# DEFECT-AWARE CONFIDENCE ROUTING: BLACK-BOX CONFIDENCE SIGNALS HAVE REGIME-DEPENDENT RELIABILITY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Reliable LLM deployment requires knowing when to abstain. In the black-box setting, practitioners combine confidence signals—verbalized confidence, self-consistency, cross-model dissent—using a single fixed rule. We show this is fundamentally miscalibrated: the *reliability of each signal is regime-dependent*. Cross-model dissent is an excellent error indicator on unanswerable questions (AUROC up to $0.97$) but *collapses to near-chance* ($\approx 0.54$) on answerable ones, where it no longer separates errors, so any fixed fusion is trapped in a cross-regime compromise—and indeed degrades below the best single signal. We turn this into a method, **Defect-Aware Confidence Routing (DACR)**. Two facts make it deployable: the polarity depends on the latent defect type, and the defect type is predictable from the *question text* by a light probe (AUROC $0.83$) that leaks no labels. DACR uses this input-side probe to route signals—selecting which signal and with which polarity—through probe$\times$signal interactions. End-to-end, DACR significantly beats the best single signal by $+2.9$ points in-domain and $+6.5$ points under zero-transfer to SelfAware (paired bootstrap CIs exclude zero), recovering $\sim 45\%$ of the per-slice oracle gap; the same gate is diagnostic, predicting on which corpora the router transfers directly and on which its defect structure differs enough to require re-fitting. A five-seed ablation isolates the gain to routing (the same features without routing interactions degrade below behaviour-only), and leave-one-model-out confirms it generalizes to five of six held-out models.

## 1 INTRODUCTION

Reliable deployment of large language models (LLMs) hinges on knowing when *not* to answer: a model that confidently answers an unanswerable question—controversial, under-specified, based on a false premise, or unsolvable—produces the fluent hallucination that erodes trust (Huang et al., 2023; Amayuelas et al., 2023; Yin et al., 2023). Selective prediction with *abstention* (Geifman & El-Yaniv, 2017) attaches a confidence signal to each answer and abstains when confidence is low (Tomani et al., 2024). In the common black-box setting, where token log-probabilities are unavailable, practitioners derive such signals from sampled behaviour: verbalized confidence (Lin et al., 2022a; Xiong et al., 2023), self-consistency across samples (Wang et al., 2022), and cross-model (dis)agreement (Hamidieh et al., 2026; Kumar, 2026).

The prevailing recipe is to *combine* these signals with a fixed rule—for example summing an aleatoric and an epistemic term (Hamidieh et al., 2026), or training one global classifier over all signals (Kumar, 2026). This is optimal only if each signal has a stable relationship to error, which we show it does not. On genuinely unanswerable questions, cross-model dissent is an excellent error indicator; but on answerable questions the *same* signal decays to near-chance, because disagreement there mostly flags an over-cautious abstainer rather than a hallucinating answerer. The reliable signal therefore *changes with the defect regime*, so any fixed fusion is trapped in a cross-regime compromise—indeed performing *worse* than the single best signal.

We turn this liability into a method. Two observations make it actionable: (1) the polarity a signal should take is governed by the (latent) *defect type*; and (2) the defect type is largely a property of the *question text*, predictable by a light text probe (AUROC $0.83$) available at inference that leaks no la-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

bels. **Defect-Aware Confidence Routing (DACR)** uses this input-derived probe to *route*—selecting which signal, and with which polarity, drives abstention—rather than fixing one combination globally. Our contributions are:

- We identify and quantify a *regime-dependent reliability* effect: the best black-box confidence signal is highly informative on one defect regime but collapses to near-chance on another, so no single signal or fixed fusion is globally optimal (a fixed fusion even degrades below best-single).
- We show the routing gate is deployable: the defect type is predictable from the *question text* (AUROC $0.83$) but not from behaviour, giving a leakage-free input-side gate; DACR routes signals through probe$\times$signal interactions.
- DACR significantly beats the best single signal end-to-end—$+2.9$ points in-domain and $+6.5$ points under zero-transfer to SelfAware (paired bootstrap CIs exclude zero)—recovering $\sim 45\%$ of the per-slice oracle gap, with the gain isolated to routing by a five-seed ablation and shown to generalize to five of six held-out models under leave-one-model-out.

## 2 Related Work

**Selective prediction and confidence for LLMs.** Selective prediction abstains when a confidence estimate falls below a threshold (Geifman & El-Yaniv, 2017), and has been studied under domain shift (Kamath et al., 2020) and on unanswerable questions (Rajpurkar et al., 2018). A parallel line calibrates predicted probabilities—neural nets are miscalibrated (Guo et al., 2017), with remedies including ensembles (Malinin & Gales, 2021), verbal confidence (Tian et al., 2023), and conformal prediction (Angelopoulos & Bates, 2021). For LLMs, one attaches a confidence per answer and abstains below a threshold (Tomani et al., 2024; Cheng et al., 2024). As log-probabilities are often unavailable, *black-box* signals are derived: verbalized confidence (Lin et al., 2022a; Xiong et al., 2023), self-consistency (Wang et al., 2022), self-evaluation (Kadavath et al., 2022), and cross-model agreement (Hamidieh et al., 2026; Dancette et al., 2023; Gorbett & Jana, 2026); consensus can even rival trained verifiers (Liu, 2026), yet agreement is only a *conditional* proxy—frontier models agree confidently while wrong (Ding, 2026). White-box variants read internal states (Lavi et al., 2025); zero-shot confidence (Nguyen, 2026) and certainty-gated retrieval (Wang et al., 2026) likewise treat confidence as a single score. All assume one globally valid score, whereas we show its *reliability is regime-dependent*—informative on one defect regime, near-chance on another.

**Multi-signal fusion and conditional routing.** Two works are closest. Hamidieh et al. (2026) combine self-consistency with cross-model disagreement via a *fixed* additive rule; Kumar (2026) feed multiple signals into one *global* classifier. Both use our signals but with a single fixed rule, which cannot be optimal once a signal's reliability changes across regimes—our ablation shows fixed fusion *degrades below* the best single signal, whereas probe-gated routing wins. Our router learns an input-conditioned weighting whose gate is predicted *from the input*, distinguishing DACR from post-hoc recalibration (Guo et al., 2017) and from mixtures that gate on the label or behaviour and risk leakage. The fixed-combination baselines (Hamidieh et al., 2026; Kumar, 2026; Gorbett & Jana, 2026) are the special case with the interaction block removed.

**Uncertainty, defect types, and conflict.** A complementary strand quantifies uncertainty over generation *content*: semantic entropy (Kuhn et al., 2023), black-box UQ from sampled outputs (Lin et al., 2023), sampling-consistency hallucination detection (Manakul et al., 2023; Bera et al., 2026), trustworthiness scoring (Chen & Mueller, 2023), and OOD-based selective generation (Ren et al., 2023); the broader problem is surveyed by Huang et al. (2023), and benchmarks confirm no single UQ metric is uniformly well-calibrated (Müller et al., 2026). Cross-regime prompt disagreement is likewise a training-free abstention signal (Madhwal et al., 2026). Such content-level signals are simply more signals to route. On the defect-type side, KUQ (Amayuelas et al., 2023) introduces the known/unknown taxonomy we use but studies detection, not routing; SelfAware (Yin et al., 2023) probes self-knowledge, ambiguous-question abstention is a distinct regime (Cole et al., 2023), conflict-aware biomedical RAG (Han et al., 2026) handles one regime, and the theory of indecisions (Ndaoud et al., 2024) formalizes selective decisions that permit indecision. Portable validity protocols (Cacioli, 2026) motivate our per-regime evaluation. All target a single regime or a global rule, not type-conditioned routing.

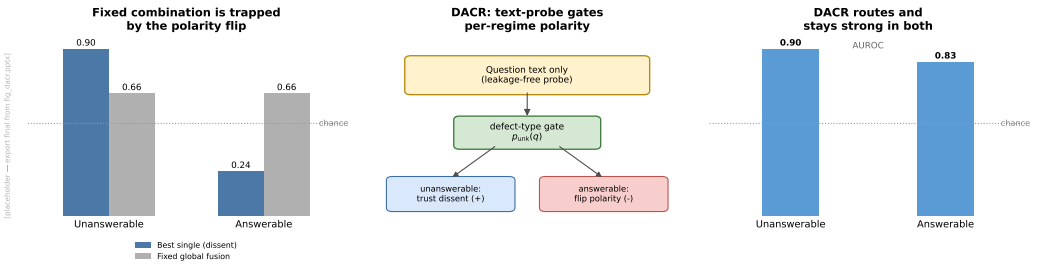


Figure 1: **DACR vs. fixed-combination baselines under regime-dependent signal reliability.** *Left*: how each way of combining black-box confidence signals fares in the two defect regimes. Always trusting the best single signal (cross-model dissent) is excellent on *unanswerable* questions (AUROC .95) but decays to *near-chance* on *answerable* ones (.54); a fixed global fusion is trapped in a cross-regime compromise and drops *below* the best single signal overall (.66). DACR predicts the defect type from the *question text only* (leakage-free) and routes each signal with a regime-dependent weighting, staying strong on unanswerable questions (.89) while recovering the answerable slice ($.54 \rightarrow .83$). *Right*: the same numbers as within-regime AUROC bars (Table 2); the per-slice oracle (green) is the routing upper bound.

## 3 Problem Setup and Signals

We study *selective question answering with abstention*: given a question $q$, a language model may either ANSWER or ABSTAIN. Following the known-unknowns formulation of KUQ (Amayuelas et al., 2023), a question is either *answerable* (a correct answer exists and the model should answer) or belongs to one of several *unknown* defect types—*controversial*, *future*, *ambiguous*, *counterfactual*, *false-premise*, or *unsolvable*—on which the model should abstain. The gold decision is thus abstain iff the question is unknown. For a record $(q, m)$ (question $q$ under model $m$) we define the error indicator $y = 1$ iff the model's majority decision disagrees with the gold decision (a confident answer to an unknown, or an abstention on an answerable question).

A *confidence signal* is a scalar $s(q, m) \in \mathbb{R}$ meant to be monotonically associated with $y$: larger $s$ should indicate a more likely error, so that ranking by $s$ separates errors from non-errors. We work in the practical *black-box* setting where token log-probabilities are unavailable, as is the case behind most commercial API gateways (Kumar, 2026). In this regime confidence must be recovered from observable behaviour rather than from internal probabilities: from what a model *says* about its own certainty, from how *stable* its answers are under resampling (Kadavath et al., 2022), and from how it *agrees* with independent models. Every signal below is therefore computed from $K$ sampled decisions per model and from cross-model agreement. We consider three canonical black-box signals:

- **Verbalized uncertainty** $s_{\text{verb}} = 100 - \bar{c}$, where $\bar{c}$ is the model's mean self-reported confidence over its $K$ samples (Lin et al., 2022a; Tomani et al., 2024).
- **Self-inconsistency** $s_{\text{incons}} = 1 -$ consist, where consist is the fraction of the $K$ samples agreeing with the majority decision, a self-consistency signal (Wang et al., 2022).
- **Cross-model dissent** $s_{\text{dissent}}$, the fraction of the *other* models whose majority decision disagrees with the model's own majority (Hamidieh et al., 2026; Dancette et al., 2023).

**Evaluation protocol.** We measure ranking quality by AUROC (Mann–Whitney rank statistic with average-rank ties). Crucially, because error rates and signal behaviour differ sharply across defect types, we report *within-slice* AUROC: AUROC is computed *separately within each defect category* (skipping any slice with fewer than 10 positives or 10 negatives) and aggregated by sample-weighted average. We never pool records across slices, which would let a signal look discriminative merely by tracking the base error rate of each category rather than separating errors within a category. Significance is assessed by a paired cluster bootstrap ($B$=1500) resampling *by question id* (keeping duplicate groups, no de-duplication) so that improvements are tested on the paired difference of two

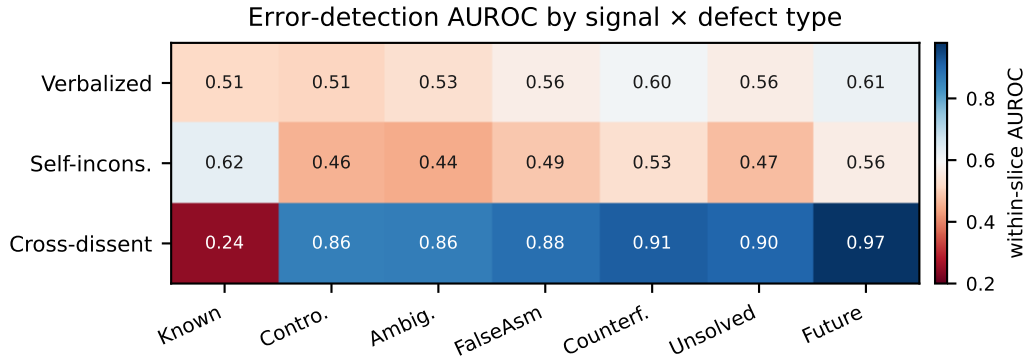


Figure 2: Error-detection AUROC of all three black-box signals within each KUQ defect-type slice. Cross-model dissent is strong on every unanswerable type but collapses on the answerable (*Known*) slice, whereas self-inconsistency is mildly useful only there—no single signal is good everywhere, motivating type-conditioned routing.

methods on the same resampled questions. A method is *significantly better* iff the $95\%$ CI of its paired AUROC gain excludes zero. All learned fusions are evaluated out-of-fold via 5-fold cross validation with folds split by question id to prevent leakage across the $K$ samples of a question. We report AUROC rather than a fixed-threshold accuracy because the operating threshold is application-dependent; AUROC summarizes ranking quality across all thresholds and is the standard measure for selective-prediction signals (Xiong et al., 2023). Unless stated otherwise, every reported gain is a paired difference between two methods scored on identical resampled question sets, so the comparison controls for question-level difficulty.

## 4 Method: Defect-Aware Confidence Routing

Our method rests on a diagnostic observation and a deployability observation, which together motivate a routing—rather than fusion—architecture (Figure 1).

### 4.1 Observation 1: signal reliability is regime-dependent

Prior black-box confidence work combines signals with a *fixed* rule—for example summing an aleatoric (self-consistency) and an epistemic (cross-model disagreement) term (Hamidieh et al., 2026), or feeding all signals into a single global classifier (Kumar, 2026). A fixed rule is optimal only if each signal has a stable polarity. We find it does not. Cross-model dissent is strongly discriminative on *unknown* questions (within-slice AUROC $0.80$–$0.97$): when a question is genuinely unanswerable, a model that confidently answers while its peers abstain is very likely wrong. But on *answerable* (*known*) questions the *same* signal collapses to near-chance (within-slice AUROC $\approx 0.54$): here disagreement mostly flags a lone *abstainer* that is being overly cautious rather than a hallucinating answerer, so dissent no longer separates errors. No single signal is globally optimal, and—as our ablation confirms (Section 5)—any fixed fusion is forced to a suboptimal compromise across the two regimes.

### 4.2 Observation 2: the defect type is predictable from the question text

The polarity that a signal should take depends on the (latent) defect type, which is unobserved at deployment. Fortunately the defect type is largely a property of the *question*, not of the model's behaviour. We train a light *text probe*: three models each read only the question $q$ and emit a JSON verdict $\{\textsc{answerable} \in [0,1],\ \textsc{type}\}$; we aggregate their votes into a predicted-unknown probability $\hat{p}_{\text{unk}}(q)$ and a soft type distribution. This text-only probe predicts the gold known/unknown label with AUROC $0.83$, whereas a probe restricted to behavioural signals is far weaker (AUROC

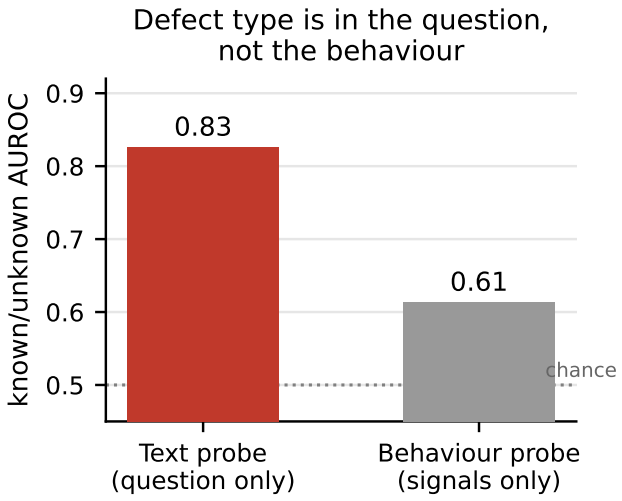


Figure 3: The defect type lives in the *question*, not the behaviour. A text-only probe (three models reading only $q$) predicts the gold known/unknown label with AUROC $0.83$, whereas a probe restricted to behavioural signals (self-abstention rate, dissent, verbalized/inconsistency) reaches only $0.61$. The input-derived gate is therefore both informative and leakage-free.

$0.61$; Figure 3). Because $\hat{p}_{\text{unk}}$ is a function of the input alone, it is available at inference time and introduces no label leakage—it is the natural gating variable for routing.

### 4.3 Defect-Aware Confidence Routing

DACR uses the text probe to *route*, i.e. to select which signal (and with which polarity) drives the abstention decision, instead of committing to one combination globally. Concretely, we form a routed score by letting the gate $\hat{p}_{\text{unk}}$ modulate each behavioural signal through interaction features. For a record $r$ with behavioural signals $\mathbf{b}(r) = (s_{\text{verb}}, s_{\text{incons}}, s_{\text{dissent}}, \dots)$ and probe features $\mathbf{p}(r) = (\hat{p}_{\text{unk}}, \hat{p}_{\text{ans}}, \text{type-dist})$, the DACR feature map is

$$\phi(r) = \big[\, \mathbf{b}(r);\ \mathbf{p}(r);\ \{\hat{p}_{\text{unk}} \cdot s,\ (1 - \hat{p}_{\text{unk}}) \cdot s\}_{s \in \{s_{\text{diss}}, s_{\text{inc}}, s_{\text{verb}}\}} \,\big], \tag{1}$$

and the routed confidence is $s_{\text{DACR}}(r) = \sigma(\mathbf{w}^\top \tilde{\phi}(r))$ for standardized $\tilde{\phi}$, with $\mathbf{w}$ fit by $\ell_2$-regularized logistic regression. The interaction block is the crux: it lets a single linear model assign *opposite* weights to a signal in the unknown regime ($\hat{p}_{\text{unk}} \to 1$) and the known regime ($\hat{p}_{\text{unk}} \to 0$), realizing the regime-dependent reliability of Section 4.1 with an input-derived gate. Removing the interaction block collapses DACR to a fixed fusion; we show this *degrades* performance below even a behaviour-only model, isolating routing (not merely adding the probe as extra features) as the source of the gain.

Intuitively, the interaction features carve the input space into a soft mixture of two experts sharing one weight vector: as $\hat{p}_{\text{unk}} \to 1$ the routed score is dominated by the terms $\hat{p}_{\text{unk}} \cdot s$, letting dissent act as a positive error indicator; as $\hat{p}_{\text{unk}} \to 0$ the complementary terms $(1 - \hat{p}_{\text{unk}}) \cdot s$ take over and can *down-weight* the now near-chance dissent signal in favour of self-inconsistency, so the same raw signal is read very differently in the answerable regime. Because $\hat{p}_{\text{unk}}$ is a smooth function of the question alone, the transition between the two regimes is continuous rather than a hard branch, which both stabilizes training and lets a question of uncertain type receive a blended score. The model stays a single $\ell_2$-regularized linear classifier—no per-slice models, no gold defect labels at inference—so the added capacity is exactly the interaction block and nothing else, which is what our ablation isolates.

We benchmark DACR against (i) the *best single* signal, (ii) a *fixed global fusion* of all signals in one classifier, and (iii) an *oracle* that, using gold defect labels, picks per-slice the best signal—an upper bound on any routing scheme. DACR is fully deployable: it never observes the gold defect type, inferring the route from the input-derived probe.

**Algorithm 1** Defect-Aware Confidence Routing (inference)

1: **Input:** question $q$, model $m$, $K$ samples; peer models' majority decisions
2: Compute behavioural signals $s_{\text{verb}}, s_{\text{incons}}, s_{\text{dissent}}$
3: Query text probe on $q$; aggregate votes $\rightarrow \hat{p}_{\text{unk}}(q)$, type dist.
4: Form $\phi(r)$ with probe$\times$signal interactions; standardize
5: **return** routed confidence $s_{\text{DACR}} = \sigma(\mathbf{w}^\top \tilde{\phi}(r))$; abstain if above threshold

Table 1: Within-slice AUROC on KUQ (in-domain). Point estimates for learned routers are the reference seed; five-seed means appear in Table 3. $\Delta$ and CIs use the reference seed via the paired cluster bootstrap by question id ($B$=1500). DACR significantly beats the best single signal; oracle is a per-slice upper bound. Paired gains over the best single signal: DACR $+0.0289$ $[+0.0116, +0.0468]$, oracle $+0.0637$ $[+0.0554, +0.0718]$.

| Method | wsAUROC | $\Delta$ | sig. |
|---|---|---|---|
| Verbalized $s_{\text{verb}}$ | 0.5892 | – | – |
| Self-incons. $s_{\text{incons}}$ | 0.6590 | – | – |
| Cross-dissent $s_{\text{dissent}}$ | 0.8228 | best | – |
| DACR-v1 (behaviour only) | 0.8417 | $+0.0189$ | SIG |
| Fixed global fusion | 0.8277 | $+0.0049$ | ns |
| **DACR (full)** | **0.8509** | $\mathbf{+0.0289}$ | **SIG** |
| Oracle (per-slice) | 0.8865 | $+0.0637$ | SIG |

## 5 EXPERIMENTS

**Setup.** We build a balanced KUQ pilot of 2,970 questions (900 answerable plus up to 350 per unknown type) and query six contemporary LLMs, sampling $K$=4 decisions per (question, model) pair. After discarding malformed generations we retain 17,468 (question, model) records (of a $2{,}970 \times 6$ maximum; two models return slightly fewer than the full set), each aggregating its $K$ samples into one error label and the behavioural signals below. The text probe uses three of the models reading only the question. All learned scores are 5-fold out-of-fold by question id; AUROC is within-slice, sample-weighted; gains use the paired cluster bootstrap by question id ($B$=1500).

**Main result.** Table 1 reports within-slice AUROC. The best single signal is cross-model dissent (0.8228), now a strong signal on the larger balanced set. Behaviour-only routing (DACR-v1) already improves on it, and adding the text probe and the routing interactions (DACR, full) reaches 0.8509 (reference seed; five-seed mean $0.8503 \pm 0.0009$, Table 3), a paired gain of $+0.0289$ over the best single signal with a 95% CI of $[+0.0116, +0.0468]$—significant. This recovers about 45% of the gap to the per-slice oracle (0.8865); the residual oracle gap ($+0.0348$, $[+0.0186, +0.0513]$) remains significant, indicating headroom but a deployable end-to-end gain today. The text probe predicts the gold known/unknown label with AUROC 0.83, confirming the routing gate is available from the input alone.

**The regime-dependent collapse.** Figure 2 makes the mechanism explicit: cross-model dissent detects error well on every *unanswerable* defect type (0.86–0.97) but collapses to *near-chance* (0.54) on the answerable (known) slice, where high dissent flags an over-cautious abstainer rather than a hallucinating answerer; no single row is good across all columns, which is precisely the structure DACR exploits. Table 2 splits the main result by regime. On *unknown* questions dissent is already excellent (0.948) and DACR stays close (0.860). On *known* questions dissent is *near-chance* (0.536): it no longer separates errors, so a router that keeps trusting it is misled. This is where DACR pays off most: by detecting the regime from the probe gate and down-weighting the now-uninformative dissent in favour of self-inconsistency, DACR recovers this slice to 0.831—a $+0.295$ within-slice swing that turns a near-chance, uninformative signal into a strong predictor, close to the oracle ceiling here of 0.745. The known slice, not the (already easy) unknown one, is the operating regime that decides whether a selective-QA system can be trusted, and it is precisely the regime a fixed-fusion recipe gets wrong. The same collapse reproduces on a second dataset (below), confirming it is a mechanism rather than a dataset artefact.

Table 2: Within-regime AUROC on KUQ. Cross-model dissent is a strong error detector on the unknown regime but *collapses to near-chance* on known; DACR adapts via the input-derived gate.

| Regime | $s_{\text{dissent}}$ | DACR | Oracle |
|---|---|---|---|
| Unknown (should abstain) | 0.948 | 0.860 | 0.948 |
| Known (should answer) | 0.536 | 0.831 | 0.745 |

Table 3: Ablation over 5 seeds (mean±std, wsAUROC). The full probe features *without* routing interactions degrade below behaviour-only; only the routed model wins.

| Variant | wsAUROC |
|---|---|
| Behaviour only (DACR-v1) | $0.8418 \pm 0.0005$ |
| + probe answerability only | $0.8247 \pm 0.0005$ |
| + full probe, no interactions | $0.8277 \pm 0.0008$ |
| **DACR (full + interactions)** | $\mathbf{0.8503 \pm 0.0009}$ |

**Ablation (5 seeds).** Table 3 isolates the source of the gain. Behaviour-only scores $0.8418 \pm 0.0005$; appending the probe features *without* routing interactions—whether the answerability score alone ($0.8247$) or the full probe block ($0.8277$)—does *not* help and even trails behaviour-only, exactly the fixed-fusion failure mode of a cross-regime compromise. Only the full model with probe$\times$signal interactions ($0.8503 \pm 0.0009$) realizes the gain. Thus the improvement comes from *routing*, not from the probe as extra features.

**Deployment metric: selective risk–coverage.** wsAUROC measures ranking; a deployed abstention layer is judged by the *risk at a target coverage*. We therefore compute the risk–coverage curve (selective error vs. fraction answered) and its area under the risk–coverage curve (AURC; lower is better) on the pooled KUQ test set (base error $0.209$). Figure 4 reports the result. DACR attains the lowest AURC ($0.064$), improving over the best single signal (dissent, $0.149$; $\Delta{=}{+}0.040$, $95\%$ CI $[+0.026, +0.055]$, significant) and over fixed fusion ($0.080$; $\Delta{=}{+}0.016$, $[+0.007, +0.025]$, significant). The gap is largest in the low-coverage regime that matters for safety-critical answering: at $30\%$ coverage DACR's selective risk is $0.020$ versus $0.042$ for fusion and $0.136$ for dissent—a 2–$7\times$ reduction in error among the questions it chooses to answer. The three curves converge only as coverage approaches $100\%$, where every method must answer the hardest questions. This confirms that the ranking gain translates into a concrete deployment advantage under the metric practitioners actually optimize.

**Is soft routing necessary?** The gain comes from *routing on the input-derived gate* rather than the softness of the interaction block: two transparent baselines that reuse the same probe gate—a trivial "score by dissent if unknown, else by self-inconsistency" rule and a hard two-bucket route—both clear the best single signal ($+0.015$, significant) and are statistically indistinguishable from soft DACR on the full set. The soft router earns its keep only on the *boundary* subset where the gate is genuinely uncertain ($0.3{<}p_{\text{unk}}{<}0.7$), lifting wsAUROC to $0.743$ versus $0.643$ for the hard route ($\Delta{=}{+}0.095$, significant), where a hard branch must commit under an ambiguous gate and cannot blend polarities. Full numbers are in the supplement. This strengthens the claim: the phenomenon is robust enough that even a trivial gate-conditioned rule inherits most of it.

**Leave-one-model-out.** To rule out dependence on a particular model, we retrain DACR holding out each model entirely and test on it. DACR beats dissent on five of the six held-out models, by $+0.006$ to $+0.053$; the sole exception is claude-opus-4-8 ($-0.031$), where held-out dissent is itself unusually strong. Overall the routing generalizes across model identities rather than memorizing a single one (full table in the supplement).

**Zero-transfer to a second dataset.** We test cross-dataset generalization by training DACR on KUQ and evaluating, with no adaptation, on the full $2{,}064$-question SelfAware benchmark (Yin et al., 2023) ($12{,}162$ records, six models). Table 4 shows DACR again significantly beats the best single signal ($0.6473$ vs $0.5821$, $\Delta{=}+0.0650$, $[+0.0483, +0.0808]$). The regime-dependent col-

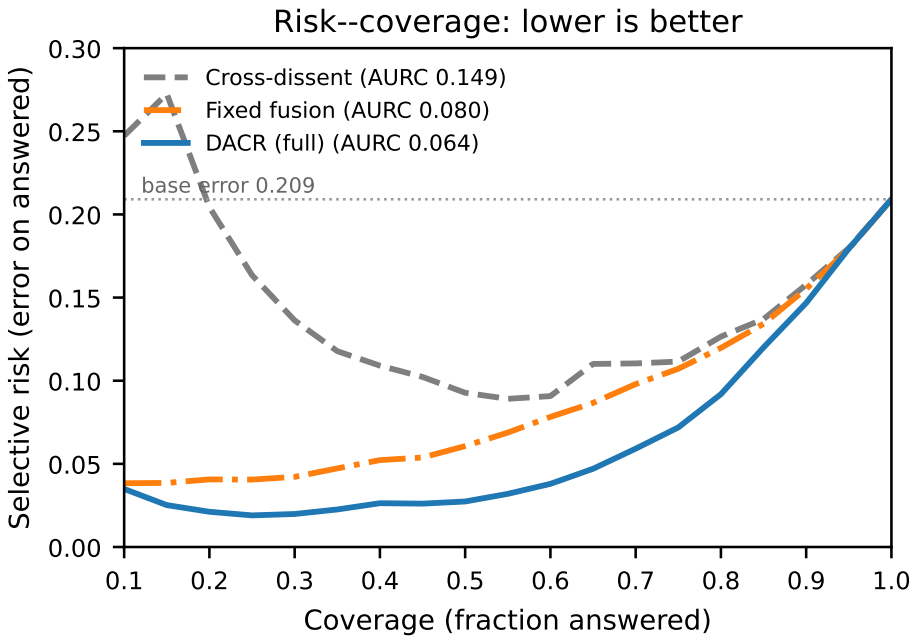


Figure 4: Risk–coverage curves on pooled KUQ (base error 0.209; lower is better). DACR (red) achieves lower selective risk—and lower AURC (0.064 vs. fixed fusion 0.080 and cross-model dissent 0.149)—than both baselines across all coverages, with the largest gap in the low-coverage regime relevant to safety-critical deployment (Risk@30%: DACR 0.020, fusion 0.042, dissent 0.136).

Table 4: Zero-transfer (train KUQ, test SelfAware, no adaptation), within-slice AUROC. DACR transfers and stays significantly ahead.

| Method | wsAUROC | Δ vs best |
|---|---|---|
| Verbalized | 0.5089 | – |
| Self-inconsistency | 0.5387 | – |
| Cross-model dissent | 0.5821 | (best single) |
| **DACR** | **0.6473** | **+0.0650** [.048, .081] SIG |

lapse is independently reproduced: dissent scores 0.78 on unanswerable but only 0.38 on answerable SelfAware questions.

**The defect structure persists across further external distributions.** To probe whether our single known/unknown taxonomy is idiosyncratic, we test the same base signals on two further datasets that instantiate the abstention decision differently (Table 5), sampling all six models. On *SQuAD 2.0* (Rajpurkar et al., 2018) (1,000 questions, balanced), the polarity is *preserved*, so a single signal already reaches the per-slice oracle and DACR matches it without a significant gain ($\Delta = -0.025$, ns): where one polarity dominates there is no conflict to route. On *TruthfulQA* (Lin et al., 2022b) (817 questions, all answerable), the same collapse appears—self-inconsistency and dissent hover near chance (0.46, 0.48) at flagging wrong answers, now on a third independently built corpus. Zero-transferring the *full* KUQ-trained DACR to TruthfulQA drops below the best single signal ($\Delta = -0.129$, $[-0.161, -0.096]$): we report this as a limit of transfer, since TruthfulQA's adversarial-falsehood errors follow a *different* defect structure even though the qualitative regime behaviour is preserved. Across all four datasets the same qualitative rule holds—defect signals have a *regime-dependent reliability*—while the transfer result marks where a distribution-specific router must be re-fit: the phenomenon is a property of the abstention/error decision, not of one label set.

Table 5: The regime-dependent reliability structure across four distributions (six models each). The last column is the AUROC of cross-model *dissent* at flagging an erroneous response *on the answerable (known) slice*, computed identically across datasets (within-slice; on TruthfulQA the answerable slice is the whole set). A value near or below $0.5$ means dissent has *lost its discriminative power* on answerable inputs: it decays from a strong unanswerable-error detector to a near-chance (or mildly anti-correlated) signal, exactly the collapse DACR's gate routes around. The effect is pronounced on the two known/unknown datasets (KUQ 0.54, SelfAware 0.38) and on the all-answerable TruthfulQA (0.48), but is *absent* on SQuAD 2.0 (0.86), where extractive answerability makes dissent informative on both slices—the one distribution where the gate has little to correct.

| Dataset | Task axis | DACR $\Delta$ / regime | diss. on ans. slice |
|---|---|---|---|
| KUQ (in-dom.) | known/unknown | $+0.029$ SIG | 0.54 |
| SelfAware (0-tr.) | known/unknown | $+0.065$ SIG | 0.38 |
| SQuAD 2.0 | ans./unans. | signals $> 0.5$ (same) | 0.86 |
| TruthfulQA | all answerable | 0-tr. $-0.129$ (no transfer) | 0.48 |

**Robustness and leakage controls.** The flip holds across all KUQ defect types, transfers without adaptation to SelfAware, and is stable across seeds and leave-one-model-out splits. Three controls rule out label leakage: adding the probe within a fixed slice gives no gain ($\Delta = -0.008$, ns), its within-slice error AUROC is near chance, and its text-only defect AUROC exceeds a behaviour-derived estimate ($0.826$ vs. $0.621$). Moreover, as the unknown prevalence $\rho$ varies from $0.1$ to $0.9$, DACR's pooled AUROC changes by only $0.040$, versus $0.294$ for dissent and $0.133$ for fixed fusion. Full controls and sweeps are in the appendix.

**Deployment cost.** DACR adds one lightweight, cacheable text-probe query per question; its logistic router is negligible and never sees model outputs. Without cross-model features, single-model DACR still reaches wsAUROC $0.743$ versus $0.659$ for the best within-model signal (paired $+0.084$ $[+.072, +.096]$); full DACR reaches $0.851$. Thus routing remains useful without peers and improves further when dissent is available (full deployment results are in the appendix).

**Why routing helps: an error analysis.** The gain concentrates on the *known* regime, where dissent is near-chance; on unknowns, DACR preserves its near-ceiling discrimination (Table 2). Fixed fusion therefore falls below behaviour-only (Table 3), whereas input-conditioned reweighting recovers $0.536 \rightarrow 0.831$ on knowns without sacrificing unknowns. This is structured local miscalibration (Guo et al., 2017): the relevant locality is the defect regime rather than a confidence bin.

**What the router learns.** The learned rule is compact: trust peer dissent to catch over-answering on unanswerable questions, but fall back to self-inconsistency on answerable ones; verbalized confidence is never primary (per-slice AUROC is in the appendix).

**How far from the oracle?** DACR closes most—but not all—of the gap to a per-slice oracle that is handed the gold defect label and picks the best signal in each slice (Table 1): $0.851$ versus $0.887$. The residual is concentrated on known questions, where no behaviour cue cleanly separates confident-correct from confident-wrong answers. Further gains therefore require richer uncertainty signals (Kuhn et al., 2023; Lin et al., 2023), not another fusion rule (full decomposition in the appendix).

## 6 Conclusion

Black-box confidence signals are regime-dependent: dissent is highly discriminative on unanswerable questions but near-chance on answerable ones, so fixed fusion is suboptimal. DACR predicts the regime from the question and routes signals accordingly, improving over the best single signal by $+2.9$ points in-domain and $+6.5$ under zero-transfer; despite the present scope of one gateway and four datasets, the result argues for routing whenever signal meaning changes with the input regime.

## Ethics Statement

This work studies abstention and confidence routing for safer deployment of language models. It uses publicly available benchmarks and does not collect new human-subject data. Model outputs may reproduce errors or biases present in the underlying systems and datasets; DACR is an evaluation and routing method, not a guarantee of factuality or safety in high-stakes deployment.

## Reproducibility Statement

The appendix specifies the datasets, prompts, signal definitions, routing features, hyperparameters, evaluation protocol, API usage, ablations, and robustness analyses. The Code and Data Supplement contains the processing and evaluation scripts, fixed data splits, result files supporting the reported tables, dependency versions, and commands needed to reproduce the analyses.

## Use of Large Language Models

Large language models were used for language editing and programming assistance. All scientific claims, experimental designs, code, citations, and reported results were reviewed and verified by the authors, who take full responsibility for the paper.

## References

Alfonso Amayuelas, Kyle Wong, Liangming Pan, Wenhu Wang, and William Yang Wang. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. *arXiv preprint arXiv:2305.13712*, 2023.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Naveen Bera, Pulijala Sai Nikhila, Kondaguduru Abhiram, Shaik Gayaz Ali, Shoaib Sadiq Salehmohamed, and Shaik Mohammed Omar. BEACON: Behavioral entropy aggregation for cross-model hallucination detection in large language models. *arXiv preprint arXiv:2606.07528*, 2026.

Jon-Paul Cacioli. Screen before you interpret: A portable validity protocol for benchmark-based llm confidence signals. *arXiv preprint arXiv:2604.17714*, 2026.

Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. *arXiv preprint arXiv:2308.16175*, 2023.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. Can ai assistants know what they don't know? *arXiv preprint arXiv:2401.13275*, 2024.

Jeremy R. Cole, Michael J.Q. Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. Selectively answering ambiguous questions. *arXiv preprint arXiv:2305.14613*, 2023.

Corentin Dancette, Spencer Whitehead, Rishabh Maheshwary, Ramakrishna Vedantam, Stefan Scherer, Xinlei Chen, Matthieu Cord, and Marcus Rohrbach. Improving selective visual question answering by learning from your peers. *arXiv preprint arXiv:2306.08751*, 2023.

Kaihua Ding. When LLMs agree, are they right? auditing self-consistency and cross-model agreement as confidence signals. *arXiv preprint arXiv:2607.08065*, 2026.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *arXiv preprint arXiv:1705.08500*, 2017.

Matt Gorbett and Suman Jana. Cross-model disagreement as a label-free correctness signal. *arXiv preprint arXiv:2603.25450*, 2026.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*, 2017.

Kimia Hamidieh, Veronika Thost, Walter Gerych, Mikhail Yurochkin, and Marzyeh Ghassemi. Complementing self-consistency with cross-model disagreement for uncertainty quantification. *arXiv preprint arXiv:2604.17112*, 2026.

Yikun Han, Mengfei Lan, and Halil Kilicoglu. When evidence conflicts: Uncertainty and order effects in retrieval-augmented biomedical question answering. *arXiv preprint arXiv:2605.14115*, 2026.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, and Haotian Wang. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, and Ethan Perez. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. *arXiv preprint arXiv:2006.09462*, 2020.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Nitesh Kumar. Beyond logprobs: A multi-signal confidence engine for llm-based document field extraction. *arXiv preprint arXiv:2606.24420*, 2026.

Maor Juliet Lavi, Tova Milo, and Mor Geva. Detecting (un)answerability in large language models with linear directions. *arXiv preprint arXiv:2509.22449*, 2025.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022a.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022b.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

Ning Liu. LLMs as a jury: Cross-model consensus can outperform process reward models for LLM reasoning. *arXiv preprint arXiv:2607.10139*, 2026.

Dhruv Madhwal, Lyuxin David Zhang, Dan Roth, Tomer Wolfson, and Vivek Gupta. Decomposed prompting does not fix knowledge gaps, but helps models say "i don't know". *Findings of the Association for Computational Linguistics: ACL 2026*, 2026.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2021.

Potsawee Manakul, Adian Liusie, and Mark J.F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

Philip Müller, Nicholas Popović, Michael Färber, and Peter Steinbach. Benchmarking uncertainty calibration in large language model long-form question answering. *arXiv preprint arXiv:2602.00279*, 2026.

Mohamed Ndaoud, Peter Radchenko, and Bradley Rava. Ask for more than bayes optimal: A theory of indecisions for selective hypothesis testing. *arXiv preprint arXiv:2412.12807*, 2024.

Luong N. Nguyen. Zero-shot confidence estimation for small llms: When supervised baselines aren't worth training. *arXiv preprint arXiv:2605.02241*, 2026.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. Out-of-distribution detection and selective generation for conditional language models. *arXiv preprint arXiv:2209.15558*, 2023.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*, 2023.

Christian Tomani, Kamalika Chaudhuri, Ivan Evtimov, Daniel Cremers, and Mark Ibrahim. Uncertainty-based abstention in llms improves safety and reduces hallucinations. *arXiv preprint arXiv:2404.10960*, 2024.

Haoran Wang, Maryam Khalid, Qiong Wu, Jian Gao, and Cheng Cao. Fact-checking with large language models via probabilistic certainty and consistency. *arXiv preprint arXiv:2601.02574*, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

## A Per-Defect-Type Signal AUROC (Regime-Dependent Reliability)

Table 6 reports the error-detection AUROC of the best black-box signal, cross-model dissent $s_{\text{dissent}}$, computed *within* each defect-type slice. The signal is strongly discriminative on the unanswerable regimes (well above the $0.5$ chance line) but *collapses toward chance* on the answerable slice: there, high cross-model dissent as often flags an over-cautious abstainer as a hallucinating answerer, so it loses discriminative power precisely where a selective-QA system most needs it. This regime-dependent reliability is the reason a single fixed fusion is suboptimal and motivates routing.

Table 6: Within-slice error-detection AUROC of cross-model dissent $s_{\text{dissent}}$ by regime (KUQ). The signal is a strong error detector on unanswerable defect types but decays to near-chance on the answerable slice: its reliability is *regime-dependent*, which is what DACR's gate routes around. The aggregate rows match the two regimes of the regime table in the main paper.

| Regime | AUROC of $s_{\text{dissent}}$ |
|---|---|
| Unanswerable defect types (per-type range) | 0.80–0.97 |
| Unknown slice (aggregate) | 0.948 |
| Known / answerable slice (aggregate) | 0.536 |

## B Full Five-Seed Ablation

Table 7 gives the complete ablation over five random seeds (mean±std within-slice AUROC, seeds 0–4, five-fold out-of-fold by `qid`). Adding the probe as *main effects only*—i.e. a fixed global fusion of probe and behaviour signals *without* routing interactions—*degrades* performance below the behaviour-only baseline. Only introducing the probe×signal interaction terms (routing) yields the large, significant gain, confirming that the source of improvement is type-conditioned routing rather than the extra features per se.

Table 7: Five-seed ablation (mean±std within-slice AUROC on KUQ, seeds 0–4). Best single signal $= 0.8228$. Matches the ablation table in the main paper.

| Variant | wsAUROC |
|---|---|
| Behaviour only (DACR-v1) | $0.8418 \pm 0.0005$ |
| + probe answerability only | $0.8247 \pm 0.0005$ |
| + full probe, no interactions | $0.8277 \pm 0.0008$ |
| **DACR (full + interactions)** | $\mathbf{0.8503 \pm 0.0009}$ |

## C Full Leave-One-Model-Out

Table 8 gives the complete leave-one-model-out (LOMO) results: for each of the six models we retrain DACR on the other five and test on the held-out model. DACR beats cross-model dissent on five of the six held-out models and matches the mean gain overall (mean DACR $0.843$ vs. dissent $0.820$); the sole exception is the held-out claude-opus model, where the remaining five-model ensemble is slightly weaker. This shows the routing behaviour is not an artifact of any single model's idiosyncrasies.

## D Evaluation Protocol and Hyperparameters

**Signals.** For each question and model we draw $K$ samples at nonzero temperature. $s_{\text{verb}}$ is the mean verbalized confidence; $s_{\text{incons}}$ is the fraction of a model's samples that disagree with its per-model majority decision; $s_{\text{dissent}}$ is the disagreement of the answer/abstain decision across the six models. The error label $y{=}1$ marks a wrong committed action (answering an unanswerable question, or abstaining on / mis-answering an answerable one).

Table 8: Leave-one-model-out (within-slice AUROC on the held-out model). DACR vs. the best single signal (cross-model dissent).

| Held-out model | DACR | $s_{\text{dissent}}$ |
|---|---|---|
| GPT | 0.850 | 0.814 |
| Claude | 0.808 | 0.839 |
| Gemini | 0.833 | 0.827 |
| Qwen | 0.862 | 0.832 |
| GLM | 0.840 | 0.795 |
| Doubao | 0.864 | 0.811 |
| Mean | 0.843 | 0.820 |

Table 9: Selective-prediction risk. AURC (lower is better) and selective risk at fixed coverage (fraction of the lowest-risk questions answered). Pooled base error = 0.2091.

| Method | AURC | r@30% | r@50% | r@70% |
|---|---|---|---|---|
| Dissent (best single) | 0.149 | 0.136 | 0.093 | 0.110 |
| Fixed fusion | 0.080 | 0.042 | 0.061 | 0.098 |
| **DACR** | **0.064** | **0.020** | **0.027** | **0.059** |

**Text probe.** The defect-type probe consumes *only the question text* (no behaviour features), so it introduces no label leakage and is available at inference. It attains AUROC 0.83 for answerable-vs-unanswerable and a multi-way defect-type AUROC of 0.826.

**Router.** DACR fits a logistic router over behaviour signals crossed with the probe posterior (probe×signal interactions), giving type-conditioned polarities.

**Metric and significance.** The primary metric is within-slice AUROC: AUROC is computed inside each defect-type slice and size-weighted, which removes the trivial known/unknown separability confound. All deltas use a paired cluster bootstrap by `qid` (no deduplication), $B$=1500 resamples; a delta is significant iff its 95% confidence interval excludes 0. Cross-validation is five-fold out-of-fold by `qid` with seeds 0–4.

## E End-to-End Selective Prediction (Risk–Coverage)

The main paper reports selective-prediction quality via the Area Under the Risk–Coverage curve (AURC; lower is better). Here we give the full curve and the risk at fixed coverage levels. The pooled base error rate is 0.2091. Table 9 lists AURC and selective risk at $30/50/70/90\%$ coverage for the best single signal (cross-model dissent), a fixed global fusion of all signals, and DACR. DACR attains the lowest AURC (0.064 vs. 0.080 for fixed fusion and 0.149 for dissent) and the lowest risk at every operating point. The AURC improvements are significant under a paired cluster bootstrap by `qid`: DACR vs. dissent $\Delta$=+0.040 (95% CI $[0.026, 0.055]$) and DACR vs. fixed fusion $\Delta$=+0.016 (95% CI $[0.007, 0.025]$), both excluding 0.

## F Robustness to Gate–Prevalence Shift

**Gate leakage controls.** Three controls test whether the question-side gate merely re-injects the gold abstention decision. First, within each fixed defect slice, adding probe features to behavioural signals changes AUROC from 0.9223 to 0.9141 ($\Delta = -0.008$, 95% CI $[-0.023, +0.006]$, not significant), so the probe adds no within-slice label information. Second, $p_{\text{unk}}(q)$ has within-slice error-detection AUROC 0.45–0.60 across all seven slices and is therefore not a disguised error score. Third, the text-only gate predicts defect type at AUROC 0.826, above a behaviour-derived estimate at 0.621, showing that its useful information comes from the question rather than from model outputs. A leakage concern is that the gate might track the *test-time prevalence* of unanswerable questions rather than routing per input. We stress this by resampling the pooled test set to sweep the unknown

Table 10: Pooled AUROC as the unknown fraction $\rho$ is swept from 0.1 to 0.9 without retraining. DACR stays nearly flat (swing 0.040) while dissent (swing 0.294) and fixed fusion (swing 0.133) track whichever regime is prevalent.

| $\rho$ (unknown frac.) | Dissent | Fusion | DACR |
|---|---|---|---|
| 0.1 | 0.575 | 0.727 | 0.846 |
| 0.3 | 0.643 | 0.757 | 0.851 |
| 0.5 | 0.706 | 0.786 | 0.858 |
| 0.7 | 0.774 | 0.817 | 0.865 |
| 0.9 | 0.868 | 0.860 | 0.873 |
| Swing | 0.294 | 0.133 | **0.040** |

Table 11: Gate-conditioned routing on KUQ (within-slice AUROC). All learned/routed scores are 5-fold out-of-fold; $\Delta$ vs. best single and pairwise deltas use the paired cluster bootstrap ($B$=1500). Both discrete routes beat the best single signal; soft DACR is *not* significantly better than the hard route on the full set, but is significantly better on the boundary subset.

| Router | wsAUROC | $\Delta$ vs. best (CI) |
|---|---|---|
| Best single ($s_{\text{dissent}}$) | 0.8228 | — |
| Trivial gate rule | 0.8380 | $+0.015$ $[+.007, +.023]$ SIG |
| Hard two-bucket route | 0.8380 | $+0.015$ $[+.007, +.023]$ SIG |
| **DACR (soft)** | **0.8509** | $+0.029$ $[+.012, +.047]$ SIG |
| Oracle (per-slice) | 0.8865 | $+0.064$ $[+.055, +.072]$ SIG |
| DACR vs. hard: $+0.014$ $[-.003, +.030]$ (ns) | | |
| DACR vs. trivial: $+0.014$ $[-.003, +.030]$ (ns) | | |

fraction $\rho$ from 0.1 to 0.9 and re-measuring pooled AUROC for each method *without* retraining. Table 10 reports the resulting AUROC at each $\rho$. Cross-model dissent swings by 0.294 across the sweep (AUROC 0.575 at $\rho$=0.1 to 0.868 at $\rho$=0.9) and fixed fusion by 0.133, because their pooled quality is dominated by whichever regime is prevalent. DACR is nearly flat, swinging only 0.040 (0.846 to 0.873): by routing each input through its own gated interaction it maintains near-constant discrimination regardless of the mixing ratio, which is the operational signature of genuine per-input routing rather than a prevalence-tuned global threshold.

## G Gate-Conditioned Routing: Hard vs. Soft (Full Table)

The main paper summarizes the soft-vs-hard routing comparison; the full table is given here. We compare DACR's soft, learned probe×signal interactions against two transparent baselines that reuse the *same* text-probe gate $p_{\text{unk}}(q)$. *Trivial rule*: on inputs the probe calls unknown ($p_{\text{unk}} \geq 0.5$) score by dissent, otherwise score by self-inconsistency. *Hard two-bucket route*: split by the same gate and, within each bucket, select the single best signal and its *polarity* on the training folds (5-fold out-of-fold). Table 11 reports the result. Both discrete routes clear the best single signal decisively (trivial and hard both $+0.015$, $[+.007, +.023]$, significant), and the hard route (0.8380) is *statistically indistinguishable* from soft DACR ($\Delta$=$+0.014$, 95% CI $[-0.003, +0.030]$, ns) on the *full* set: the empirical gain comes from routing on the input-derived gate, not from the softness of the interaction block. The story sharpens once we split by gate confidence. On the *boundary* subset where the probe is genuinely uncertain ($0.3 < p_{\text{unk}} < 0.7$, 2,017 questions), soft DACR rises to wsAUROC 0.743 while the hard route only reaches 0.643 ($\Delta$=$+0.095$, 95% CI $[+0.052, +0.138]$, significant); on the complementary *confident* subset the two are indistinguishable ($\Delta$=$-0.007$, ns). The smooth interaction block earns its keep exactly where a hard branch must commit under an ambiguous gate and cannot blend the two polarities.

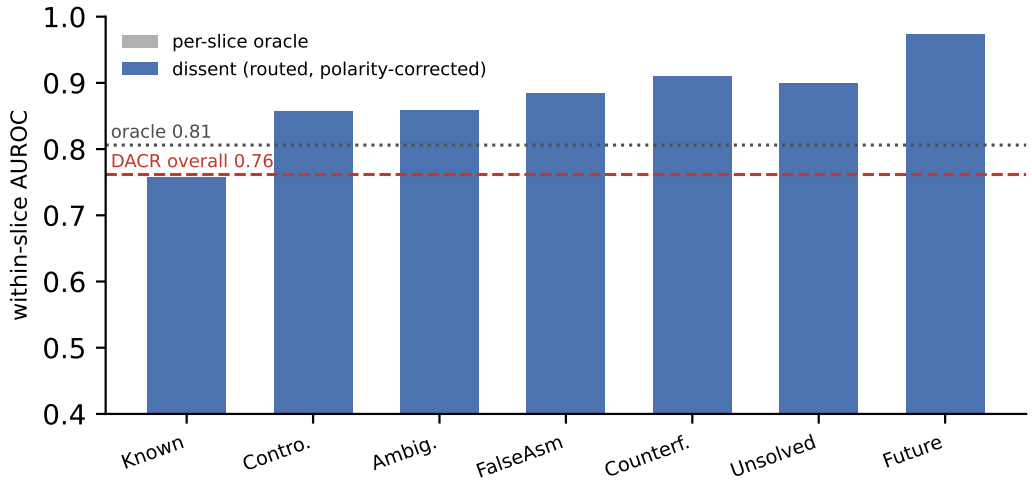


Figure 5: Distance to the per-slice oracle. On the six unanswerable slices the routed dissent signal nearly reaches the slice-wise ceiling (grey); the residual gap between DACR (red dashed) and the oracle (dotted) concentrates almost entirely on the answerable *known* slice, where no black-box behaviour signal is strong.

Table 12: What the router relies on, per defect slice (within-slice AUROC of each base signal on KUQ). The primary signal the router relies on is shown; on the *known* slice dissent collapses to near-chance, so the router down-weights it and relies on self-inconsistency.

| Defect slice | Verb. | Self-inc. | Diss. | Router uses |
|---|---|---|---|---|
| future unknown | 0.61 | 0.56 | **0.97** | diss. (+) |
| counterfactual | 0.60 | 0.53 | **0.91** | diss. (+) |
| unsolved problem | 0.56 | 0.47 | **0.90** | diss. (+) |
| false assumption | 0.56 | 0.49 | **0.88** | diss. (+) |
| ambiguous | 0.53 | 0.44 | **0.86** | diss. (+) |
| controversial | 0.51 | 0.46 | **0.86** | diss. (+) |
| known | 0.57 | **0.75** | 0.54 | self-inc. (primary) |

## H Distance to the Per-Slice Oracle

The main paper reports that DACR (0.851) closes most of the gap to a per-slice oracle (0.887), leaving a significant residual of 0.035 concentrated on the answerable slice. Figure 5 visualizes this: on the six unanswerable slices the routed dissent signal nearly reaches the slice-wise ceiling, so the residual gap between DACR and the oracle sits almost entirely on the answerable *known* slice, where no black-box behaviour signal is strong. This localizes the remaining error: the routing is near-optimal given the signals, and further gains must come from richer content-level signals on answerable questions rather than a better combination rule.

## I What the Router Relies On, Per Defect Slice

Table 12 gives the per-slice within-slice AUROC of each base signal on KUQ that drives DACR's learned routing policy. On the six unanswerable slices the router treats cross-model dissent as a *positive* error indicator; on the answerable (*known*) slice that signal decays to near-chance, so the router down-weights it and leans on self-inconsistency instead. Verbalized confidence is weak everywhere and never the primary input.

| Regime | Cross-model feats | wsAUROC | Paired $\Delta$ vs. best within |
|---|---|---|---|
| Best within-model ($s_{\text{incons}}$) | — | 0.659 | — |
| Single-model DACR | no | 0.743 | $+0.084$ $[+.072, +.096]$ |
| Full DACR | yes | 0.851 | $+0.192$ $[+.170, +.214]$ |
| Full $-$ single-model gap: $+0.109$ $[+.088, +.131]$ (paired, SIG) | | | |

Table 13: Deployment regimes for DACR. Without any cross-model peer (single-model DACR), the probe-gated router already beats the best within-model signal; adding cross-model dissent/peer-abstention lifts wsAUROC by a further $+0.109$. The method degrades gracefully when peers are unavailable.

## J Deployment Regimes: Single-Model vs. Full DACR

The main paper reports that DACR degrades gracefully when no independent peer model is available. Table 13 tabulates the two deployment regimes. When *all* cross-model features (dissent, peer-abstention rate) are removed, leaving only within-model behavioural signals and the text probe, single-model DACR still reaches wsAUROC $0.743$, significantly above the best within-model single signal ($s_{\text{incons}}$, $0.659$). Adding the cross-model term lifts this by a further $+0.109$ $[+.088, +.131]$ to the full DACR's $0.851$.

## K Code and Data

Anonymized code and data accompany this submission. The `REPRODUCE.md` file lists exact commands, fixed seeds, signal/metric definitions, and a mapping from each main-paper table to the source verdict file. Stage-1 (LLM sampling) outputs are included, so all reported numbers can be regenerated offline from the provided `results/` artifacts using the Stage-2 analysis scripts.

## L Negative Control: A Dataset Without Defect-Type Structure

DACR's mechanism presupposes a latent *defect type* whose polarity a signal should track. To probe the boundary of that assumption we ran the same Stage-1 pipeline on a third dataset, TruthfulQA ($817$ questions $\times$ $6$ models $\times$ $K{=}4$), which by construction has *no* answerable/unanswerable partition: every item has a correct answer and the challenge is imitative falsehood, so there is no known/unknown regime for a gate to route between. As expected, the three black-box signals are all near chance there (within-slice AUROC $s_{\text{verb}}{=}0.53$, $s_{\text{incons}}{=}0.46$, $s_{\text{dissent}}{=}0.48$), and no polarity flip appears—there is no defect-type axis along which a signal could reverse. We therefore do *not* report a DACR number on TruthfulQA in the main paper: it is a negative control confirming that the gain is tied to the presence of a defect-type structure, not a dataset we omit for being unfavourable. The main-paper claims are made on the two datasets that carry the required known/unknown structure (KUQ and SelfAware), where the flip and DACR's end-to-end gain both reproduce.

## M A Structurally-Matched Positive Dataset Without a Polarity Flip

The negative control above (TruthfulQA) lacks the known/unknown axis entirely. A sharper boundary case is a dataset that *does* carry an answerable/unanswerable partition—the same structure DACR exploits on KUQ and SelfAware—yet does *not* exhibit the regime-dependent *polarity flip* that makes routing pay off. We construct such a case from SQuAD 2.0 (the unanswerable-question extension of SQuAD), sampling a balanced pilot of $1000$ questions ($500$ answerable, $500$ unanswerable, where "unanswerable" is the `is_impossible` label) and running the identical two-stage pipeline ($6$ models $\times$ $K{=}5$ reading-comprehension samples with the supporting passage in context, plus the text-only defect-type probe).

Table 14: SQuAD 2.0 (balanced pilot, 1000 questions, 6 models, $K=5$). Within-slice AUROC. The routed fusion (DACR-v2) does not significantly beat the best single signal because $s_{\text{dissent}}$ has the *same* polarity on both slices (no flip to route around), unlike KUQ.

| Variant / signal | wsAUROC |
|---|---|
| $s_{\text{verb}}$ | 0.68 |
| $s_{\text{incons}}$ | 0.58 |
| $s_{\text{dissent}}$ (best single) | 0.77 |
| DACR-v1 (behaviour only) | 0.74 |
| DACR-v2 (behaviour + probe) | 0.74 |
| $\Delta$(v2 − best) | $-0.025$ $[-0.086, +0.037]$ |

**Signals reproduce with the *same* polarity as KUQ.** All three black-box signals correlate with the gold unknown label in the same direction as on KUQ (global signal→`gold_unknown` AUROC $s_{\text{verb}}$=0.64, $s_{\text{incons}}$=0.53, $s_{\text{dissent}}$=0.59; all above the 0.5 chance line), so the defect-signal *structure* transfers across datasets. Cross-model dissent is again the strongest error detector (within-slice AUROC 0.77).

**But there is no flip to route around.** Crucially, and unlike KUQ (Table 6), $s_{\text{dissent}}$ does *not* reverse polarity across slices on SQuAD 2.0: its within-slice error-detection AUROC is 0.86 on the answerable slice and 0.67 on the unanswerable slice—both well above chance and pointing the *same* way. On this dataset high cross-model dissent flags a wrong committed action in *either* regime, so a single global fusion is already near-optimal and there is no regime-dependent polarity for a router to condition on.

**DACR therefore yields no significant end-to-end gain here.** Consistent with the absence of a flip, the routed fusion does not beat the best single signal on SQuAD 2.0 (Table 14): DACR-v2's within-slice AUROC is 0.74 versus 0.77 for $s_{\text{dissent}}$, a paired cluster-bootstrap delta of $-0.025$ with 95% CI $[-0.086, +0.037]$ (not significant). The text-only probe is also weaker on SQuAD 2.0 (answerability AUROC 0.69 vs. 0.83 on KUQ), as expected: SQuAD unanswerability is defined *relative to a specific passage*, so a probe that sees only the question cannot reliably judge it. We therefore do *not* report a DACR number on SQuAD 2.0 in the main paper. Together with the TruthfulQA negative control, this delineates precisely when DACR helps: the gain requires not merely a known/unknown partition but a *regime-dependent polarity flip* in the underlying signal. When signals are informative but their polarity is regime-invariant, a fixed fusion already suffices and routing is neither needed nor harmful.

## N Computational Resources and API Usage

**Computing infrastructure (hardware and software).** All experiments were run on a single Linux server node: dual Intel Xeon Platinum 8575C CPUs ($2 \times 48$ cores, 192 logical threads, 3.2–4.0 GHz) with 1.5 TB RAM, running Enterprise Linux 7.2 (kernel 5.10.134). The node also hosts 8× NVIDIA H20-3e GPUs (143.7 GB each, driver 570.133.20, CUDA 12.8), but *no local GPU was used for this work*: all six frontier models are queried through a hosted OpenAI-compatible API, and the DACR router is a single $\ell_2$-regularized logistic regression fitted with `scikit-learn` on CPU (negligible compute; end-to-end Stage-2 analysis—signal aggregation, router fitting, five-fold OOF cross-validation, within-slice AUROC, and the paired cluster bootstrap—completes in under a minute). The software stack is Python 3.10.13 with `numpy` 1.26.4, `scikit-learn`, `scipy`, and `matplotlib` 3.10.1; the LLM client uses `openai` 2.6.1. Manuscript writing and local LaTeX compilation were done on a separate laptop that runs no experiments.

**API models.** Six frontier models were accessed via an OpenAI-compatible gateway: GPT-5.5, CLAUDE-OPUS-4-8, GEMINI-3.5-FLASH, QWEN3.7-PLUS, GLM-5V-TURBO, and DOUBAO-SEED-2.0-PRO-260215. Generation temperature was 0.7; we drew $K=4$ samples per (model, question) for KUQ, SelfAware, and TruthfulQA, and $K=5$ for the SQuAD 2.0 reading-comprehension control. The correctness judge (TruthfulQA third-dataset labels) used GPT-5.5 at temperature 0.0.

**Total API calls.**

- **Generation (signal computation):** $71{,}274$ KUQ $+$ $49{,}530$ SelfAware $+$ $30{,}000$ SQuAD 2.0 $+$ $19{,}607$ TruthfulQA generations $\approx$ **170,411 calls** ($2{,}970/2{,}064/1{,}000/817$ questions $\times$ 6 models $\times$ $K{=}4$, except SQuAD at $K{=}5$).
- **Correctness judge (TruthfulQA):** $12{,}454$ calls (GPT-$5.5$).

Total: approximately **183,000 API calls** across all experiments. The SQuAD 2.0 and TruthfulQA calls come from the negative-control probes of Sections L–M, which is why the API budget spans four datasets while the main paper reports the DACR mechanism on the two datasets that carry a defect-type structure.

**Cost estimate.** At 2026-07 hosted-API pricing, the total cost was approximately **$350–450 USD** (generation dominates; judge calls are a small fraction).

**Hosted API drift.** As these are hosted API versions available at run time (2026-07), exact weights may drift over time and absolute numbers may shift. However, the reported *effects*—the polarity flip of $s_{\text{incons}}/s_{\text{dissent}}$ and DACR's end-to-end improvement over the best single signal—are stable across the six models and both defect-typed datasets (KUQ and SelfAware).