to APS saturation (Sec. 3.2).

Fig. 2 displays the reliability ordering at $\alpha_h = 0.30$ (**RQ2**): EMIT ($n = 42$, 69% coverage) achieves accuracy 0.786; the two attribution categories that escalate to human review — ESCALATE-RETRIEVAL ($n = 6$, would-be accuracy 0.33) and ESCALATE-MULTIPLE ($n = 13$, would-be accuracy 0.46) — isolate

items substantially harder than the unconditional baseline of 0.672. The attribution sources route to human review precisely those items the pipeline is least likely to classify correctly. ESCALATE-READER never fires at any tested α_h (the reader is not the strict-tail violator alone on this benchmark); ESCALATE-VERIFIER-CONFLICT *does* flag items where $s_4 = 1$ but the upstream is clean ($n = 13$ at $\alpha_h = 0.30$), yet the framework emits these verdicts because the upstream behaved nominally and the verifier saturation reflects modal agreement rather than disagreement; we group them with EMIT in the headline accuracy.

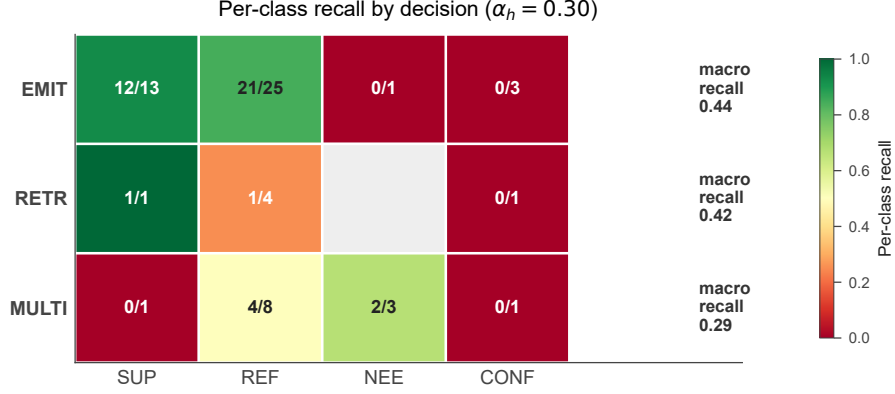


Figure 2: Reliability ordering at $\alpha_h = 0.30$. Bar widths proportional to the fraction of the $n = 61$ test set.

5.2. Baselines and risk-coverage analysis (RQ1)

Tab. 2 compares the multi-module conformal framework against three single-signal baselines: random selection at matched coverage, retrieval-only ($s_2 < 1.0$), and reader-only (s_3 below its 70th calibration percentile).

Table 2

Emit accuracy with 95% Wilson CIs. Top: single-signal baselines. Bottom: multi-module conformal framework.

Method	n (cov.)	Acc	Wilson 95% CI
Unconditional (no filter)	61 (100%)	0.672	[0.55, 0.78]
Random@69%	42 (69%)	0.672 [†]	—
Retrieval-only ($s_2 < 1$)	44 (72%)	0.750	[0.61, 0.85]
Reader-only (s_3 P70)	48 (79%)	0.729	[0.59, 0.83]
Conformal $\alpha_h = 0.30$	42 (69%)	0.786	[0.64, 0.88]
Conformal $\alpha_h = 0.35$	44 (72%)	0.750	[0.61, 0.85]
Conformal $\alpha_h = 0.40$	56 (92%)	0.696	[0.57, 0.80]

[†]Expected over 10,000 random draws.

At $\alpha_h = 0.30$ the multi-module framework achieves the highest emit accuracy (0.786) among the methods compared. We test pairwise differences using McNemar’s test, which is the standard statistical test for comparing two classifiers on the same labelled test set because it conditions on the discordant pairs (items where the two methods disagree) and is therefore unaffected by items both methods classify identically. Applied on shared-emit subsets, it returns $p = 0.50$ versus retrieval-only and $p = 0.82$ versus reader-only: the differences lie within sample noise at $n = 61$. We therefore present the ordering as indicative rather than statistically established. Two structural observations remain robust to this caveat. First, of the 19 items the framework escalates at $\alpha_h = 0.30$, 13 are caught only by the retrieval signal ($s_2 = 1$, s_3 below P70), 2 only by the reader signal, and 4 by both — the conformal escalation set is contained in the union of single-signal escalations.

The framework does not surface items neither baseline could find; what it provides is a calibrated rule combining the two signal types with explicit attribution to the firing module. Second, conformal calibration carries a distribution-free per-module guarantee and a principled α that transfers across pipelines, neither of which the heuristic single-signal thresholds enjoy. Fig. 3 shows the risk–coverage curve (AURC = 0.200; oracle 0.061).

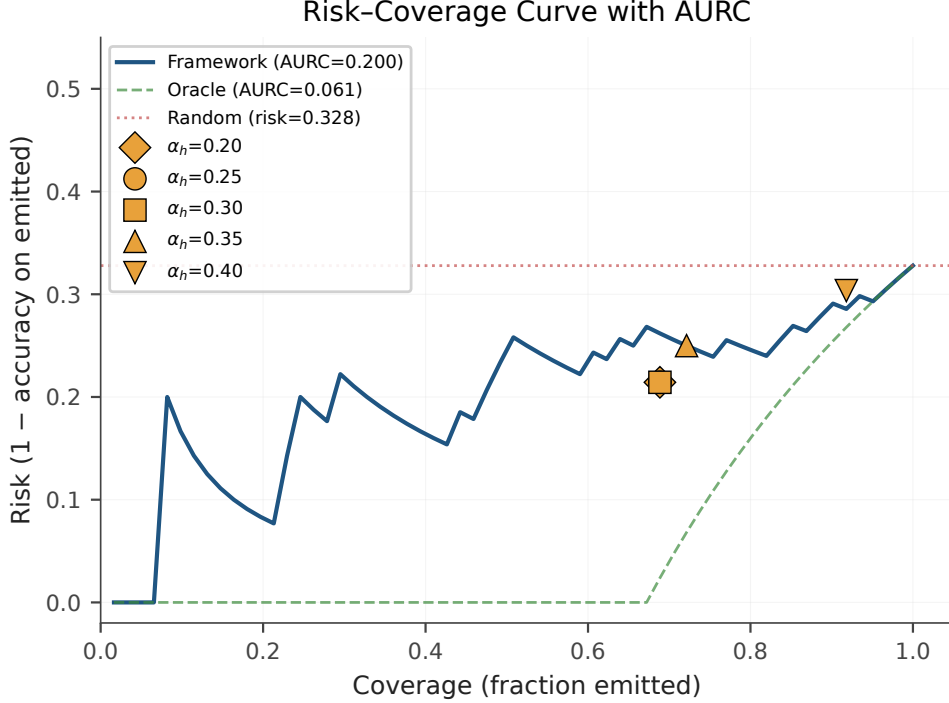


Figure 3: Risk–coverage curve. AURC = 0.200 ; oracle 0.061. Markers show the α_h operating points.

5.3. Per-class analysis: the accuracy–recall asymmetry (RQ3)

Tab. 3 decomposes performance by class across all methods. The selective framework improves accuracy (0.672 $\rightarrow$ 0.786) but *reduces* class-coverage macro recall from 0.517 to 0.342. We omit per-class CIs because per-class denominators (4 and 5 for the minority classes) make the Wilson intervals span almost the full $[0, 1]$ range and would be uninformative.

Table 3

Class-coverage recall by method (= correct in emit/total in class). Full per-class F1 is not computable because predicted labels for wrong items are not stored; class-coverage recall is the standard selective-classification metric [2].

Method	n	Acc	R_{SUP}	R_{REF}	R_{NEE}	R_{CONF}	MacroR
Unconditional	61	.672	.87	.70	.50	.00	.517
Reader-only	48	.729	.73	.59	.50	.00	.457
Retrieval-only	44	.750	.80	.57	.00	.00	.342
Conformal .30	42	.786	.80	.57	.00	.00	.342
Mondrian-M3	35	.800	.67	.49	.00	.00	.288

A striking pattern emerges: the *reader-only* baseline preserves NEE recall (0.50) because the reader is consistent on NEE items ($s_3 \approx 0.03$ – 0.09 , below the threshold). It is the *retriever* that flags NEE: 3 of 4 NEE items have $s_2 = 1.0$ (complete retrieval failure). The multi-module framework captures this signal and escalates all three: 2 of these are correct (the framework over-escalates) and 1 is wrong

(correctly caught). Among the 19 escalated items at $\alpha_h = 0.30$, 8 would have been correct, including both correct NEE items. The accuracy–recall trade-off is module-specific: combining more signals improves majority-class accuracy but amplifies minority-class over-escalation.

5.4. Mondrian extension: a cost-aware routing analysis

Fig. 4 visualises the class-conditional structure within Stratum A (Eq. 7) at $\alpha_h = 0.30$: NEE items cluster at low $s_3 \approx 0.04$, Conflicting at moderate $s_3 \approx 0.07$ – 0.13 , and majority classes spread across $\tau_A = 0.109$ (bootstrap 95% CI [0.102, 0.132]).

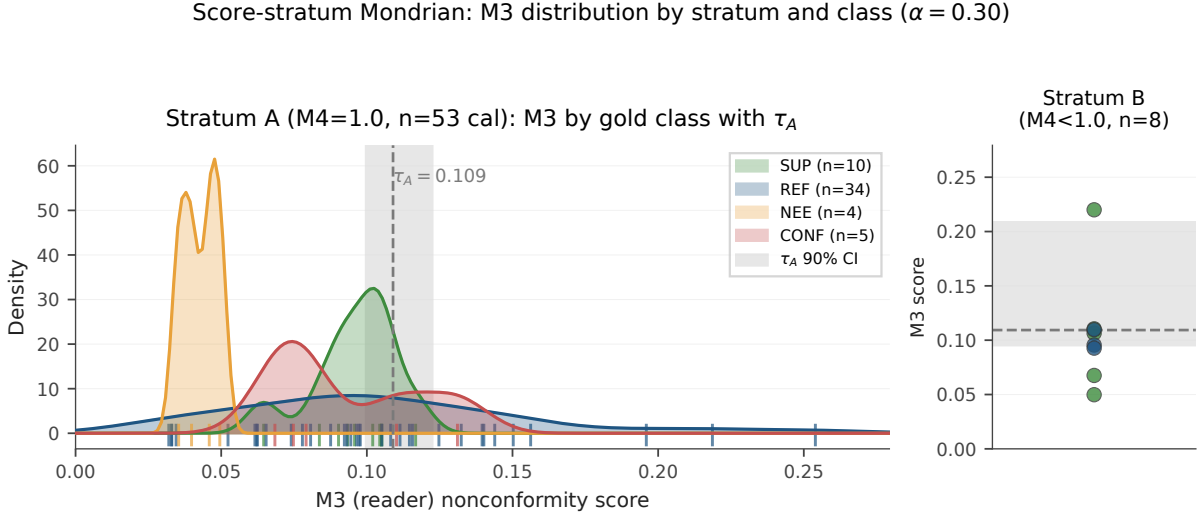


Figure 4: Reader score s_3 for each calibration item in Stratum A ($n=53$), grouped by gold class. Each marker is one calibration item; points are jittered vertically only for readability. The dashed line marks $\tau_A = 0.109$. Items with $s_3 > \tau_A$ are diverted to ESCALATE-MONDRIAN-M3.

Applying the diversion rule (Eq. 8) to the 42 EMIT items diverts 7: 2 wrong Conflicting items (2 of 3 in EMIT, exploiting the high- $M3$ tail) and 5 correctly-classified majority items diverted as the cost. The filter does not recover the 2 correct NEE items already escalated by marginal scoring; those sit in the low- $M3$ region and are unreachable by construction. EMIT shrinks to 35 items, accuracy rises from 0.786 to 0.800, but the Wilson 95% CIs [0.64, 0.88] and [0.64, 0.90] overlap, so the gain is not statistically distinguishable at this sample size.

Cost-aware reframe. We do not present Mondrian-M3 as an accuracy improvement. Instead, we evaluate it under a deployment-cost view. Letting $R = c_{FE}/c_{FX}$ denote the cost ratio of a wrong emit to a false escalate of a correct item, total expected cost is $9R$ for marginal (no false-escalates, 9 wrong emits) and $7R + 5$ for Mondrian-M3 (7 wrong emits, 5 false-escalates). Equating gives a break-even ratio of $R^* = 2.5$: Mondrian-M3 is preferred whenever publishing a wrong verdict costs more than $2.5\times$ asking a human to glance at a correct one (Tab. 4).

Table 4

Cost-weighted comparison at $\alpha_h = 0.30$ ($c_{FX} = 1$, $R = c_{FE}/c_{FX}$; break-even $R^* = 2.5$).

R	Marg.	Mond.	Pref.
1	9	12	Marg.
3	27	26	Mond.
5	45	40	Mond.
10	90	75	Mond.

In fact-checking deployment R is typically well above this threshold, making Mondrian-M3 a cost-favourable routing strategy in spite of the inconclusive accuracy result.

Tab. 5 compares Mondrian-M3 against joint copula and two-sided Mondrian variants on the same EMIT subset; among multivariate recalibrations of the marginal scheme, one-sided Mondrian-M3 is the only variant that diverts more than one minority item without reducing accuracy below baseline.

Table 5

Ablation at $\alpha_i = 0.30$. Min. div. = minority items diverted from EMIT (out of 4). CIs overlap; ordering is indicative.

Method	n (cov.)	Acc	Min. div.
Marginal	42 (69%)	0.786	0
+ Joint copula	34 (56%)	0.765	0
+ Mondrian-M3	35 (57%)	0.800	2
+ Mondrian (2-sided)	31 (51%)	0.774	1

6. Discussion

For **RQ1**, the multi-module conformal framework achieves the highest emit accuracy among the methods tested (0.786 vs. 0.750 retrieval-only and 0.729 reader-only); McNemar’s test returns $p > 0.5$ against both, so the ordering is indicative rather than statistically established at $n = 61$. Two structural benefits remain robust to this caveat: a distribution-free per-module guarantee that transfers across pipelines, and explicit attribution to the firing module thereby facilitating AI system modular accountability. Of the 19 escalated items, 17 overlap with retrieval-only and 6 with reader-only; the conformal escalation set is contained in the union of the two baselines, i.e. the framework does not surface items neither baseline could find but combines both signal types into one calibrated rule.

For **RQ2**, attribution categories isolate genuinely harder items (would-be accuracy 0.33 for ESCALATE-RETRIEVAL and 0.46 for ESCALATE-MULTIPLE). ESCALATE-READER never fires because s_3 is never the strict-tail violator alone on this benchmark; ESCALATE-VERIFIER-CONFLICT does flag items but the framework emits them, since s_4 saturation reflects sample agreement rather than genuine verifier uncertainty when the upstream is clean. These channels would route to human review on pipelines with non-saturated verifier sampling.

For **RQ3**, selective abstention under imbalance reduces class-coverage macro recall from 0.52 to 0.34 because minority items are flagged as unusual regardless of correctness. The Mondrian extension does not statistically improve emit accuracy, but under a cost-aware view it becomes preferred whenever the cost of publishing a wrong verdict exceeds $\sim 2.5\times$ the cost of falsely escalating a correct one (Sec. 5.4); it addresses the high- $M3$ (Conflicting) end of the asymmetry but cannot reach the low- $M3$ (NEE) end by construction.

Limitations. The combined multi-threshold rule does not inherit a closed-form joint coverage guarantee; formal joint coverage for multi-module compositions under dependence remains an open problem. The 61/61 split limits statistical power: minority-class findings ($n = 4$ NEE, $n = 5$ Conflicting in test) are indicative rather than statistically definitive, and the $0.786 \rightarrow 0.800$ Mondrian gain is not distinguishable from sample noise. We report verdict accuracy rather than the joint AVeriTeC score. Scores s_1 and s_2 require gold annotations at calibration and test time; self-supervised replacements (decomposer self-consistency for s_1 , retrieval-confidence gaps for s_2) would make the framework fully deployable. Future work should pursue class-aware scoring that distinguishes “unusual and correct” from “unusual and wrong”: prediction-conditional calibration via a lightweight class estimator, cost-sensitive conformal thresholds, and exogenous evidence-quality scores calibrated as a fifth module independent of the verifier’s APS saturation. Finally, deploying the framework across institutions—where calibration data is distributed and cannot be pooled for privacy reasons—connects selective abstention to federated and

one-shot learning paradigms [19, 20] and to trustworthy distributed learning more broadly [21], an avenue we leave to future work.

7. Conclusion

We presented a selective abstention framework for agentic LLM pipelines that combines per-module conformal calibration with attribution to the upstream module triggering each escalation. On AVeriTeC, the framework achieves the highest emit accuracy among the methods tested while exposing a structural accuracy–recall asymmetry under class imbalance: minority items are flagged as unusual regardless of correctness, reducing macro recall from 0.52 to 0.34. We reframe a Mondrian extension through a cost lens rather than as an accuracy improvement. The asymmetry finding and the cost frame for evaluating abstention extensions represent a contribution for practitioners deploying selective abstention in data imbalanced conditions.

Acknowledgments

This work was partially supported by PNRR MIUR Project PE0000013-FAIR. The FAIR project is committed to promoting an advanced vision of Artificial Intelligence, driving research and development in this crucial field and constantly keeping ethical, legal and sustainability considerations in mind.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) in order to: Drafting content, Improve writing style, Grammar and spelling check. After using these tools/services, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication’s content.

References

- [1] R. El-Yaniv, Y. Wiener, On the foundations of noise-free selective classification, *Journal of Machine Learning Research* 11 (2010) 1605–1641.
- [2] Y. Geifman, R. El-Yaniv, Selective classification for deep neural networks, in: *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [3] V. Vovk, A. Gammerman, G. Shafer, *Algorithmic Learning in a Random World*, Springer, New York, 2005.
- [4] A. N. Angelopoulos, S. Bates, A gentle introduction to conformal prediction and distribution-free uncertainty quantification, *Foundations and Trends in Machine Learning* 16 (2023) 494–591.
- [5] B. Kumar, C. Lu, G. Gupta, A. Palepu, D. Bellamy, R. Raskar, A. Beam, Conformal prediction with large language models for multi-choice question answering, in: *ICML Workshop on Neural Conversational AI*, 2023.
- [6] V. Quach, A. Fisch, T. Schuster, A. Yala, J. H. Sohn, T. S. Jaakkola, R. Barzilay, Conformal language modeling, in: *International Conference on Learning Representations (ICLR)*, 2024.
- [7] F. Ye, M. Yang, J. Pang, L. Wang, D. F. Wong, E. Yilmaz, S. Shi, Z. Tu, Benchmarking LLMs via uncertainty quantification, 2024. [arXiv:2401.12794](https://arxiv.org/abs/2401.12794).
- [8] M. Schlichtkrull, Z. Guo, A. Vlachos, AVeriTeC: A dataset for real-world claim verification with evidence from the web, in: *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.
- [9] V. Vovk, D. Lindsay, I. Nouretdinov, A. Gammerman, *Mondrian Confidence Machine*, Technical Report, Royal Holloway, University of London, 2003.
- [10] Y. Romano, M. Sesia, E. J. Candès, Classification with valid and adaptive coverage, in: *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

- [11] S. Ravfogel, Y. Goldberg, J. Goldberger, Conformal nucleus sampling, in: Findings of the Association for Computational Linguistics (ACL), 2023, pp. 27–34.
- [12] J. Ren, Y. Zhao, T. Vu, P. J. Liu, B. Lakshminarayanan, Self-evaluation improves selective generation in large language models, in: ICML Workshop, 2023.
- [13] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, et al., Language models (mostly) know what they know, 2022. [arXiv:2207.05221](https://arxiv.org/abs/2207.05221).
- [14] A. Kamath, R. Jia, P. Liang, Selective question answering under domain shift, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 5684–5696.
- [15] N. Varshney, S. Mishra, C. Baral, Investigating selective prediction approaches across several tasks in IID, non-IID, and adversarial settings, in: Findings of the Association for Computational Linguistics (ACL), 2022, pp. 1995–2002.
- [16] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and verification, in: Proceedings of NAACL-HLT, 2018.
- [17] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, *Advances in Neural Information Processing Systems (NeurIPS)* (2020).
- [18] M. Sadinle, J. Lei, L. Wasserman, Least ambiguous set-valued classifiers with bounded error levels, *Journal of the American Statistical Association* 114 (2019) 223–234.
- [19] S. A. Huda, M. Canzaniello, D. Annunziata, F. Piccialli, Graph-based federated learning for smart healthcare: A comprehensive survey, *Future Generation Computer Systems* 183 (2026) 108524. URL: <https://www.sciencedirect.com/science/article/pii/S0167739X26001585>. doi:<https://doi.org/10.1016/j.future.2026.108524>.
- [20] F. Amato, L. Qiu, M. Tanveer, S. Cuomo, D. Annunziata, F. Giampaolo, F. Piccialli, Towards one-shot federated learning: Advances, challenges, and future directions, *Neurocomputing* 664 (2026) 132088. URL: <https://www.sciencedirect.com/science/article/pii/S0925231225027602>. doi:<https://doi.org/10.1016/j.neucom.2025.132088>.
- [21] D. Annunziata, P. Qi, K. Li, F. Piccialli, Towards trustworthy iiot: A reinforcement learning framework with federated unlearning, *IEEE Internet of Things Journal* (2026) 1–1. doi:10.1109/JIOT.2026.3697565.