JuryProbe: An Empirical Consensus-Risk Diagnostic for Routing Reference-Free Factuality Judge Panels to Grounded Verification

Tianxin Zhou *zhoutx0@gmail.com*
Independent Researcher

Ruixi Lin *ruix.lin@northeastern.edu*
Independent Researcher

Abstract

Model-based evaluation systems increasingly use panels of inexpensive LLM judges to make accept-or-escalate decisions. In factuality settings, accepting a claim because several reference-free judges agree can create a hidden risk: agreement may arise from shared false-negative blind spots rather than independent evidence. We introduce JuryProbe, an empirical consensus-risk diagnostic for reference-free factuality judge panels, paired with a calibration-based routing policy. JuryProbe estimates panel-level consensus risk from a labeled calibration probe using false-negative-only (FN-only) judge correlation and false-consensus lift; when a panel is flagged high-risk, reference-free majority accepts are routed to the same judge panel with trusted references. Using frozen FEVER-supported claim pools, fixed-seed construction, and audited Number and Entity corruption families, we show that reference-free panels exhibit substantial correlated false negatives (FN-only correlations of 0.402 and 0.368; false-consensus lifts of 3.13× and 18.13×). Under a trusted-reference best-case diagnostic, unanimous false consensus drops to zero both on minimal-pair FEVER corruptions and on non-minimal-pair scientific evidence. In flagged settings, the routed policy is, by construction, equivalent to grounding every reference-free majority accept (verified in 34/34 flagged splits). Its improvement therefore comes from accept-conditioned grounding, while the diagnostic's operational role is to determine whether to activate this policy. A fixed, pre-specified rule produces both labels out of sample: it flags synthetic, benchmark-authored, and scientific claim families in 8–10 of 10 splits each, while standing down on the negative control in 0 of 10 splits. There, standing down avoids grounding roughly 28% of deployment claims at a 0.004 absolute increase in false-accept rate. Stress tests with BM25-retrieved references and distribution-shifted deployment streams show that false-accept reduction persists under weak retrieval, while true-accept coverage degrades and stale stand-down labels require periodic labeled recalibration. JuryProbe provides no formal risk guarantee, nor do we establish reliable stand-down on natural panels; the supported contribution is an empirical diagnostic of high-risk panel error dependence.

1 Introduction

In many model-evaluation pipelines, the most consequential action is not producing a score, but deciding whether an output should be accepted without further verification. A cheap reference-free judge panel makes this decision attractive: several models can be queried quickly, and majority or unanimous agreement appears to provide redundancy. Yet this deployment logic relies on an assumption that is rarely tested: agreement is informative only when judge errors are sufficiently independent. If the judges share the same false-negative blind spots, the panel may accept a corrupted claim precisely when it appears most confident. In this setting, agreement is not merely noisy evidence; it can become the event that triggers unsafe acceptance.

This paper studies that failure mode as a measurement-and-routing problem. We introduce JuryProbe, an empirical consensus-risk diagnostic with calibration-based routing for reference-free factuality judge panels. Prior work has established LLM-as-a-judge evaluation as a practical paradigm (Zheng et al., 2023; Wang et al., 2024; Zhu et al., 2025), and recent work has explored replacing single judges with panels of smaller models (Verga et al., 2024). JuryProbe asks a narrower deployment question: when should agreement from a cheap reference-free factuality panel be trusted, and when should it be routed to grounded verification?

JuryProbe estimates consensus risk from a calibration probe using two complementary panel-level statistics. FN-only correlation measures dependence: whether judges fail on the same corrupted claims. False-consensus lift measures consequence: whether unanimous false acceptance occurs more often than expected from the judges' marginal false-negative rates. Together, these statistics define a panel-level risk regime rather than a sample-level classifier.

This framing differs from uncertainty or disagreement-based escalation. Selective prediction and escalation methods route uncertain cases to stronger systems or humans (Geifman & El-Yaniv, 2017; 2019; Jung et al., 2025). Those methods are valuable, but false consensus is different: the dangerous cases are precisely those in which the reference-free judges agree. Disagreement-based escalation cannot catch unanimous false acceptance by construction. JuryProbe therefore routes accept decisions when calibration shows that the panel is in a high-risk consensus regime, not merely when the judges disagree on a particular item. The diagnostic is empirical: it does not certify that an unflagged panel is safe, and we evaluate both rule outcomes explicitly.

We evaluate JuryProbe on audited Number and Entity corruptions from frozen FEVER pools and six additional claim families. Reference-free panels exhibit substantial consensus risk, with no unanimous false consensus observed under trusted references. In held-out evaluation, gains come from accept-conditioned grounding: in flagged splits, the routed policy is by construction identical to grounding every reference-free majority accept, while the diagnostic only determines whether to activate it. On a negative control, the rule stands down in all splits at a small false-accept cost.

This paper makes three contributions.

- We define *consensus risk* for reference-free factuality judge panels and show that FN-only correlation and false-consensus lift are complementary panel-level signals that capture dependence and its deployment consequence, a failure mode that is invisible by construction to disagreement-based escalation.
- We provide a trusted-reference grounding diagnostic: holding the judge panel fixed, the unanimous false consensus observed under reference-free judging is not observed when the same judges are given trusted references, both for minimal-pair FEVER corruptions and for non-minimal-pair scientific evidence. This diagnostic represents a best-case assessment under trusted references rather than a causal isolation of why grounding improves reliability.
- We evaluate JuryProbe-Routed with an explicit policy-identity analysis: in every flagged split, it coincides with the no-estimator Ground-All-Accepts baseline (verified in 34/34 flagged splits), so the diagnostic's operational role is to determine whether accept-conditioned grounding is activated. A fixed, pre-specified rule produces both labels across eight claim families, while stand-down on a negative control avoids approximately 28% of reference acquisitions.

Together, these results suggest that reliability in factuality judging can be improved by risk-aware grounding rather than by relying on reference-free agreement or simply adding more reference-free judges.

2 Related Work

2.1 LLM-as-a-Judge and Judge Panel

LLM-as-a-judge systems have become a practical alternative to human evaluation for open-ended model outputs. Zheng et al. (2023) introduce MT-Bench and Chatbot Arena as scalable evaluation setting based

on LLM judgments, while Wang et al. (2024) and Zhu et al. (2025) study automatic and fine-tuned LLM evaluators. Verga et al. (2024) further motivate panels of smaller, diverse judges as an alternative to a single expensive judge. JuryProbe builds on this deployment setting, but studies a different question: not whether judge panels are useful on average, but when reference-free panel agreement should no longer be treated as reliable evidence because judge errors may be correlated on the same factual corruptions.

2.2 Judge Correlation and Dependence

The value of a judge panel depends on whether additional judges provide additional independent evidence. Kohli (2026) studies this issue directly, showing that nominally large LLM judge panels can yield far fewer effective independent votes when judge errors are correlated. This motivates JuryProbe: if judges fail on the same examples, then majority or unanimous agreement cannot be interpreted as independent confirmation.

Our contribution is complementary. Kohli (2026) studies judge independence and effective votes in LLM evaluation panels, whereas JuryProbe studies reference-free factuality as a routing problem. We narrow the failure mode to correlated false negatives on corrupted factual claims, measure the downstream consequence through false-consensus lift, use grounding as a trusted-reference diagnostic and an escalation target, and evaluate a routed policy that sends high-risk accept decisions to grounded verification. JuryProbe therefore treats correlated failures as an operational risk signal that drives verification decisions, rather than solely as a property of panel composition.

2.3 Factuality Verification and Grounding

Factuality evaluation has been studied through atomic fact decomposition, hallucination detection, sampling-based consistency, and retrieval-augmented verification (Min et al., 2023; Wei et al., 2024; Manakul et al., 2023; Li et al., 2023).

JuryProbe builds on the observation that grounding can improve factuality judgments, but it does not propose a new factuality verifier. Instead, grounding is used as an intervention and an escalation target. The question is not whether grounded verification is useful in general, but when a cheap reference-free panel should be routed to it.

2.4 Selective Evaluation and Escalation

Selective prediction studies how models can abstain or defer on unreliable inputs, trading coverage for reliability (Geifman & El-Yaniv, 2017; 2019). In LLM evaluation, Jung et al. (2025) study escalation based on estimated agreement with human judgment. JuryProbe likewise treats evaluation as a decision problem, but differs in the routing signal: it routes based on measured consensus risk rather than uncertainty, confidence, or disagreement. This distinction matters because disagreement-based escalation cannot catch unanimous false acceptance by construction. These methods are complementary: Jung et al. (2025) and Geifman & El-Yaniv (2017) give formal guarantees for judge–human agreement and single-model selective prediction, respectively, whereas JuryProbe studies panel error dependence without distribution-free guarantees. Its risk thresholds were fixed before held-out evaluation (Section 3.2).

Unlike prior work, JuryProbe combines judge panels, factuality verification, grounded escalation, and consensus-risk-aware routing in a single diagnostic and routing framework. Its goal is not to improve factuality verification itself, but to determine when reference-free agreement should no longer be trusted without grounding. A compact comparison with related work is provided in Appendix A.

3 JuryProbe Framework

JuryProbe is an empirical consensus-risk diagnostic, paired with a routing policy, for reference-free factuality judging. Its goal is not to replace a judge panel with a new factuality verifier, but to decide when agreement from a reference-free panel should no longer be treated as sufficient evidence for accepting a claim.

3.1 Problem Setup

We consider a factuality decision setting in which a claim must be either accepted as factual or rejected as non-factual. Each example consists of a claim c_i and a ground-truth factuality label $y_i \in \{0, 1\}$, where $y_i = 1$ denotes a factual claim and $y_i = 0$ denotes a corrupted claim. When grounded verification is used, a trusted reference r_i is additionally provided.

A judge receives a claim and returns a binary decision. For a panel of m judges, let $z_{ij}^{\mathrm{RF}} \in \{0, 1\}$ denote whether judge j accepts item i in the reference-free setting, where 1 means accept and 0 means reject. Throughout this work, an accept decision means that the judge treats the claim as factual. The panel majority decision is

$$A_i^{\mathrm{RF}} = \mathbb{I}\left[\sum_{j=1}^{m} z_{ij}^{\mathrm{RF}} \geq \left\lceil \frac{m}{2} \right\rceil\right]. \tag{1}$$

In our main experiments, $m = 3$, so majority acceptance means that at least two judges accept the claim.

The critical failure mode is false acceptance of corrupted claims. For a corrupted item, we use the term false negative in the corruption-detection sense: the judge fails to detect the corruption and incorrectly accepts the claim as factual. Thus, for a corrupted item ($y_i = 0$), judge j makes a false negative when $z_{ij}^{\mathrm{RF}} = 1$. A false-consensus event occurs when all judges accept the same corrupted claim:

$$C_i^{\mathrm{RF}} = \mathbb{I}\left[y_i = 0 \ \wedge \ \sum_{j=1}^{m} z_{ij}^{\mathrm{RF}} = m\right]. \tag{2}$$

Disagreement-based escalation cannot detect this event: the panel appears maximally reliable while all judges make the same error. JuryProbe is motivated by the observation that it can occur at rates exceeding the independent-error expectation.

3.2 Consensus Risk

JuryProbe estimates consensus risk on a calibration probe before applying any routed policy to deployment examples. The calibration probe consists of corrupted claims from a fixed corruption family and the reference-free decisions of a judge panel on those claims. The resulting risk estimate is a panel-level statistic: it characterizes whether a particular judge panel tends to share false-negative failures on a class of factual corruptions. It is not a classifier that predicts whether an individual deployment item is safe. Let $\mathcal{N}_{\mathrm{cal}} = \{\, i : y_i = 0 \,\}$ denote the corrupted items in the calibration probe. For each corrupted item $i \in \mathcal{N}_{\mathrm{cal}}$ and judge j, define the false-negative indicator $e_{ij} = \mathbb{I}\left[z_{ij}^{\mathrm{RF}} = 1\right]$.

JuryProbe uses two complementary statistics. First, FN-only correlation measures whether judges fail on the same corrupted items. For each judge pair (j, k), we compute the Pearson correlation between their false-negative vectors: $\rho_{jk} = \mathrm{corr}\left(e_{\cdot j}, e_{\cdot k}\right)$. Because e_{ij} is binary, this equals the phi coefficient. The panel-level FN-only correlation is the average pairwise correlation,

$$\rho_{\mathrm{FN}} = \frac{2}{m(m-1)} \sum_{1 \leq j < k \leq m} \rho_{jk}. \tag{3}$$

Positive FN-only correlation indicates that judges tend to miss the same corruptions rather than making independent errors.

Second, false-consensus lift measures the consequence of that dependence for unanimous false acceptance. Let $q_{\mathrm{obs}} = \frac{1}{|\mathcal{N}_{\mathrm{cal}}|} \sum_{i \in \mathcal{N}_{\mathrm{cal}}} \mathbb{I}\left[\sum_j e_{ij} = m\right]$ denote the observed false-consensus rate and p_j each judge's marginal false-negative rate; under independent errors the expected rate is $q_{\mathrm{ind}} = \prod_{j=1}^{m} p_j$. False-consensus lift is then

$$L_{\mathrm{FC}} = \frac{q_{\mathrm{obs}}}{q_{\mathrm{ind}}} \tag{4}$$

Values above 1 indicate excess unanimous false acceptance relative to independence; thus, lift captures the consequence of dependence rather than dependence itself.

Together, the two statistics define a consensus-risk regime for a judge panel. Importantly, consensus risk is assessed at the panel level and remains fixed throughout deployment; it is not recomputed for individual claims.

To assess whether the observed unanimous false-consensus rate is stronger than expected from marginal false-negative rates alone, JuryProbe uses a permutation test with 3000 permutations. The test preserves each judge's marginal false-negative rate while shuffling the locations of false negatives across calibration items. This produces a null distribution in which judges have the same individual error rates but no item-level alignment. The p-value is computed as the fraction of shuffled panels whose unanimous false-consensus rate is at least as large as the observed rate, with one-count smoothing.

A panel is treated as high-risk when the calibration probe satisfies three conditions:

$$\rho_{\mathrm{FN}} > 0.15, \quad L_{\mathrm{FC}} > 1.5, \quad p < 0.05.$$

These thresholds define a risk regime, not an item-level decision rule. They were selected prior to deployment evaluation and are examined through a threshold-sensitivity analysis in Section 5.6. The risk regime is estimated only on calibration data and then carried forward to deployment evaluation. In the JuryProbe policy, the high-risk label means that reference-free agreement should no longer be treated as sufficient evidence for accepting claims without grounding.

Threshold provenance. The three constants were selected from initial Number-family calibration diagnostics and frozen on 2026-06-09, before held-out multiseed evaluation. Thus, Number is the threshold-development setting, while Entity, Attribute, both controls, FEVER-Refutes, SciFact, and CREAK were evaluated out of sample under the unchanged rule. No held-out outcome informed threshold selection. The rule is empirical, not a guarantee: it cannot target a specified error rate, and the permutation test does not bound deployment error.

3.3 Grounded Verification

Grounded verification defines the reference-augmented protocol used for the trusted-reference grounding diagnostic and routed escalation. The same judge panel evaluates the same claim with access to a trusted reference; judge identities, binary output space, and majority aggregation are kept fixed.

We denote reference-free decisions by z_{ij}^{RF} and grounded decisions by z_{ij}^{G}. In the reference-free setting, judge j receives only the claim c_i and returns $z_{ij}^{\mathrm{RF}} \in \{0, 1\}$. In the grounded setting, the same judge receives (c_i, r_i), where r_i is the trusted reference, and returns $z_{ij}^{\mathrm{G}} \in \{0, 1\}$. The ground-truth label y_i is not provided to the grounded verifier.

Grounded majority acceptance A_i^{G} and grounded false consensus C_i^{G} are defined analogously to Eqs. 1 and 2, replacing z_{ij}^{RF} with z_{ij}^{G}.

Section 5.2 compares paired reference-free and grounded judgments under a fixed judge panel to assess whether the false-consensus events observed without references persist when trusted references are supplied. Since the references are trusted by construction, we treat this comparison as a *trusted-reference best-case diagnostic*, rather than evidence for a mechanism-level account of why grounding improves judgments.

3.4 JuryProbe-Routed Policy

JuryProbe uses the consensus-risk estimate to decide whether reference-free accept decisions require grounded verification. Let $h \in \{0, 1\}$ denote the panel-level risk label estimated on the calibration probe, where $h = 1$

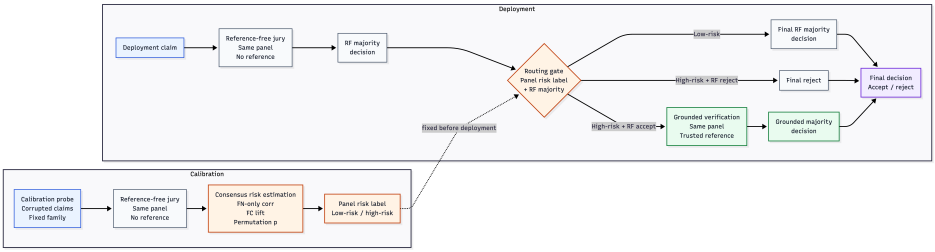

Figure 1: JuryProbe-routed policy. Consensus risk is estimated once on a calibration probe and fixed during deployment.

indicates that the panel is high-risk according to the thresholds in Section 3.2. This risk label is fixed before deployment evaluation. It is not recomputed for each deployment item.

For a deployment claim i, the reference-free panel first produces the majority decision A_i^{RF} defined in Eq. 1. The final JuryProbe decision is

$$D_i^{\mathrm{JP}} = \begin{cases} A_i^{\mathrm{RF}}, & h = 0, \\ A_i^{\mathrm{G}}, & h = 1 \ \wedge \ A_i^{\mathrm{RF}} = 1, \\ 0, & h = 1 \ \wedge \ A_i^{\mathrm{RF}} = 0. \end{cases} \tag{5}$$

Figure 1 summarizes the routed policy.

This asymmetric design is intentional. JuryProbe is optimized to prevent unsafe acceptance of corrupted claims rather than to maximize acceptance coverage. In factuality verification, accepting a corrupted claim is the safety-relevant error: it allows a false statement to pass as factual. By contrast, a reference-free reject already blocks the claim. JuryProbe therefore does not use grounded verification to recover additional accepts; it uses grounding to prevent high-risk false accepts. One-sided accept protection also has a structural property: because the routed policy only overturns accepts, its false-accept rate cannot exceed the reference-free rate by construction. Full decision replacement offers no such protection; Section 5.3 reports a control family where Always Grounded increases false accepts.

Unlike disagreement-based escalation, JuryProbe can route even unanimous accepts, with verification driven by calibration rather than item-level disagreement.

4 Experimental Setup

We evaluate JuryProbe in a controlled factuality setting where clean claims are supported by trusted references and corrupted claims have known non-factual labels. This section describes the frozen corruption datasets, judge panels, metrics, and held-out protocol used to separate risk estimation from policy evaluation.

4.1 Claim Pools and Corruption Families

We construct examples from FEVER-supported claims (Thorne et al., 2018). To avoid adaptive example selection, claim pools, corruption generation, and audits are frozen before judge evaluation.

The confirmatory evaluation uses two audited corruption families. Number corruptions modify numerical facts such as years, counts, or quantities, while Entity corruptions replace named entities with plausible alternatives. Each confirmatory family contains 300 clean and 300 corrupted examples, supporting equal-sized calibration and deployment splits. This evaluation budget was fixed before judge evaluation.

We additionally construct an Attribute family for replication analysis. Relation corruptions were explored but excluded before judge evaluation because audits revealed unstable fluency, syntax, and semantic-naturalness artifacts.

Beyond the FEVER-derived corruption families, we evaluate the fixed risk rule on five additional frozen families (Section 5.4). Two are FEVER-pool controls: a Self-Contained Contradiction control (300 clean / 300 corrupted), where corruptions append one of 40 self-contained numeric contradictions, and an Obvious-Number boundary control with blatant numeric errors. Three are benchmark-authored: FEVER-Refutes, a frozen atomic SciFact subset of 190 supported and 190 contradicted expert-written claims (Wadden et al., 2020), and an unfiltered CREAK (Onoe et al., 2021) commonsense sample. Construction details are in the appendix.

4.2 Judge Panels

Our main judge panel consists of three relatively small LLM judges: Llama-3.1-8B-Instruct, Qwen-2.5-7B-Instruct, and Gemma-3-12B-IT. The same three judges are used for both reference-free judging and grounded verification. In the reference-free setting, judges receive only the claim; in the grounded setting, they receive the same claim with a trusted reference and determine whether the claim is fully consistent with that reference. Holding judge identities fixed keeps judge composition constant while the protocol changes by adding trusted references. All outputs are converted into binary accept/reject decisions. We additionally evaluate a pre-specified larger-capacity panel (Llama-3.3-70B-Instruct, Qwen-2.5-72B-Instruct, and Gemma-2-27B-IT) once on fixed SciFact claims, using a grouped five-fold protocol frozen before any judge calls (Section 5.6).

4.3 Metrics

We report metrics for three purposes: consensus-risk estimation, the grounding diagnostic, and policy evaluation.

For consensus-risk estimation, we report FN-only correlation (ρ_{FN}), false-consensus lift (L_{FC}), and the permutation-test p-value defined in Section 3.2.

For the grounding diagnostic, we compare reference-free and grounded judging using FN-only correlation and unanimous false-consensus rate.

For policy evaluation, we report false accept rate, true accept rate, residual false-consensus rate, extra verifier items, and extra verifier calls. False accept rate is the fraction of corrupted claims accepted by the final policy. True accept rate is the fraction of clean claims accepted by the final policy. Residual false-consensus rate is the fraction of corrupted claims unanimously accepted by the reference-free panel and not blocked by the final policy. Extra verifier items denote the number of deployment examples routed to grounded verification, and extra verifier calls denote the corresponding number of grounded judge calls. Tables report raw counts and rates where space permits; Wilson 95% intervals and per-split ranges for key rates are given in the appendix. Lift is comparable only under a common independence null and is flagged as unstable when the permutation null approaches zero.

4.4 Held-out Evaluation Protocol

For policy evaluation, we separate risk estimation from policy measurement using held-out splits. For each corruption family and split seed, the 300 clean and 300 corrupted examples are partitioned into equal-sized calibration and deployment splits, each containing 150 clean claims and 150 corrupted claims. The calibration split is used only to estimate the panel-level risk label h using the thresholds in Section 3.2. The resulting risk label is fixed before deployment evaluation and is not recomputed on deployment examples.

We repeat this procedure over 10 split seeds and report mean and standard deviation across splits. This multi-seed protocol tests whether the high-risk label and policy outcomes are stable under different calibration/deployment partitions.

We compare JuryProbe-Routed against six baselines: Reference-Free Majority, Reference-Free Unanimity, Always Grounded, Ground-All-Accepts, Disagreement-Routed, and Random-Routed. Always Grounded grounds all deployment claims, while Ground-All-Accepts grounds every reference-free majority accept without the risk estimator; it is JuryProbe-Routed's no-estimator counterpart and coincides with it by

construction in high-risk splits. Disagreement-Routed escalates only when reference-free judges disagree. Random-Routed uses the same grounded-verification budget as JuryProbe-Routed but selects deployment items uniformly at random from the deployment split, isolating the effect of consensus-risk routing from simply spending additional verifier calls.

All policies are evaluated on the same deployment splits using the metrics in Section 4.3. Missing grounded decisions, if any, are treated as missing rather than replaced with gold labels, ensuring that evaluation reflects actual verifier outputs rather than oracle fallback.

5 Results

We structure the results into six parts: consensus-risk estimation, the trusted-reference grounding diagnostic, policy evaluation with identity and stand-down analyses, cross-family evaluation of the fixed rule, a retrieval stress test, and robustness analyses.

5.1 Consensus Risk in Reference-Free Judging

We first test whether reference-free judge panels exhibit consensus risk before any grounded intervention is applied. Table 1 reports the calibration statistics for the audited corruption families. FN-only correlation is computed over all corrupted claims. False-consensus rate, false-consensus lift, and permutation p-value are reported on a detectable corrupted subset. A corrupted claim is considered detectable if an external high-capability factuality detector (GPT-4o) correctly rejects the corruption.

Table 1: Reference-free consensus-risk statistics. N_{corr} is the number of corrupted claims, N_{det} is the detectable corrupted subset used for difficulty-controlled false-consensus analysis, Corr. denotes FN-only correlation, and FC denotes false consensus.

Family	N_{corr}	N_{det}	Corr.	FC Rate	FC Lift	p
Number	300	214	0.402	0.159	3.13×	< 0.001
Entity	300	286	0.368	0.031	18.13×	< 0.001
Attribute	300	296	0.263	0.003	54.55×	0.018

Both confirmatory families exhibit substantial consensus risk, with FN-only correlations of 0.402 and 0.368 and lifts of 3.13× and 18.13×, respectively.

Attribute reproduces the signal ($\rho_{\text{FN}} = 0.263$, lift 54.55×, $p = 0.018$), but its low absolute false-consensus rate (0.003) limits its utility for paired grounding analysis.

Together, these results show that reference-free panel agreement can reflect correlated false-negative failures rather than independent confirmation, even when multiple judges agree. This motivates the next question: whether the false-consensus pattern remains when the same judges receive trusted references.

5.2 Trusted-Reference Grounding Diagnostic

We next test whether the false-consensus pattern observed in reference-free judging remains when the same judges receive trusted references. This analysis uses the two confirmatory families, Number and Entity; Attribute is reported as replication evidence but is not used for this paired analysis, because its absolute false-consensus event rate is too low to provide a stable paired comparison.

Table 2 compares reference-free and grounded judging on the same GPT-4o detectable corrupted claims. The reference-free statistics in this table are computed on that subset and therefore differ slightly from the all-corrupted FN-only correlation reported in Table 1. The judge identities and aggregation rule are held fixed, while the protocol changes by adding trusted references.

For both confirmatory families, the false-consensus pattern is not observed under grounding. Number falls from RF correlation 0.386 and false-consensus rate 0.159 to grounded correlation 0.000 and false-consensus

Table 2: Trusted-reference grounding diagnostic on the confirmatory families. RF denotes reference-free judging, Corr. denotes FN-only correlation, and FC denotes unanimous false consensus. The same judge panel evaluates the same corrupted claims with and without trusted references. Grounded FC lift is not reported because, in both families, the grounded false-consensus rate and the corresponding independence baseline are zero, making the lift ratio undefined.

Family	RF Corr.	Grounded Corr.	RF FC Rate	Grounded FC Rate	RF FC Lift
Number	0.386	0.000	0.159	0.000	3.13×
Entity	0.393	-0.003	0.031	0.000	18.13×

rate 0.000. Entity shows the same pattern, with RF correlation 0.393 and false-consensus rate 0.031 falling to approximately zero under grounding.

These results provide a trusted-reference best-case diagnostic: with judges and aggregation fixed, the false-consensus pattern under reference-free judging is not observed with benchmark evidence. Because FEVER corruptions are minimal edits directly contradicted by their references, may partly reflect the simple consistency check.

Beyond minimal-pair edits, we evaluate 190 contradicted SciFact claims against the benchmark rationale and full published abstract (274 tokens on average). Neither reference was constructed from the claim. No three-judge false consensus is observed under either reference (0/190 vs. 10/190 reference-free), and per-judge false accepts drop substantially; mean pairwise FN correlation also falls but remains nonzero (0.247 with rationales; 0.108 with abstracts; Appendix E). Thus, grounding reduces but does not eliminate correlated errors, and these comparisons remain diagnostic rather than causal.

5.3 Policy Evaluation: Baselines, Identity, and Stand-Down

We next evaluate whether consensus-risk routing improves deployment decisions. For each family, risk is estimated only on the calibration split and policy metrics are reported on held-out deployment examples over 10 split seeds. In both Number and Entity, JuryProbe detects a high-risk panel in all 10 splits.

Policy identity in flagged splits. In every high-risk split, JuryProbe-Routed and Ground-All-Accepts are identical by construction: both ground exactly the reference-free majority accepts. We verify this on frozen caches: across all 34 flagged splits in which both policies were evaluated (Number 10, Entity 10, Attribute 8, boundary control 6), every reported field matches with zero discrepancies. Thus, improvements in flagged settings arise from accept-conditioned grounding, while the estimator's distinct operational role is determining whether to activate it. The policies differ only when the rule stands down, evaluated below on the negative control.

Table 3 compares JuryProbe-Routed with reference-free, grounded, no-estimator, disagreement-based, and budget-matched routing baselines. JuryProbe-Routed eliminates false accepts and residual false consensus in both confirmatory families (0.427 $\rightarrow$ 0.000 for Number; 0.119 $\rightarrow$ 0.000 for Entity) while preserving the reference-free true-accept rate by routing accepts only.

Disagreement-Routed leaves residual false consensus because unanimous false accepts provide no disagreement signal. Random-Routed fails to eliminate false accepts despite using the same verifier-call budget as JuryProbe-Routed. Always Grounded also eliminates false accepts in these two families but requires verification of every deployment item and achieves full clean coverage. JuryProbe-Routed uses 49.6% fewer verifier calls for Number and 62.1% fewer for Entity, while accepting 58–64% of clean claims versus 100%. This coverage loss is the cost of one-sided accept protection.

Stand-down on the negative control. The fixed rule stands down in 10/10 Self-Contained Contradiction splits, issuing zero grounded calls. For an empirical no-estimator comparison, we grounded all reference-free-majority accepts appearing in at least one deployment split (165 unique items: 161 clean, 4 corrupted; 495

Table 3: Held-out policy evaluation over 10 split seeds. Values are mean ± standard deviation across splits. Residual FC denotes residual false-consensus rate after policy decisions. Verifier Calls denotes the number of grounded judge calls used by the policy. Always Grounded uses actual grounded verifier outputs from the same judge panel, not oracle labels. The Ground-All-Accepts rows match JuryProbe-Routed exactly because every split in both families is flagged high-risk, making the two policies identical by construction.

Family	Policy	False Accept	True Accept	Residual FC	Verifier Calls
Number	RF Majority	0.427±0.029	0.581±0.029	0.201±0.022	0±0
Number	RF Unanimity	0.201±0.022	0.273±0.027	0.201±0.022	0±0
Number	Disagreement-Routed	0.201±0.022	0.762±0.029	0.201±0.022	420±19
Number	Random-Routed	0.212±0.013	0.792±0.019	0.100±0.009	454±14
Number	Ground-All-Accepts	0.000±0.000	0.581±0.029	0.000±0.000	454±14
Number	**JuryProbe-Routed**	**0.000±0.000**	0.581±0.029	**0.000±0.000**	454±14
Number	Always Grounded	0.000±0.000	1.000±0.000	0.000±0.000	900±0
Entity	RF Majority	0.119±0.014	0.639±0.027	0.035±0.011	0±0
Entity	RF Unanimity	0.035±0.011	0.394±0.027	0.035±0.011	0±0
Entity	Disagreement-Routed	0.035±0.011	0.841±0.015	0.035±0.011	306±10
Entity	Random-Routed	0.074±0.008	0.776±0.021	0.022±0.007	341±13
Entity	Ground-All-Accepts	0.000±0.000	0.639±0.027	0.000±0.000	341±13
Entity	**JuryProbe-Routed**	**0.000±0.000**	0.639±0.027	**0.000±0.000**	341±13
Entity	Always Grounded	0.000±0.000	1.000±0.000	0.000±0.000	900±0

Table 4: Stand-down comparison on the Self-Contained Contradiction negative control (10 splits; 300-item deployment halves). All values are executed grounded-panel results, not oracle labels. References / Split counts unique claims requiring a trusted reference.

Policy	False Accept	True Accept	References / Split
RF Majority	0.013±0.006	0.552±0.023	0
JuryProbe-Routed (stand-down)	0.013±0.006	0.552±0.023	**0**
Ground-All-Accepts	0.009±0.005	0.552±0.023	84.7±4.0
Always Grounded	0.138±0.025	1.000±0.000	300

calls; zero parse failures) and ran Always Grounded on all 600 claims (1,800 verdicts; zero parse failures). Table 4 reports the results.

Stand-down saves 84.7 ± 4.0 references per split (about 28% of deployment claims) relative to Ground-All-Accepts, which reduces false accepts from 0.013 to 0.009 by overturning 1 of 4 RF-accepted corruptions and none of 161 accepted clean items. These corruptions are self-contained contradictions appended to the claims, so reference consistency does not target them. Always Grounded accepts 42/300 corrupted claims (false accept 0.138) while accepting all clean claims; this is the failure mode that one-sided accept protection avoids by construction (Section 3.4). An earlier oracle-label simulation suggesting 0.000 for Always Grounded on this control is documented and corrected in Appendix J. This control demonstrates stand-down and its control-specific tradeoff, not natural-panel safety.

Boundary variance. Attribute is flagged in 8/10 splits. In the two non-flagged splits (FN-correlation 0.083 and 0.119), JuryProbe retains reference-free decisions (false accept 0.047 and 0.040; residual unanimous false consensus 0.013), while Ground-All-Accepts would reduce false accepts to at most 0.013 and 0.007 at 127 and 117 references. This illustrates boundary variance near the 0.15 threshold.

5.4 Cross-Family Evaluation of the Fixed Rule

We apply the unchanged thresholds and splitwise estimator to eight claim families with 10 splits each, using disjoint calibration and deployment items (Table 5). A split is flagged when all correlation, lift, and permutation criteria are met; mean lift is omitted when near-zero permutation nulls make it unstable, with

Table 5: Cross-family evaluation of the fixed rule (frozen thresholds, no per-family tuning): flagged splits out of 10, with cross-split mean ± standard deviation of calibration-side FN-only correlation and lift. Entries marked – are unstable under near-zero permutation nulls (Appendix F).

Family	Type	Flagged	FN Corr	FC Lift
Number	programmatic FEVER corruption	10/10	0.385±0.028	2.60±0.20
Entity	programmatic FEVER corruption	10/10	0.348±0.062	11.44±5.27
Attribute	programmatic FEVER corruption	8/10	0.265±0.112	–
SciFact	external scientific benchmark	10/10	0.319±0.022	5.57±0.70
FEVER-Refutes	benchmark-authored refutations	8/10	0.306±0.073	23.55±13.18
CREAK	benchmark-authored commonsense	8/10	0.318±0.100	–
Obvious-Number	boundary control	6/10	0.330±0.167	–
Self-Contained Contradiction	negative control	0/10	0.082±0.054	0.00 (9), undef. (1)

per-split values in Appendix F. Unlike Table 1, which reports full-family statistics, this table is splitwise and calibration-side, so the aggregation scopes differ.

The fixed rule yields both labels: signal families are flagged in 8–10/10 splits, the negative control in 0/10, and the boundary control in 6/10. A 95-clean/95-false SciFact sensitivity analysis remains flagged in 10/10 splits (FN correlation 0.305 ± 0.043; lift 5.11 ± 0.78). These results extend evaluation beyond the submitted FEVER corruptions but remain replication when calibration is drawn from the evaluated family; they do not establish generalization to error families absent from calibration, which we test under distribution shift in Section 6.1.

5.5 Imperfect References: A BM25 Retrieval Stress Test

The zero false-accept rates in Table 3 use trusted benchmark FEVER evidence and are best-case results. We therefore stress-test routing on SciFact with a fixed, unsupervised, claim-only BM25 retriever over abstract sentences; labels and gold annotations are used only for evaluation. Retrieval is weak, with corrupt-side any-evidence recall of 0.300 at k=1 and 0.437 at k=3. On the frozen subset (flagged in 10/10 splits, so routed rows equal Ground-All-Accepts), false accepts fall from 0.211 under reference-free majority to 0.016 (BM25@1) and 0.021 (BM25@3), versus 0.026 with benchmark rationales, while true-accept coverage falls from 0.595 to 0.295 and 0.326. The disjoint held-out analysis is consistent (BM25@3: false accept 0.020 ± 0.023, true accept 0.310 ± 0.027). Surviving unanimous false accepts fall from 10/190 to at most 2/190. Full results are in Appendix I. The reduction persists under weak retrieval, but does not establish BM25 sufficiency, retrieval-independent safety, or robustness to incomplete, outdated, adversarial, or production references.

5.6 Robustness and Replication

We perform four additional analyses to evaluate whether the main findings are sensitive to corruption family, judging condition, threshold choice, or judge-panel composition. These analyses support the same interpretation: the observed consensus-risk pattern is not driven by a single corruption family, a grounded-estimator artifact, a narrow threshold choice, or one particular judge-panel composition. Detailed robustness results are reported in Appendix B.

Two pre-specified evaluations probe natural settings. On fixed SciFact claims, both judge panels (Section 4.2) are flagged in 5/5 grouped folds; the larger panel yields 44/190 reference-free false accepts and 21/190 unanimous false consensus, while full-abstract Ground-All-Accepts reduces false accepts to 2/190 using 183 references (Appendix H). CREAK is a boundary case: 8/10 splits flagged, with reference-free false accept 0.029 ± 0.010 and true accept 0.767 ± 0.024. We retain both outcomes without evaluating replacement panels; neither establishes reliable stand-down on natural data (Section 6.2).

6 Discussion

JuryProbe treats reference-free agreement as reliable only after the dependence structure of judge errors has been characterized. This section discusses the operating regime of the routed policy, its behavior under distribution shift, the limitations of the current study, and directions for future work.

6.1 Operating Regime, Costs, and Distribution Shift

Costs. LLM-call cost is negligible across evaluated policies (about 10^{-5} USD per claim for a three-judge evaluation at current API prices), so the relevant deployment-scaling cost is trusted-reference acquisition, reported per policy in Tables 3 and 4. The one-time labeled calibration probe (300 labeled items per family) is reported separately and amortizes over deployment volume. A parameterized utility analysis (Appendix J) makes the operating regime explicit: given the measured policy rates, routing accepts to grounding is preferable to grounding every claim when a trusted reference costs more than 0.29–0.42 times the value of a true accept, and preferable to reference-free operation when a reference costs less than 0.16–0.42 times the cost of a false accept. On the negative control, the diagnostic stands down, and stand-down is preferred unless a reference costs less than about 0.6% of a false accept. JuryProbe is therefore useful in a conditional regime—false accepts are costly, trusted references are available but expensive to acquire universally, and some true-accept coverage can be traded for fewer reference acquisitions—rather than as an unconditional improvement over Always Grounded.

Distribution shift and recalibration. A stand-down label estimated on a calibration probe can become stale under distribution shift. We test this directly with a frozen stress test: an initial 95-clean/95-false negative-control probe stands down in 0/10 splits; we then construct disjoint shifted evaluation streams in which Number, Entity, or SciFact false claims replace increasing fractions of the false claims, while a separate 95/95 labeled audit recomputes the unchanged rule. At target shares of 0/10/25/50/100%, the recalibration audit flags 0/9/10/10/10 splits for Number, 0/3/6/6/10 for Entity, and 0/4/5/7/10 for SciFact (full grid in Appendix G). Without recalibration, the initial stand-down label remains stale, while the reference-free false-accept rate on the shifted streams rises to 0.429 (Number), 0.111 (Entity), and 0.198 (SciFact) at 100% shift. These results support periodic labeled recalibration under the tested shifts; they do not establish zero-shot drift detection or provide guarantees under arbitrary open-world drift.

6.2 Limitations

No formal guarantee. JuryProbe provides no distribution-free guarantee. Its thresholds define an empirical protocol: users cannot specify a target error rate and obtain a rule guaranteed to meet it, and the permutation test provides no bound on deployment error. A split not flagged as high-risk is therefore not a zero-error certificate; it only indicates that the pre-specified calibration criteria for grounded routing were not satisfied.

Natural stand-down and scope. Stand-down is demonstrated as a specificity result on a constructed negative control. On natural data, CREAK represents a boundary case, while both SciFact panel evaluations are flagged in every fold. Thus, this work does not establish when agreement from a natural factuality panel can be trusted without grounding. Our conclusions are limited to short, self-contained binary factuality claims evaluated by small open-weight panels under the stated corruption and reference conditions. Consensus-risk estimates are family-dependent and may not transfer to unseen corruption types, long-form or free-form outputs, or multilingual claims. We exclude relation corruptions from judge evaluation because preliminary audits revealed unstable fluency, syntax, and semantic-naturalness artifacts.

The detectable-subset analysis uses GPT-4o as an external, analysis-time detector. Although it is not part of the JuryProbe panel, grounded verifier, or routing policy, its use is still a modeling choice. Appendix D.4 therefore reports all-corrupted and grounded-detectable variants to assess whether the qualitative conclusions depend on this choice.

Reference quality. All grounded results outside Section 5.5 should be interpreted as trusted-reference best-case diagnostics. The BM25 stress test considers one deliberately weak retriever on a single benchmark;

behavior under degraded or adversarial reference sources beyond this setting remains untested. The negative control further demonstrates that grounded verification is effective for errors that can be resolved against the reference, but does not necessarily address errors arising from a claim's internal logic (Section 5.3).

Diagnostic, not causal. The paired reference-free and grounded comparisons hold the judge panel fixed while changing the input protocol by adding trusted references. They therefore diagnose the reference-augmented protocol rather than isolate the effects of reference wording, prompt framing, or other protocol-level factors. Moreover, pairwise FN correlation persists under grounding on SciFact.

Finally, JuryProbe is designed as an accept-protection policy: the routed policy guards against unsafe acceptance of corrupted claims but does not seek to recover additional true accepts from reference-free rejects. Although the detectable-subset analysis reduces the influence of shared difficulty, it cannot fully eliminate latent sources of item difficulty, a limitation also noted in studies of correlated judge errors (Kohli, 2026).

6.3 Future Work

Long-form, free-form, and multilingual factuality settings are natural directions for future work, as is determining when natural judge panels can be reliably deemed safe to trust without grounding. Extending this framework beyond factuality is another important avenue, particularly for planning, forecasting, recommendation, and other applications where trusted evidence may be incomplete or unavailable. Developing risk-aware routing and intervention strategies for such evidence-sparse settings remains an open challenge. More broadly, the central lesson is that agreement should not be equated with reliability without understanding the dependence structure of judge errors.

7 Conclusion

This paper introduced JuryProbe, an empirical diagnostic of consensus risk in reference-free factuality judge panels, together with a calibration-based routing policy. Rather than assuming that agreement implies reliability, JuryProbe assesses whether a panel exhibits correlated false-negative errors that can give rise to false consensus: a failure mode that disagreement-based escalation cannot detect by construction.

Our supported findings are as follows. Reference-free panels exhibit substantial false-negative dependence on audited Number and Entity corruptions, while the false-consensus pattern is not observed when the same judges are provided with trusted references, for both minimal-pair and non-minimal-pair evidence. In flagged settings, the routed policy is identical by construction to grounding every reference-free majority accept (verified in 34/34 flagged splits): the resulting improvement comes from accept-conditioned grounding, while the diagnostic determines whether that policy should be activated. A fixed, pre-specified rule yields both labels across eight claim families, standing down on a negative control in every split and avoiding roughly 28% of reference acquisitions at a 0.004 increase in false-accept rate. False-accept reduction persists under a deliberately weak BM25 retriever, albeit with substantial coverage loss, and stale stand-down labels are corrected through periodic labeled recalibration under the evaluated distribution shifts.

These conclusions remain subject to important bounds: JuryProbe provides no formal risk guarantee, does not establish reliable stand-down for natural judge panels, and relies on trusted-reference best-case grounding outside the retrieval stress test. Within these limits, reliability in LLM factuality judging depends not only on individual judge accuracy or panel agreement, but also on the dependence structure of judge errors. Agreement alone should therefore not be treated as sufficient evidence for acceptance unless that error dependence is understood.

8 Broader Impact Statement

Three cautions should guide the use of this work. First, a stand-down decision is specific to the evaluated judge panel and data distribution; any change in models, prompts, retrieval, or data requires recalibration with a fresh labeled audit, and a stale label cannot detect open-world distribution shift (Section 6.1). Second,

trusted references may themselves be incomplete, incorrect, outdated, or manipulated; our retrieval stress test measures degradation under one deliberately weak retriever but does not establish robustness to adversarial or production reference sources. Third, automated factuality judging should not replace expert review in high-stakes applications without domain-specific validation.

References

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *International Conference on Learning Representations*, 2017. doi: 10.5555/3295222.3295241. URL https://dl.acm.org/doi/abs/10.5555/3295222.3295241.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: LLM judges with provable guarantees for human agreement. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=UHPnqSTBPO.

Guneet Kohli. Nine judges, two effective votes: Correlated errors undermine LLM evaluation panels, 2026. URL https://arxiv.org/abs/2605.29800.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=bxsrykzSnq.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=RwzFNbJ3Ez.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FACTSCORE: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023. URL https://aclanthology.org/2023.emnlp-main.741/.

Yasumasa Onoe, Michael J. Q. Zhang, Eunsol Choi, and Greg Durrett. CREAK: A dataset for commonsense reasoning over entity knowledge. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://openreview.net/forum?id=mbW_GT3ZN-.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models, 2024. URL https://arxiv.org/abs/2404.18796.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5Nn2BLV7SB.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Zixia Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=4M9f8VMt2C.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Proceedings of the NeurIPS 2023 Datasets and Benchmarks Track*, 2023. URL https://neurips.cc/virtual/2023/poster/73434.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. JudgeLM: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xsELpEPn4A.

A Related Work Positioning

Table 6: Positioning of JuryProbe relative to related work. Panel = LLM judge panel; Fact. = factuality-oriented evaluation; Grounded = grounded verification; Escal. = selective escalation; Consensus Risk = routing based on measured correlated-failure risk.

Work	Panel	Fact.	Grounded	Escal.	Consensus Risk
Kohli (2026)	✓	x	x	x	Partial
Trust or Escalate	✓	x	Partial	✓	x
PoLL	✓	x	x	x	x
LongFact	x	✓	x	x	x
SAFE	x	✓	✓	x	x
FActScore	x	✓	✓	x	x
HaluEval	x	✓	x	x	x
JuryProbe	✓	✓	✓	✓	✓

B Additional Robustness and Artifact Details

Table 7 reports the robustness checks summarized in Section 5.6. These checks test whether the main conclusions depend on threshold choice, grounded-output specificity, random-routing variance, grounded evaluation integrity, or judge-panel composition. All checks are computed from frozen datasets and cached model outputs. Note that the grounded-specificity check is based on grounded verifier outputs: it shows that the estimator is not an always-on trigger, but does not test the reference-free stand-down branch. That branch is evaluated by the negative control in Section 5.3.

C Representative False-Consensus Cases

Table 8 shows representative false-consensus cases. In each case, all three reference-free judges accept the corrupted claim, even though the trusted reference contradicts the modified factual element. When the same judges receive the reference, grounded verification rejects the claim. These examples illustrate why disagreement-based routing is insufficient: the reference-free panel does not disagree on these corrupted claims; it unanimously accepts them.

Table 7: Additional robustness and artifact checks. Threshold sensitivity reports high-risk detections over correlation thresholds from 0.10 to 0.25 and lift thresholds from 1.25 to 2.00. Grounded specificity reports high-risk detections when the same risk estimator is applied to grounded verifier outputs.

Check	Result	Purpose
Threshold sensitivity	Number: 10/10; Entity: 10/10	Not threshold-fragile
Grounded specificity	Number: 0/10; Entity: 0/10	Not always high-risk
Random-Routed stability	100 trials per split	Stable budget-matched baseline
Grounded evaluation integrity	No oracle, no gold fallback, no missing imputation	Actual verifier outputs only
Strong judge slice	$\rho_{\mathrm{FN}} = 0.252$, $L_{\mathrm{FC}} = 2.17\times$, $p < 0.001$	Stronger RF judges do not remove risk

Table 8: Representative false-consensus cases. RF denotes reference-free judging, and G denotes grounded judging with the same judge panel and trusted reference.

Family	Corrupted claim	Reference	Outcome
Number	Pacific Rim was released July 14, 2013.	Pacific Rim was released July 12, 2013.	RF: 3/3 accept; G: 0/3 accept
Number	Alex Rodriguez was suspended for 169 games.	Alex Rodriguez was suspended for 211 games.	RF: 3/3 accept; G: 0/3 accept
Entity	Steve Jobs's birth date is October 28, 1955.	Bill Gates's birth date is October 28, 1955.	RF: 3/3 accept; G: 0/3 accept
Entity	Elton John was named MusiCares' person of the year in 2013.	Bruce Springsteen was named MusiCares' person of the year in 2013.	RF: 3/3 accept; G: 0/3 accept

D Additional Experimental Details

D.1 Judge Prompts and Output Parsing

All reference-free and grounded judgments are converted into binary decisions before evaluation. In the reference-free condition, each judge receives only the claim and is asked to decide whether the statement is factually correct. The prompt requires a one-word output, `true` if the statement is fully correct and `false` if it contains any factual error. In the grounded condition, each judge receives a trusted reference together with the claim and is asked whether the claim is fully consistent with the reference. The grounded prompt also requires a one-word `true`/`false` output. Outputs are parsed using a strict string matcher for `true` or `false`. Responses that do not contain a valid binary verdict are marked as parse failures rather than interpreted manually. For the final grounded-verifier cache used in the policy evaluation, parse failures are retried with the same grounded prompt. The validated cache contains the expected number of grounded decisions for both confirmatory families: 1800 grounded judge outputs for Number and 1800 for Entity, corresponding to 600 examples times three judges. The final validated cache has zero remaining parse failures. No oracle replacement, gold-label fallback, or missing-item imputation is used. Grounded decisions are produced by the same judge panel used in reference-free evaluation, with the only protocol change being the addition of the trusted reference to the judge input. The ground-truth label is never provided to the judge and is used only for evaluation.

D.2 Corruption Audit Protocol

All corruption families are frozen before judge evaluation. Number and Entity are used as confirmatory families because author audits found them reliable enough for the main consensus-risk, grounding-diagnostic, and policy evaluations. Number corruptions modify numerical facts such as years, counts, or quantities, while Entity corruptions replace named entities with plausible alternatives. Attribute is reported as an additional replication family because it satisfies the consensus-risk criteria but has a very low absolute false-consensus event rate, limiting its usefulness for the paired grounding diagnostic and policy evaluation.

Relation corruptions were explored during construction but excluded from judge evaluation because audits revealed unstable fluency, syntax, and semantic naturalness artifacts. This exclusion was made before using Relation in any main judge-panel evaluation.

D.3 Detectable Subset Analysis

False-consensus lift and the paired grounding diagnostic are reported on a detectable corrupted subset to reduce the influence of items that may be difficult or ambiguous even for a stronger external factuality detector. Detectability is estimated using GPT-4o as an analysis-time external detector. GPT-4o is not part of the JuryProbe judge panel, is not used as the grounded verifier, and is not used by the routed policy. A corrupted item is included in the detectable subset if GPT-4o correctly rejects the corrupted claim. This analysis-only filter focuses the false-consensus analysis on corrupted claims whose factual error is externally detectable, while still evaluating whether the reference-free judge panel unanimously accepts them. The detectable subset contains 214, 286, and 296 corrupted examples for Number, Entity, and Attribute, respectively. FN-only correlation is reported over all corrupted claims, while false-consensus rate, false-consensus lift, and the permutation-test p-value are reported on the detectable corrupted subset. The paired grounding diagnostic uses the same detectable corrupted claims in paired reference-free and grounded conditions.

D.4 Alternative Detectability Definitions

The main analysis uses the GPT-4o detectable subset because this filter is independent of both the JuryProbe judge panel and the grounded verifier. To check that the conclusions do not depend on this particular filter, we also recompute the reference-free false-consensus statistics under two alternative subsets: all corrupted claims, and a grounded-detectable subset consisting of corrupted claims for which at least one grounded judge rejects the claim. The grounded-detectable variant is reported only as a robustness check, not as the primary subset for the grounding diagnostic, because it is defined using grounded verifier outputs.

Table 9: Alternative detectability definitions. Across all detectability definitions, the qualitative conclusion remains unchanged: reference-free panel exhibits excess false consensus, while no grounded false consensus is observed.

Family	Subset	N	RF Corr.	RF FC Rate	RF FC Lift	Grounded FC
Number	GPT-4o detectable	214	0.386	0.159	3.13×	0.000
Number	All corrupted	300	0.402	0.193	2.69×	0.000
Number	Grounded-detectable	300	0.402	0.193	2.69×	0.000
Entity	GPT-4o detectable	286	0.393	0.031	18.13×	0.000
Entity	All corrupted	300	0.368	0.030	13.19×	0.000
Entity	Grounded-detectable	300	0.368	0.030	13.19×	0.000

E Grounding Diagnostic with Non-Minimal-Pair Scientific Evidence

Table 10 presents the full metrics for the SciFact grounding diagnostic summarized in Section 5.2: 190 contradicted claims evaluated under reference-free judging, with the benchmark-annotated rationale, and with the full published abstract (274 tokens on average, with no marked rationale span). These references were not constructed by editing the claims and encode semantic rather than single-token contradictions. For example, a false claim states that aPKCz causes tumour enhancement through glutamine metabolism, whereas the evidence shows that *loss* of aPKCz enhances tumorigenesis and characterizes aPKCz as a metabolic tumor suppressor.

Table 10: SciFact grounding diagnostic (190 contradicted / 190 supported claims). Per-judge false-accept rates list the three judges in the order Llama / Qwen / Gemma.

Metric	Reference-free	Benchmark rationale	Full abstract
All-3 false consensus	0.053 (10/190)	0.000 (0/190)	0.000 (0/190)
FC lift	5.56×	0.00×	0.00×
Mean pairwise FN correlation	0.312	0.247	0.108
Per-judge false accept	0.23 / 0.10 / 0.43	0.07 / 0.01 / 0.06	0.06 / 0.02 / 0.06
Clean true accept (per-judge mean)	0.598	0.737	0.791
Final parse failures	1/1,140	0/1,140	0/1,140

F Cross-Family Evaluation: Split-Level Results and Full Statistics

Table 11 presents the complete cross-split statistics underlying Table 5, including the lift values omitted from the latter and the corresponding permutation p-values. Lift is defined as a ratio relative to the independence null q_{ind}; when the marginal false-negative rates are very low, q_{ind} approaches zero on the 95-item calibration halves, and a single false-consensus event can therefore yield an extreme ratio. As a result, cross-split mean lifts become unstable for Attribute, CREAK, and the boundary control; one boundary-control split even has an infinite lift. This instability does not affect the splitwise high-risk decision, which is determined by a per-split conjunction rather than by a cross-split mean. In the negative-control split containing a constant all-zero false-negative vector for one judge, pairwise FN correlations involving that judge are assigned 0.0 under the zero-variance convention used in the implementation, while lift is undefined because the corresponding independence baseline is zero.

Table 11: Full cross-family splitwise statistics (mean ± standard deviation over 10 splits). Lift values marked unstable are dominated by near-zero independence nulls and are not comparable across families.

Family	Flagged	FN Corr	Lift	Permutation p
Number	10/10	0.385±0.028	2.60±0.20	0.000±0.000
Entity	10/10	0.348±0.062	11.44±5.27	0.011±0.016
Attribute	8/10	0.265±0.112	79.55±87.23 (unstable)	0.209±0.417
SciFact	10/10	0.319±0.022	5.57±0.70	0.000±0.000
FEVER-Refutes	8/10	0.306±0.073	23.55±13.18	0.016±0.031
CREAK	8/10	0.318±0.100	92.97±64.29 (unstable)	0.205±0.419
Obvious-Number	6/10	0.330±0.167	611.11±952.88 (+1 inf; unstable)	0.401±0.516
Self-Contained Contradiction	0/10	0.082±0.054	0.00 (9 splits), undef. (1)	1.000±0.000

On CREAK, the reference-free deployment rates are false accept 0.029 ± 0.010 and true accept 0.767 ± 0.024. FEVER-Refutes has no trusted reference in our setup, so neither grounded policy is evaluable there; the Obvious-Number boundary control has no grounded cache, so its grounded-policy rates are analytic intervals rather than measured results and are not quoted as policy performance.

G Distribution-Shift Recalibration Grid

Table 12 presents the full stress-test grid summarized in Section 6.1. The initial calibration relies only on the negative control and stands down in 0/10 splits across all audit sizes. Each cell reports the number of splits (out of 10) flagged by the separate labeled recalibration audit, as a function of the target family's share of false claims in the shifted stream and the audit's labeled-item budget (25, 50, or 95 corrupted items). Stale RF FA denotes the reference-free false-accept rate on the shifted stream while the initial stand-down label is retained. The main-text counts report the 95-item audit column.

This stress test evaluates listed benchmark shifts after a new labeled audit; it does not establish zero-shot safety under arbitrary open-world drift.

Table 12: Recalibration flags (out of 10 splits) by target share and labeled-audit size, with the stale reference-free false-accept rate of the shifted stream.

Target family	Share	Audit 25	Audit 50	Audit 95	Stale RF FA
Number	0%	0	0	0	0.013±0.008
Number	10%	2	7	9	0.054±0.032
Number	25%	6	8	10	0.116±0.038
Number	50%	7	10	10	0.217±0.044
Number	100%	8	10	10	0.429±0.051
Entity	0%	0	0	0	0.013±0.008
Entity	10%	0	1	3	0.021±0.015
Entity	25%	1	4	6	0.033±0.018
Entity	50%	4	4	6	0.063±0.023
Entity	100%	3	6	10	0.111±0.018
SciFact	0%	0	0	0	0.013±0.008
SciFact	10%	2	2	4	0.033±0.018
SciFact	25%	2	3	5	0.072±0.026
SciFact	50%	2	4	7	0.113±0.032
SciFact	100%	4	6	10	0.198±0.031

H SciFact Panels: Fold-level Results and Wilson Intervals

This appendix expands the SciFact panel evaluations of Section 5.6. Table 13 reports pooled out-of-fold rates with Wilson 95% intervals for both panels, and Table 14 reports fold-level calibration statistics. Ground-All-Accepts outcomes are actual grounded-panel decisions with the full published abstract as reference, never gold-label substitutions. A non-flagged fold would exercise the stand-down branch; none occurred.

Table 13: Pooled out-of-fold rates on 190 contradicted / 190 supported SciFact claims, with Wilson 95% intervals.

Panel	Metric	Events	Rate	Wilson 95%
Main	RF false accept	40/190	0.211	[0.159, 0.274]
Main	RF clean accept	113/190	0.595	[0.524, 0.662]
Main	RF all-3 false consensus	10/190	0.053	[0.029, 0.094]
Main	Ground-All-Accepts false accept	3/190	0.016	[0.005, 0.045]
Main	Ground-All-Accepts clean accept	96/190	0.505	[0.435, 0.576]
Larger	RF false accept	44/190	0.232	[0.177, 0.297]
Larger	RF clean accept	139/190	0.732	[0.664, 0.790]
Larger	RF all-3 false consensus	21/190	0.111	[0.073, 0.163]
Larger	Ground-All-Accepts false accept	2/190	0.011	[0.003, 0.038]
Larger	Ground-All-Accepts clean accept	121/190	0.637	[0.566, 0.702]

I BM25 Retrieval Stress-Test Details

Table 15 reports the full reference-source comparison underlying Section 5.5, using the frozen subset (190 clean / 190 corrupted; flagged in 10/10 splits, so the routed rows also coincide with the Ground-All-Accepts rows). Descriptively, the BM25@3 verifier-majority false-accept rate is 0.012 on retrieval hits (83 false claims) and 0.047 on misses (107 false claims). Because hit status is not randomized, this stratification reflects an association rather than a mechanism test.

Table 16 reports the within-split held-out check: the routed policy is recomputed on disjoint deployment halves (40 clean / 40 corrupted per split), while the risk flag is estimated using the calibration half only

Table 14: Fold-level calibration statistics for both SciFact panels (grouped five-fold protocol; all folds flagged).

Panel	Fold	FN corr	All-3 count	Lift	p	RF FA	GAA FA
Main	1	0.296	7	6.34	0.0003	13/38	0/38
Main	2	0.357	10	5.91	0.0003	5/38	0/38
Main	3	0.228	5	4.60	0.0023	12/38	1/38
Main	4	0.372	10	5.98	0.0003	4/38	1/38
Main	5	0.297	8	4.66	0.0003	6/38	1/38
Larger	1	0.413	13	7.60	0.0003	14/38	1/38
Larger	2	0.471	18	6.77	0.0003	7/38	1/38
Larger	3	0.416	15	7.16	0.0003	12/38	0/38
Larger	4	0.471	20	6.28	0.0003	6/38	0/38
Larger	5	0.492	18	6.65	0.0003	5/38	0/38

Table 15: SciFact reference-source stress test (full frozen subset). Surviving RF-unanimous FA counts corrupted claims unanimously accepted reference-free that also pass grounded majority (count in parentheses). Grounded Claims counts items routed to grounded verification.

Policy	False Accept	True Accept	Surviving RF-unanimous FA	Grounded Claims
RF Majority	0.211	0.595	0.053 (10)	0
RF Unanimity	0.053	0.368	0.053 (10)	0
Routed + benchmark rationale	0.026	0.479	0.005 (1)	153
Routed + benchmark full abstract	0.016	0.505	0.005 (1)	153
Routed + BM25@1	0.016	0.295	0.005 (1)	153
Routed + BM25@3	0.021	0.326	0.011 (2)	153

(flagged in 10/10 splits). Every full-subset figure falls within one split-level standard deviation of the held-out mean, and the policy has no fitted parameters beyond the binary flag.

Table 16: SciFact held-out deployment-half policy check versus full-subset rates.

Policy	Held-out FA	Full-subset FA	Held-out TA	Full-subset TA
RF Majority	0.193±0.041	0.211	0.580±0.071	0.595
Routed + rationale	0.028±0.025	0.026	0.490±0.076	0.479
Routed + full abstract	0.010±0.013	0.016	0.502±0.080	0.505
Routed + BM25@1	0.013±0.018	0.016	0.278±0.025	0.295
Routed + BM25@3	0.020±0.023	0.021	0.310±0.027	0.326

J When Is Consensus-Risk Routing Worth It? A Parameterized Utility Analysis

J.1 Model

For a deployment stream of V claims with corrupted fraction π, define the per-claim utility of policy P as

$$U(P) = v_{\mathrm{TP}}\,\mathrm{TP}(P) - c_{\mathrm{FA}}\,\mathrm{FP}(P) - c_{\mathrm{ref}}\,\mathrm{Refs}(P) - c_{\mathrm{LLM}}(P) - c_{\mathrm{cal}}(P)/V,$$

where $\mathrm{TP}(P) = (1-\pi)\,\mathrm{TA}_P$ and $\mathrm{FP}(P) = \pi\,\mathrm{FA}_P$ are obtained from the class-conditional accept rates measured in our experiments ($\pi = 0.5$ on the frozen splits; other class mixes simply reweight these rates under the assumption that the class-conditional rates remain stable). Here, $\mathrm{Refs}(P)$ denotes the measured number of trusted references acquired per claim; $c_{\mathrm{LLM}}(P)$ is the judge-call cost from the run logs (3.78×10^{-6} USD per call, or approximately 10^{-5} USD per claim for every policy); and c_{cal} is the one-time calibration cost, consisting of 300 labeled items plus 0.0034 USD in reference-free calls. This calibration cost is incurred only by the routed policy and amortized over V. Rejections receive zero utility, so any loss in coverage corresponds to forgone v_{TP}.

For an alternative policy ALT, let $\Delta\text{TP} = \text{TP(ALT)} - \text{TP(JP)}$, $\Delta\text{FP} = \text{FP(JP)} - \text{FP(ALT)}$, and $\Delta\text{Refs} = \text{Refs(ALT)} - \text{Refs(JP)} > 0$. JuryProbe-Routed is preferred if and only if

$$c_{\text{ref}} > [\, v_{\text{TP}}\, \Delta\text{TP} + c_{\text{FA}}\, \Delta\text{FP} + \Delta c_{\text{LLM}} + c_{\text{cal}}/V \,]\,/\Delta\text{Refs}.$$

Thus, every decision boundary is linear in $(v_{\text{TP}}, c_{\text{FA}}, c_{\text{ref}})$, with coefficients taken directly from the measured results; no parameters are fitted.

J.2 Measured inputs ($\pi = 0.5$)

In the flagged families, JuryProbe-Routed and Ground-All-Accepts are identical, as verified split by split, so a single entry represents both policies. For Number, RF majority has TA 0.581 / FA 0.427 / 0 references per claim; JuryProbe-Routed has TA 0.581 / FA 0.000 / 0.504 references per claim; and Always Grounded has TA 1.000 / FA 0.000 / 1.000. For Entity, RF majority has TA 0.639 / FA 0.119 / 0 references per claim; JuryProbe-Routed has TA 0.639 / FA 0.000 / 0.379; and Always Grounded has TA 1.000 / FA 0.000 / 1.000.

In the stand-down regime (negative control, not flagged in 10/10 splits), all entries are fully empirical. RF majority and JuryProbe have TA 0.552 / FA 0.013 / 0 references per claim; Ground-All-Accepts has TA 0.552 / FA 0.009 / 0.282; and Always Grounded has TA 1.000 / FA 0.138 / 1.000.

J.3 Break-even boundaries

In the flagged regime, JuryProbe-Routed is preferred to Always Grounded if and only if $c_{\text{ref}} > 0.422\, v_{\text{TP}}$ for Number (split range 0.382–0.466) or $c_{\text{ref}} > 0.290\, v_{\text{TP}}$ for Entity (0.265–0.316). It is preferred to RF majority if and only if $c_{\text{ref}} < 0.423\, c_{\text{FA}}$ for Number or $c_{\text{ref}} < 0.157\, c_{\text{FA}}$ for Entity. In both comparisons, the omitted terms are negligible ($\Delta c_{\text{LLM}} \approx 10^{-5}$ USD and $c_{\text{cal}}/(V\,\Delta\text{Refs})$).

Together, these inequalities define a family-dependent operating band, $0.29\text{–}0.42\, v_{\text{TP}} < c_{\text{ref}} < 0.16\text{–}0.42\, c_{\text{FA}}$. The band is non-empty only when the cost of a false accept is at least approximately $1\times$ (Number) to $1.9\times$ (Entity) the value of a true accept. Below this band, Always Grounded dominates; above it, reference-free operation dominates.

In the stand-down regime, JuryProbe, which uses zero references, is preferred to Ground-All-Accepts if and only if $c_{\text{ref}} > 0.006\, c_{\text{FA}}$, and to Always Grounded if and only if $c_{\text{ref}} > 0.224\, v_{\text{TP}} - 0.063\, c_{\text{FA}}$. The negative coefficient on c_{FA} arises because empirical Always Grounded has a higher false-accept rate (0.138) than the reference-free panel (0.013) on this family; consequently, when $c_{\text{FA}} \geq 3.6\, v_{\text{TP}}$, JuryProbe is preferred regardless of reference cost.

Under the SciFact reference-quality variants, JuryProbe-Routed is preferred to Always Grounded if and only if $c_{\text{ref}} > 0.216\, v_{\text{TP}} - 0.009\, c_{\text{FA}}$ for rationale references or $c_{\text{ref}} > 0.184\, v_{\text{TP}} - 0.009\, c_{\text{FA}}$ for BM25@3.

J.4 Correction of an earlier reported number

An earlier version of this analysis reported the negative control's false-accept rate as 0.013 "versus 0.000 for Always-Grounded." The 0.000 value came from an oracle-label policy simulation in which the Always Grounded decision was assigned the gold label rather than generated by the grounded judge panel, and was mistakenly presented as an empirical result. The executed Always Grounded condition on all 600 control claims (1,800 verdicts; zero parse failures) yields a false-accept rate of 0.138 ± 0.025 and a clean true-accept rate of 1.000 ± 0.000 over the ten held-out splits. On the full frozen set, 42/300 corrupted claims are accepted, while 300/300 clean claims are accepted.

The claim that freely available references make Always Grounded dominant therefore holds in the evaluated Number and Entity settings, but not universally: the control's appended contradictions are absent from, rather than contradicted by, the reference, and a consistency-based grounded check accepts many items that the reference-free panel rejects. All other negative-control results remain unchanged.

J.5 Assumptions and scope

The analysis assumes a linear, risk-neutral, and per-claim-additive utility, with no distribution-free guarantee. TA and FA are class-conditional rates measured on balanced splits; reweighting by π therefore assumes that these rates remain stable across different class mixes. Boundary constants are point estimates based on the means across 10 splits, with split-level ranges reported separately. All rates are specific to the evaluated judge panel, claim families, and reference sources; recalibration is required following any change to the model, prompt, retrieval method, or data.

K Reproducibility

Judge prompts. The reference-free prompt is, verbatim:

```
You are checking a single statement for factual accuracy.
You are given ONLY the statement, with no reference answer.
Statement: {statement}
Is this statement factually correct? Consider any dates, numbers, and claims.
Respond with ONLY one word: true (if fully correct) or false (if it contains any
factual error).
Your answer:
```

The grounded prompt is, verbatim:

```
You are verifying a statement against a trusted reference.
Reference (trusted, correct): {reference}
Statement to check: {statement}
Is the Statement fully consistent with the Reference? Answer ONLY 'true'
(consistent) or 'false' (contradicts the reference).
Your answer:
```

The main panel consists of Llama-3.1-8B-Instruct, Qwen-2.5-7B-Instruct, and Gemma-3-12B-IT, while the pre-specified larger panel consists of Llama-3.3-70B-Instruct, Qwen-2.5-72B-Instruct, and Gemma-2-27B-IT. All judges are queried through the OpenRouter API at temperature 0.0 with a 64-token output limit. Outputs are parsed using a strict `true`/`false` matcher; unparseable responses are recorded as parse failures and retried with the same prompt, with no imputation.

Held-out splits use seeds 1–10 with 150/150 calibration/deployment halves (SciFact uses the grouped five-fold protocol over 190/190 claims). The permutation seed is the same as the split seed, and permutation tests use 3,000 permutations with one-count smoothing. The risk thresholds ($\rho_{\mathrm{FN}} > 0.15$, $L_{\mathrm{FC}} > 1.5$, $p < 0.05$) were frozen on 2026-06-09. SciFact selection was frozen before judging: a label-blind atomic filter was applied to both classes, 190 of 208 available CONTRADICT candidates were sampled without replacement using seed 42, and supported claims were length-matched without model-assisted selection. The Self-Contained Contradiction control appends one of 40 distinct self-contained numeric contradictions (year ordering, magnitude, or arithmetic) to clean claims; FEVER-Refutes consists of claims labeled REFUTES by FEVER annotators; and the CREAK sample was drawn without content filtering.

All datasets, judge verdicts, and grounded verdicts are cached with dated freeze manifests. Grounded caches cover 600 items each for Number and Entity (1,800 verdicts per family), all 600 negative-control claims (1,800 verdicts), and every routed SciFact item under all four reference conditions. Corruption-construction scripts, author audit records, and freeze manifests will be released upon publication. No oracle replacement, gold-label fallback, or missing-item imputation is used in any reported policy result.