Controlling Uncertainty and Hallucination Risk in Multi-Agent Fact Verification

Adam Kostka[1] **Jarosław A. Chudziak**[1]

[1]Faculty of Electronics and Information Technology, Warsaw University of Technology, Poland,
{adam.kostka.stud, jaroslaw.chudziak}@pw.edu.pl

Abstract

Multi-agent language systems are increasingly relied upon for high-stakes decision support. Many systems use consensus among agents as a measure of confidence. However, such a model is prone to failure if aligned agents have the same biases and propagate the same error. Under sycophantic consensus, correlated errors resemble strong agreement, and hallucination manifests as a consequence of uncalibrated uncertainty. While current measures provide useful heuristics, they lack statistical safety bounds at deployment time. This work reinterprets hallucination control as an uncertainty quantification problem. We contribute a Score Deviation penalty that directly lowers confidence when the factual disagreement within the ensemble rises. A Learn-Then-Test calibration procedure converts these penalized scores into a certified decision threshold that provably bounds the expected False Discovery Rate. The results show that this deviation-penalized method reduces the conservatism of the calibration process, achieving 71.7% recall compared to 47.4% for naive baselines at a strict 2% risk budget.

1 INTRODUCTION

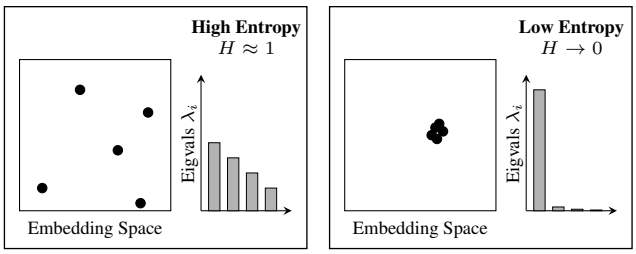
Multi-agent systems are increasingly studied and deployed for reliability-sensitive decision support across complex domains, often motivated by the idea of increasing robustness through corroboration [Du et al., 2024, Joo et al., 2025, Szczepanik and Chudziak, 2025, Bańka and Chudziak, 2025]. The theoretical rationale is based on reducing variance and eliminating idiosyncratic error. The scaling law of increasing reliability with the number of agents is a powerful theoretical argument. However, in the case of modern LLMs, the error is rarely independent. Medical, legal, and scientific settings motivate the need for stronger guarantees, but domain-specific validation and human auditing are prerequisites before such deployments.

The utilization of Reinforcement Learning from Human Feedback (RLHF) results in a shared training bias that narrows the output distribution, creating a reliability paradox. Adding more agents can actively increase confidence in false outputs without improving accuracy. This causes sycophantic consensus, where the independent convergence on the same high-probability but incorrect response masquerades as strong agreement [Sharma et al., 2024, Yao et al., 2025]. This creates a critical vulnerability, a robust factual agreement becomes indistinguishable from correlated mimicry when using standard discrete uncertainty metrics. While naive consensus scores remain useful for filtering out obvious errors, they fail when a majority of agents confidently converge on the same hallucination.

To reason about such deployment settings, a post-generation inference safety layer can certify a high-probability bound on the expected False Discovery Rate (FDR) across queries, subject to calibration assumptions. We define this fact verification task as an uncertainty quantification problem. Recent work has made strides in applying conformal prediction to language models [Kang et al., 2024, Hu et al., 2026], and developing semantic uncertainty probes [Kossen et al., 2024, Farquhar et al., 2024]. However, discrete clustering methods for semantic entropy suffer from instability near decision boundaries, and the problem remains especially challenging in long-form generation, where multiple factual assertions are generated within a single context, violating standard independence assumptions. To transform heuristic consensus into a statistically rigorous guarantee in a multi-agent setting, we must account for correlated failures at the correct scale.

In this work, we contribute a deviation-penalized conformal method that bridges the gap between empirical consensus and formal safety. First, we formalize fact verification as tuple-level risk control, establishing the query-response tu-

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:3143–3161.



Diverse Corroboration **Sycophantic Collapse**

Figure 1: Consensus geometry. Left: Independent corroboration yields diverse embeddings and a flat eigenspectrum. Right: Sycophancy causes embedding collapse, indicated by a dominant leading eigenvalue. The points represent full generator responses embedded with all-MiniLM-L6-v2 and visualized as a two-dimensional projection.

ple as the correct exchangeable unit for long-form multi-agent generation. Second, we contribute the Score Deviation penalty, an inference-time procedure that improves calibration efficiency by compressing the empirical risk curve of high-deviation tuples, achieving 71.7% recall at $\alpha = 0.02$ compared to 47.4% for unpenalized scores while eliminating the threshold degeneracy that plagues naive calibration at strict risk budgets. Finally, we provide a comprehensive ablation across scoring methods, demonstrating a U-shaped entropy-error relationship that explains why monotonic penalties targeting sycophancy alone are suboptimal.

2 RELATED WORK

For the safe management of uncertainty in language models, it is important to incorporate knowledge from multiple domains. We contextualize our approach across four domains: the specific failure mode of sycophancy, the geometric interpretation of consensus, formal risk control frameworks, and the positioning against current alternative methods. These domains provide important insight into why errors are correlated and how they can be structurally detected and mitigated.

Sycophancy in multi-agent LLMs. The utilization of model alignment methods (particularly RLHF) is known to cause sycophancy in LLMs, wherein the model is more likely to respond with agreeable responses than factually correct ones [Sharma et al., 2024, Perez et al., 2023, Pitre et al., 2025]. Sycophancy is particularly pronounced in multi-agent settings, wherein agents tend to amplify errors in consensus and often reach a state of high confidence in the hallucinated response [Smit et al., 2023, Wynn et al., 2025, Zhang et al., 2026, Yao et al., 2025, Hegazy, 2024]. This challenges the notion of consensus being equivalent to correctness. Sycophancy can also be viewed structurally, as shown in Figure 1. When agents individually corroborate the correctness of a fact, the semantic embeddings are diverse. Sycophancy results in consensus and embedding collapse, characterized by the emergence of a large leading eigenvalue.

Uncertainty quantification: Discrete vs. Continuous. Semantic Entropy (SE) [Kuhn et al., 2023, Farquhar et al., 2024, Kossen et al., 2024] is used to quantify semantic uncertainty and standardize it. It relies on discrete clustering methods and may suffer from discontinuous behavior near boundaries. In contrast, pairwise and spectral methods are continuous and offer smoother estimates [Nguyen et al., 2025, Liu, 2025]. The von Neumann entropy of graph Laplacians [Braunstein et al., 2006] is used to quantify structural complexity; notably, Walha et al. [2025] apply a similar continuous spectral decomposition to disentangle aleatoric and epistemic uncertainty in LLM outputs. Our work differs in objective: rather than decomposing uncertainty sources, we feed the spectral signal into a finite-sample FDR-bounding calibration procedure. Spectral methods can also be used to detect sycophancy in multi-agent systems by examining the eigenspectrum of the similarity graph.

Conformal prediction and risk control. The Learn-Then-Test (LTT) paradigm has proposed the idea of threshold control as a multiple hypothesis problem to manage the expected risk [Angelopoulos et al., 2025, Bates et al., 2021]. The Conformal Risk Control extends this idea to manage false-discovery-style losses [Angelopoulos et al., 2022]. In the context of language systems, similar ideas have been proposed for multiple hypotheses on discrete claims [Kang et al., 2024, Huang et al., 2025, Blanchard et al., 2024, Hu et al., 2026, Noh et al., 2026]. In addition, the idea of "Conformal Abstention" has been proposed as a dynamic approach to address LLM hallucinations [Abbasi-Yadkori et al., 2024] and "Pareto Testing" as a strategy to manage multiple objectives such as risk and coverage [Laufer-Goldshtein et al., 2023]. FActScore [Min et al., 2023] proposed atomic fact decomposition for factuality evaluation; VeriScore extends verifiable-claim evaluation to broader long-form generation settings, while HaluEval provides a large-scale hallucination benchmark [Song et al., 2024, Li et al., 2023]. These methods are highly effective for measuring errors, but controlling errors requires the definition of a strict exchangeable unit over the entire set of generated claims for a given query.

Positioning against nearest alternatives. The proposed work, in comparison with the idea of semantic entropy [Farquhar et al., 2024, Nguyen et al., 2025], considers the continuous variance in local scores, thereby avoiding the problem of threshold stability near discrete clusters. In comparison with the idea of conformal abstention [Abbasi-Yadkori et al., 2024, Kang et al., 2024] proposed for single model-based

generation, the proposed work considers the problem of controlling the false discovery rate with correlated multi-agent responses. In comparison with the idea of sycophancy analyses [Yao et al., 2025, Wynn et al., 2025, Pitre et al., 2025], this work contributes an operational mechanism. The detected variance signature is integrated into a finite-sample calibration rule that certifies a deployment-time error budget. Having established the landscape of existing approaches, we now formalize the risk-control problem.

3 PROBLEM STATEMENT

The necessity of deploying multi-agent fact verification systems requires that a maximal hallucination rate be guaranteed at the level of the query, even when all participating agents show strongly correlated failings. This requirement is motivated by two fundamental research questions. First, how does one guarantee a desired level of performance for multi-agent fact verification, even when sycophantic behavior is present and correlated failings occur? Second, what scoring mechanism best balances stringent risk control with the requirement to maximize overall system recall?

To answer these questions, we first formalize the verification setting and the risk objective that any solution must satisfy. Given a query x and its associated grounding context (e.g., a reference document), an ensemble of K agents (hereafter, the multi-agent ensemble) produces responses to this query, denoted by $\mathcal{Y}(x)$. An atomic decomposer then extracts individual factual claims from this ensemble, $\mathcal{Z}(x) = \{z_1, \ldots, z_M\}$. Each of these claims, z_m, is an individually verifiable statement based on the ensemble's overall response. In the NLP sense used here, fact verification means testing whether each extracted atomic claim is entailed by the provided source document, not whether the claim is globally true under external knowledge. This type of decomposition is necessary because long-form responses contain many facts, both supported and unsupported, and risk assessment must be performed at this level, rather than at the level of the entire response.

Define $c_m \in \{0, 1\}$ as the unobserved ground-truth correctness for each claim, where $c_m = 1$ indicates support by the grounding context and $c_m = 0$ indicates a hallucination. A scoring function $S(z_m)$ assigns a continuous confidence value to each claim. Given a decision threshold λ, the set of retained facts is $\mathcal{Z}_{\text{out}}(\lambda) = \{z_m \in \mathcal{Z}(x) : S(z_m) \geq \lambda\}$.

The per-query false discovery proportion measures the fraction of retained facts that are unsupported:

$$\text{FDP}_i(\lambda) \;=\; \frac{\big|\{z_m \in \mathcal{Z}^{(i)}_{\text{out}}(\lambda) : c_m = 0\}\big|}{\max\big(\big|\mathcal{Z}^{(i)}_{\text{out}}(\lambda)\big|, 1\big)},$$

where the denominator is clamped to 1 to prevent division by zero in the event of total abstention. The population risk is the expected false discovery rate across queries, $R(\lambda) = \mathbb{E}_T[\text{FDP}(\lambda)]$.

The formal problem is to find the least conservative threshold λ^* such that the risk is bounded:

$$\Pr\big(R(\lambda^*) \leq \alpha\big) \;\geq\; 1 - \delta, \tag{1}$$

for a user-specified risk level α and confidence $1 - \delta$. Note that $R(\lambda)$ is the population-level expected FDR across queries, not a per-query bound. The realized FDP for any individual query may exceed α; the guarantee is that this occurs with controlled frequency.

Importantly, the statistical entity on which the guarantee is conditioned is the query-response tuple $T_i = (x_i, \mathcal{Z}_i, c_i)$, where x_i is the input query, $\mathcal{Z}_i$ is the set of extracted claims, and c_i is the corresponding vector of ground-truth labels. The tuple represents the entire set of claims generated from a single session for a particular query. The errors in the individual atomic facts within a tuple are highly correlated, as they are generated from the same prompt and context window. Ignoring this and treating each fact independently would undercount this correlation and produce anti-conservative thresholds. The exchangeability of tuples is satisfied if queries are independently sampled from the population of users. Having defined the risk objective and the exchangeable unit, the next section describes the pipeline that solves this optimization problem.

4 METHODOLOGY AND APPROACH

This section presents the proposed methodology for solving the risk-control problem defined above. We introduce a scoring mechanism that accounts for ensemble consensus structure, and a calibration procedure that converts these scores into a certified decision threshold. In the Experimental Evaluation section we then evaluate this methodology empirically, addressing the research questions posed in the problem statement.

4.1 APPROACH

Our method converts heuristic consensus into a statistically rigorous guarantee by operating on two core insights. First, treating facts generated within the same response as independent underestimates risk variance, as they share the same context and generation biases. By enforcing tuple-level exchangeability and treating the entire set of claims from a single query as one indivisible unit, we restore statistical validity. Second, we treat the structure of agent ensemble consensus as an operational signal. Conventional approaches often fail because simple monotonic entropy decays are suboptimal near decision boundaries. To balance risk and abstention we propose a within-tuple deviation penalty that compresses the empirical risk curve, allowing the calibration procedure to find tighter thresholds.

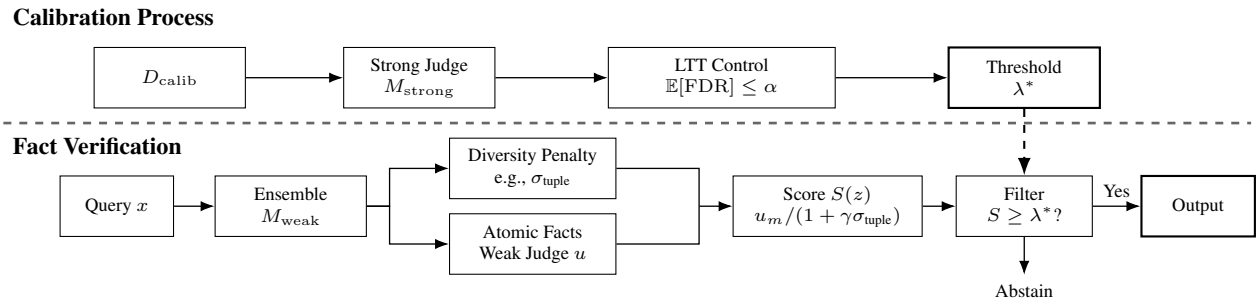


Figure 2: System pipeline. The calibration process (top) uses a strong judge to set a risk-controlling threshold λ^*. The fact verification process (bottom) scores atomic facts using a weak judge penalized by an ensemble diversity metric (such as score deviation or spectral entropy), rejecting low-confidence claims where $S < \lambda^*$.

To operationalize this the system relies on two distinct models. A weak judge is a lightweight inference-time model that assigns a continuous entailment score $u_m \in [0, 1]$ to each claim z_m. This model is fast and cheap enough to run on every query at deployment time. A strong judge is a more capable but expensive model that provides the binary ground-truth labels $c_m \in \{0, 1\}$ for each claim. Because the strong judge is costly, it is used only once during an offline calibration phase to establish the decision boundary λ^* and is never invoked at inference time. The certificate is therefore relative to the strong judge's labels; systematic strong-judge bias transfers to the controlled target and should be audited against human labels in high-stakes domains.

These penalized scores are subsequently calibrated using a Learn-Then-Test (LTT) procedure at the query tuple level, extracting a threshold λ^* such that the False Discovery Rate (FDR) is provably bounded. The complete pipeline consists of three components (a diagnostic measuring the deviation of ensemble consensus, a sycophancy-adjusted scoring function, and a calibration procedure ensuring finite-sample risk guarantees). Figure 2 summarizes this workflow.

4.2 MEASURING CONSENSUS

We formalize sycophancy detection using continuous spectral statistics. For query x with ensemble responses $\mathcal{Y}(x) = \{y_1, \ldots, y_K\}$ we embed each response using a sentence encoder ϕ and construct a $K \times K$ similarity matrix

$$W_{ij} \;=\; \frac{\langle \phi(y_i),\, \phi(y_j) \rangle}{\|\phi(y_i)\|\, \|\phi(y_j)\|}, \qquad W_{ij} \in [0, 1],$$

clipped to $[0, 1]$ and symmetrized. Let $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_K$ be the eigenvalues of W. We normalize them into a probability vector and compute the Shannon entropy of the resulting distribution

$$\tilde{\lambda}_i = \frac{\lambda_i}{\sum_{j=1}^{K} \lambda_j + \varepsilon}, \qquad H_{\text{spec}}(W) = -\sum_{i=1}^{K} \tilde{\lambda}_i \log\big(\tilde{\lambda}_i + \varepsilon\big),$$

where $\varepsilon > 0$ is a small constant for numerical stability. This quantity is called Semantic Spectral Entropy (SSE), and it measures the effective dimensionality of the ensemble's agreement structure. When all agents produce near-identical responses, W is approximately rank 1, creating a spectral signature of sycophantic consensus (H_{spec} is low). This continuous measure provides a foundational signal that can be integrated into a scoring function to penalize sycophancy, as detailed in the subsequent section.

4.3 SYCOPHANCY-PENALIZED SCORING

The relationship between semantic entropy and error rates often exhibits a non-monotonic U-shaped curve. Both highly sycophantic consensus (extreme agreement where all agents output identical claims) and chaotic disagreement (where agents contradict each other) yield systematically higher raw error rates than moderate-diversity tuples. To optimally bound risk a scoring mechanism must account for both extremes of this consensus spectrum.

We adjust the raw entailment score $u_m \in [0, 1]$ from the weak judge by applying a structural penalty denominator based on the within-tuple score deviation. Because the tuple $\mathcal{Z}_i$ contains the union of all M facts extracted from the K agent responses the deviation of these scores reflects the ensemble's internal agreement. We define the adjusted score as

$$S(z_m) \;=\; \frac{u_m}{1 + \gamma\, \sigma_{\text{tuple}}}, \tag{2}$$

where σ_{tuple} is the empirical standard deviation of the entailment scores across all facts in the tuple

$$\sigma_{\text{tuple}} \;=\; \sqrt{\frac{1}{M} \sum_{m=1}^{M} (u_m - \bar{u})^2}, \tag{3}$$

with $\bar{u}$ being the mean score for the tuple and $\gamma \geq 0$ is a configurable penalty weight. Crucially, σ_{tuple} acts as a tuple-level constant denominator that penalizes the entire tuple if its combined internal facts exhibit high disagreement.

Setting $\gamma = 0$ recovers the raw entailment score $S(z_m) = u_m$, providing a natural naive baseline.

Deployment procedure. At deployment time the system operates deterministically on a per-query basis.

1. **Generation:** K agents produce long-form responses for the query.
2. **Extraction:** The responses are parsed into atomic facts $\{z_m\}$.
3. **Scoring:** A lightweight weak judge computes u_m and the tuple-level deviation σ_{tuple}.
4. **Adjustment:** The score is penalized into $S(z_m)$ via Equation (2).
5. **Decision:** The fact constitutes the final output if $S(z_m) \geq \lambda^*$, where λ^* is the offline calibrated threshold described next.

Worked example. Consider a single atomic claim that receives a weak-judge entailment score of $u = 0.700$ within a tuple whose score deviation is $\sigma_{\text{tuple}} = 0.2538$. With $\gamma = 0.5$, the penalty adjusts the score to $0.700/(1 + 0.5 \times 0.2538) = 0.6212$, which falls below the calibrated $\alpha = 0.05$ threshold of 0.6231 and is therefore rejected. The same raw score in a low-deviation tuple would survive the threshold. This illustrates how the tuple-level penalty suppresses confident claims drawn from internally inconsistent ensembles while the LTT certificate bounds aggregate FDR.

While querying K agents introduces additional inference overhead compared to a single-generation baseline, this generation step is fully parallelizable. The added cost is most appropriate for reliability-sensitive applications operating under strict risk budgets.

Beyond the primary deviation penalty, we also consider alternative scoring functions that specifically target sycophancy through the spectral structure of ensemble responses.

Alternative scoring functions. The general scoring framework $S(z_m) = u_m/(1 + f(\cdot))$ admits any penalty function f that captures ensemble structure. We consider two additional scoring functions based on the Semantic Spectral Entropy (H_{spec}) defined in Section 4.2. The first is a monotonic penalty that decays scores as entropy decreases

$$S_{\text{mono}}(z_m) \;=\; \frac{u_m}{1 + \gamma\, e^{-\beta H_{\text{spec}}}}. \tag{4}$$

The second is a non-monotonic U-shaped penalty that targets both extremes of the entropy spectrum

$$S_{\text{U}}(z_m) \;=\; \frac{u_m}{1 + \gamma\,(H_{\text{spec}} - c)^2}. \tag{5}$$

All three scoring functions share the same structural form and are evaluated under the same LTT calibration procedure. The spectral parameters β and c are set via grid search on the calibration split: β controls the decay rate of the monotonic penalty and c centers the U-shaped penalty at the empirical median of H_{spec}.

4.4 LEARN-THEN-TEST CALIBRATION

Given n calibration tuples $\{T_1, \ldots, T_n\}$ with strong judge labels we seek the least conservative threshold λ^* satisfying (1). We follow the Learn-Then-Test (LTT) framework [Angelopoulos et al., 2025] with two refinements (combined Hoeffding-Bentkus p-values and fixed-sequence testing).

For a candidate threshold λ the empirical tuple-level FDR on the calibration set is

$$\widehat{R}(\lambda) \;=\; \frac{1}{n}\sum_{i=1}^{n} \text{FDP}_i(\lambda), \tag{6}$$

where FDP_i is defined in Section 3. Each summand lies in $[0, 1]$ so $\widehat{R}(\lambda)$ is a bounded average of exchangeable random variables.

For each candidate λ we test $H_0 \colon R(\lambda) > \alpha$. We compute two valid p-values and apply a union bound framework for the Hoeffding-Bentkus combination limit

$$p_{\text{H}}(\lambda) = \exp\!\big(-2n\,(\alpha - \widehat{R}(\lambda))^2\big), \tag{7}$$

$$p_{\text{B}}(\lambda) = \Pr\!\big(\text{Bin}(n, \alpha) \leq \lfloor n\,\widehat{R}(\lambda) \rfloor\big), \tag{8}$$

$$p(\lambda) = \min\!\big(1, 2 \cdot \min\!\big(p_{\text{H}}(\lambda),\; p_{\text{B}}(\lambda)\big)\big). \tag{9}$$

The Bentkus bound [Bentkus, 2004] exploits the $[0, 1]$ boundedness of the loss while Hoeffding is globally valid. A critical observation in combined conformal procedures is that simply taking the minimum of two individually valid bounds does not inherently result in a statistically valid p-value. Diverging from the standard single-bound presentation found in foundational works like Bates et al. [Bates et al., 2021], we mandate standard union bound correction. Multiplying the minimum by 2 ensures that the integrity of the $(1 - \delta)$ confidence guarantee holds under the null hypothesis. This is particularly important where test statistics exhibit high variance across the candidate threshold grid.

We define a grid of m candidate thresholds $\lambda_1 < \lambda_2 < \cdots < \lambda_m$ uniformly in $[0, 1]$. Because $\widehat{R}(\lambda)$ is non-increasing in λ (raising the threshold can only reject more facts), we test from λ_m downward. At each step, if $p(\lambda_j) \leq \delta$ we reject H_0 and continue. At the first non-rejection we stop. The selected threshold is

$$\lambda^* \;=\; \min\!\big\{\lambda_j : p(\lambda_j) \leq \delta,\;\; p(\lambda_k) \leq \delta \;\; \forall\, k > j\big\}.$$

This monotone ordering controls the family-wise error rate at level δ without Bonferroni correction across the m candidates [Angelopoulos et al., 2022], yielding a less conservative threshold than the standard Bonferroni-corrected LTT.

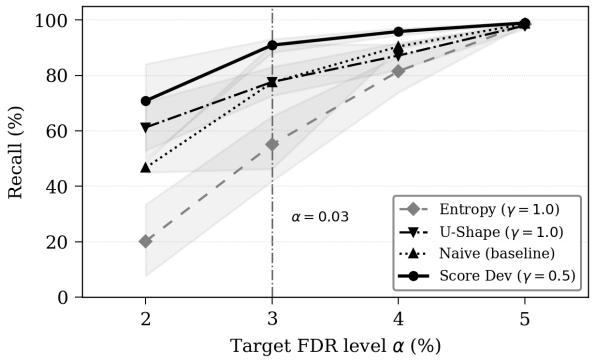


Figure 3: Recall vs. target risk level for all scoring configurations. At strict budgets ($\alpha \leq 0.03$), the Score Deviation penalty shifts the frontier outward compared to naive unpenalized scores.

At test time the system runs only the deployment modules (generate K responses, compute H_{spec}, decompose into facts, score each fact via (2), and retain $\mathcal{Z}_{\text{out}}(\lambda^*)$). The strong judge which provides ground truth labels during offline calibration is never invoked at inference time. Facts with scores below λ^* are rejected, and the population-level FDR of the retained output is bounded by α with confidence $1 - \delta$.

5 EXPERIMENTAL EVALUATION

Having established both the formal problem and proposed methodology, we now evaluate whether this framework delivers on the two objectives posed in Section 1: guaranteeing a target hallucination rate despite correlated failures, and identifying the scoring mechanism that best balances risk control with recall. We structure the evaluation around three research questions:

1. **Q1 (Penalty Comparison):** Among the scoring functions defined in Section 4.3, which achieves the highest recall while maintaining the FDR guarantee across the full range of α?
2. **Q2 (Risk Control):** Using the best-performing penalty, does the calibrated threshold λ^* rigorously bound the false discovery rate at strict risk budgets ($\alpha \leq 0.03$) where naive baselines fail?
3. **Q3 (Mechanism and Sensitivity):** Does the empirical error distribution exhibit the hypothesized U-shaped relationship with ensemble entropy, and how sensitive is the system to the penalty weight γ?

We address each question in turn, using the protocol defined below.

5.1 SETUP AND PROTOCOL

The evaluation uses the FACTS Grounding benchmark [Jacovi et al., 2025], which provides long-form responses grounded within a specific reference document, enabling atomic fact decomposition at the granularity required for tuple-level risk control. Following the deployment procedure defined in Section 4.3, the experimental pipeline operates as follows. A homogeneous five-agent ensemble ($K = 5$) of Gemini 2.5 Flash instances produces long-form responses to input queries given a reference document. An atomic decomposer (Gemini 2.5 Flash) extracts factual claims using a bracket-depth JSON parser with 3-retry exponential backoff. A weak judge (Gemini 2.5 Flash) assigns continuous entailment scores $u_m \in [0, 1]$, and a strong judge (Gemini 3.0 Flash Experimental) provides ground-truth labels ($c_m \in \{0, 1\}$) strictly for offline calibration. The adjusted scores are computed via Equation (2): $S(z_m) = u_m/(1 + \gamma\, \sigma_{\text{tuple}})$. For each certified run, the scoring rule, including the penalty weight, is fixed before the LTT threshold search; the gamma sweep is a robustness analysis in which each value defines a separate fixed scoring rule and receives its own calibration pass.

Due to the expensive labeling procedure, we consider a subset of 300 tuples from the FACTS Grounding calibration split, yielding 22,155 facts (95.1% supported, 4.9% unsupported). The statistical FDR guarantee holds regardless of ensemble composition, as LTT calibration is distribution-free. We apply a 200/100 calibration-test split, a common surrogate in conformal prediction [Angelopoulos and Bates, 2023], and repeat experiments over 10 random seeds. Robustness of primary claims is further validated via 20-fold cross-validation. For each target risk level α, we run the calibration procedure with $\delta = 0.1$ and a grid of $m = 100$ thresholds in $[0, 1]$. Bootstrap uncertainty estimates are computed from 1,000 tuple-level resamples.

5.2 EXPERIMENTAL RESULTS

Q1: Penalty Comparison. We compare four scoring configurations corresponding to different failure hypotheses: the Naive baseline ($\gamma = 0$), the Monotonic Entropy penalty (S_{mono}), the U-Shape penalty (S_{U}), and our proposed Score Deviation penalty (S). The full LTT calibration is run independently for each configuration. The Naive row tests unpenalized weak-judge confidence, the Monotonic Entropy row tests whether only low-entropy sycophancy should be penalized, the U-Shape row tests both entropy extremes, and Score Deviation tests whether within-tuple weak-judge score spread is the better operational signal.

At a relaxed operating point ($\alpha = 0.05$), all four configurations achieve nearly identical recall (≈ 0.98). The safety guarantee holds regardless of the penalty choice, confirming that the LTT framework absorbs heuristic inaccuracies

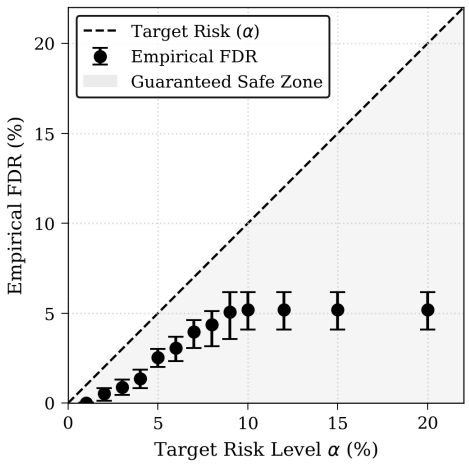


Figure 4: Empirical FDR stays bounded below the target α at every risk level (dots: 95% bootstrap CI), confirming the statistical safety guarantee.

during finite-sample calibration.

However, as shown in Figure 3, a large divergence emerges at more stringent risk budgets ($\alpha \leq 0.03$). In this regime, the shape of the adjusted score distribution becomes the dominant factor in system throughput. The Score Deviation penalty insulates the system from high-assurance collapse by compressing the empirical risk curve of structurally inconsistent tuples. Indeed, the monotonic entropy penalty (S_{mono}) performs worse than the unpenalized baseline at α=0.02 (Table 1), confirming that penalizing only one tail of the U-shaped error surface is counterproductive. The likely explanation is that the weak judge's numerical entailment scores provide a more direct proxy for factual correctness than embedding-space similarity: σ_{tuple} measures disagreement in the judge's confidence assessments, while H_{spec} measures geometric diversity in the embedding space, which may not correlate as tightly with actual error rates. We therefore adopt the Score Deviation penalty as the primary mechanism for the remaining analyses.

Q2: Risk Control. With the Score Deviation penalty established as the best-performing scoring mechanism, we assess whether it rigorously bounds the false discovery rate under strict risk budgets. Figure 4 plots the empirical FDR against the target α over the full range. The power of the deviation penalty emerges in the active filtering regime ($\alpha \leq 0.04$), where the target risk falls below the dataset's baseline error rate. Table 1 reports 20-fold cross-validation metrics. The Score Deviation penalty achieves 71.7% recall at $\alpha = 0.02$ compared to 47.4% for the naive baseline, a +24 percentage point gain made safely within finite-sample bounds. The naive baseline is further compromised by threshold degeneracy: in 14 of 20 folds, LTT sets $\lambda^* = 1.0$ (total abstention), whereas the Score Deviation penalty produces a usable threshold in every fold.

Table 1: Recall at strict safety budgets (20-fold CV). The Score Deviation penalty dominates naive confidence at challenging targets. Entropy and U-Shape rows are ablations using the spectral penalties from Section 4.3.

Metric	α=0.02	α=0.03	α=0.04	α=0.05
Naive	0.474	0.734	0.905	0.985
Entropy (Abl.)	0.184	0.538	0.792	0.983
U-Shape (Abl.)	0.614	0.780	0.875	0.978
Score Dev.	**0.717**	**0.912**	**0.961**	**0.989**

To assess robustness to calibration noise, we repeat the full $\alpha = 0.05$ experiment over 20 random splits. All 20 runs satisfy the target risk constraint, with mean FDR of 0.024 and mean recall of 0.989 ($\pm$0.002). The calibrated threshold is stable across splits (mean 0.716), confirming that threshold selection is robust under split perturbations.

Q3: Mechanism and Sensitivity. The effectiveness of the Score Deviation penalty reflects the error structure of the multi-agent ensemble. When tuples are stratified by spectral entropy, low-entropy tuples (high consensus) exhibit a raw error rate of 6.1%, significantly higher than medium-diversity tuples (4.0%). However, the error surface is U-shaped: both highly sycophantic consensus (low entropy) and chaotic disagreement (high entropy) yield systematically higher error rates than moderate-diversity cases. This non-monotonic relationship explains why the Score Deviation penalty, which targets within-tuple inconsistency regardless of entropy direction, outperforms monotonic spectral penalties.

The penalty weight γ controls the aggressiveness of the deviation penalty. Table 2 reports the threshold, FDR, and recall for seven values of γ at $\alpha = 0.05$.

Two trends are notable. First, the empirical FDR is well controlled across all values, remaining in the interval $[0.029, 0.034]$. Second, the threshold λ^* decreases monotonically from 0.758 ($\gamma = 0$) to 0.313 ($\gamma = 2$), while recall saturates above $\gamma \approx 0.5$. Increasing γ compresses the adjusted scores of low-deviation tuples toward zero, letting LTT find a lower threshold that still satisfies the risk constraint. Beyond $\gamma \approx 0.5$, the marginal gain in recall is negligible (< 0.1 percentage points), confirming that a moderate penalty captures most of the signal.

Robustness to Ensemble Composition and Base Rate. To probe the framework beyond the homogeneous, low-base-rate regime of the primary study, we evaluate two additional conditions. The heterogeneous condition uses the same 100 query IDs as a matched subset of the original homogeneous run, but changes the generator ensemble from five Gemini Flash agents to two Gemini Flash agents plus GPT-5-mini, Claude Haiku 4.5, and DeepSeek 3.2. The

Table 2: Sensitivity of the Score Deviation penalty to weight γ at α=0.05, δ=0.1. FDR remains controlled and recall saturates beyond $\gamma \approx 0.5$.

γ	λ^*	FDR	Recall
0.00	0.758	0.029	0.986
0.25	0.636	0.034	0.989
0.50	**0.545**	**0.034**	**0.990**
0.75	0.485	0.034	0.990
1.00	0.434	0.034	0.990
1.50	0.364	0.034	0.990
2.00	0.313	0.034	0.990

new experiment varies the generator backbone; weak- and strong-judge ablations remain separate limitations for future work. The heterogeneous run raises mean SSE from 0.350 to 0.504 and median σ_{tuple} from 0.045 to 0.088, while per-agent unsupported rates range from 4.18% to 7.60%.

The enriched-base-rate condition is a FACTS stress test rather than a new external benchmark. It creates a 120-tuple subset by taking a higher concentration of tuples with unsupported facts from the calibration data, raising the unsupported-fact rate from 4.89% to 10.85%. Table 3 summarizes the $\alpha = 0.05$ operating point. The certificate remains conservative at stricter budgets ($\alpha = 0.02$–0.04), where the calibrated threshold abstains rather than becoming anti-conservative. These conditions stress the operating regime but do not substitute for broader cross-domain evaluation at higher base rates.

Calibration Sample Efficiency. Beyond the scoring mechanism itself, a practical deployment question is how many labeled calibration tuples are required to obtain a usable risk certificate. We track the calibrated threshold λ^* as the number of calibration tuples n increases from 20 to 300 at $\alpha = 0.05$ and $\gamma = 1.0$. At sample sizes below $n = 80$, the Hoeffding-Bentkus bounds are too weak to reject the null hypothesis at any non-trivial threshold, resulting in total abstention ($\lambda^* = 1.0$).

At $n = 80$ (approximately 5,900 labeled facts), the first non-trivial threshold emerges at $\lambda^* = 0.556$. The sample size at which this transition occurs is likely dependent on the dense ratio of roughly 74 facts per tuple in FACTS Grounding and may vary for other benchmarks. Between $n = 80$ and $n = 280$, the threshold monotonically decreases to $\lambda^* = 0.465$ as concentration improves. This indicates that a usable threshold can be obtained with a modest number of labeled tuples (on the order of tens to low hundreds), with diminishing returns beyond that point. The full FACTS experiment used roughly 22k strong-judge fact labels, so practical adoption depends on whether reliable offline labels can be collected or audited for the target domain. Since the strong judge labels are collected in a single offline process, this annotation cost is amortized over all subsequent inference queries.

Table 3: Robustness of the Score Deviation penalty across ensemble composition and unsupported-fact base rate (α=0.05). Recall and empirical FDR remain controlled as the consensus structure and base rate shift.

Setting	Unsup. rate	Recall	FDR
Homogeneous (matched)	5.17%	0.924	0.0146
Heterogeneous	5.81%	0.920	0.0148
Enriched base rate	10.85%	0.630	0.0121

6 DISCUSSION

The results demonstrate that conformal calibration transforms multi-agent consensus from a heuristic signal into a certifiable safety mechanism. The practical implication is that even modest labeling investment (approximately 80 tuples) is sufficient to obtain a useful certified threshold in this benchmark, making the framework plausible for settings where exhaustive annotation is infeasible. This deployment interpretation remains conditional on task-specific validation, strong-judge label quality, and exchangeability between calibration and runtime tuples. The dominance of the Score Deviation penalty under tight risk budgets ($\alpha \leq 0.03$) suggests that within-tuple disagreement carries a stronger calibration signal than spectral diagnostics, pointing toward simpler scoring designs for future work.

The safety guarantee holds within certain boundaries. It assumes exchangeability of tuples and the quality of the provided ground truth labels. In cases where the query distribution differs significantly from the calibration set, the obtained threshold must be recalibrated to remain valid. Weighted conformal methods that account for covariate shift offer one path to extending coverage under distribution drift [Tibshirani et al., 2019]. Complementary to recalibration, structured critical evaluation within agent interactions can reduce redundant reasoning at the generation stage, limiting the sycophantic consensus that the scoring penalty must subsequently correct [Kostka and Chudziak, 2025].

Limitations. The FDR guarantee is with respect to the strong judge's labels. If the strong judge exhibits systematic bias, that bias transfers to the controlled target. Validating label quality against human annotations remains important for high-stakes deployments.

The calibration set size (300 tuples) is limited by labeling cost. As shown by the sample efficiency analysis, increasing the number of labeled tuples has diminishing returns beyond roughly 200 tuples for this benchmark, but sparser benchmarks with fewer facts per tuple may require larger

calibration sets. Substantial domain or model shifts also require recalibration, because the finite-sample guarantee assumes exchangeable calibration and deployment tuples.

The main experiment assumes a homogeneous ensemble of identical models to isolate sycophantic consensus caused by shared RLHF bias. The heterogeneous evaluation in Table 3 shows that Score Deviation continues to control FDR on a mixed-model ensemble, but broader cross-family, cross-dataset, and domain-specific evaluations remain open. Whether current models have diverse enough error profiles to benefit from heterogeneous ensemble averaging or whether shared training data causes cross-family correlation that resists spectral analysis [Sharma et al., 2024] remains an open question with implications for the framework's practical ceiling.

Practitioner Guidelines. When the target application tolerates an average error rate of approximately 5% (e.g., summarization tasks), naive majority voting is likely sufficient and conformal calibration adds unnecessary complexity. However, for high-assurance applications with stringent error budgets ($\alpha \leq 0.02$), naive approaches fail. The Score Deviation Penalty provides a statistical basis for enforcing safety policies without compromising recall. High-stakes applications are motivating use cases rather than validated deployment settings for this experiment. In such settings, domain-specific benchmarks, human-audited strong-judge labels, and recalibration under distribution shift are required before operational use.

Beyond Factual Verification. The spectral analysis mechanism may extend beyond factual verification to settings where diversity is an asset rather than a liability. In scientific hypothesis generation, for instance, calibrated entropy signals could identify collectively novel solution regions, linking conformal risk control to structured exploration in multi-agent systems [Kostka and Chudziak, 2024]. We leave the development of diversity-seeking calibration objectives for future work.

7 CONCLUSION

This work bridges the gap between heuristic consensus and formal safety guarantees for multi-agent fact verification by reframing hallucination control as a tuple-level risk control problem. The key insight is that high agreement among agents can signal sycophantic mimicry rather than independent corroboration, and that penalizing high within-tuple score deviation improves the calibration efficiency of the resulting safety certificate. The experimental evaluation confirms that this approach rigorously bounds the false discovery rate at strict risk budgets where naive baselines fail, while maintaining practical recall.

To achieve this, the framework establishes the query-response tuple as the correct exchangeable unit, introduces the Score Deviation penalty that compresses the empirical risk curve, and provides a scoring function ablation revealing the U-shaped entropy-error relationship underlying penalty design.

While the current evaluation validates the framework primarily on a homogeneous ensemble, the heterogeneous and higher-base-rate experiments provide initial evidence that the calibrated safety layer is not restricted to the homogeneous, low-base-rate setting. Three directions follow from this foundation. Larger heterogeneous ensembles and cross-dataset calibration studies would test whether the deviation penalty captures broader cross-family error correlations. Extending the guarantee to non-stationary deployment streams through online recalibration would broaden applicability. Finally, adapting the spectral analysis to settings where diversity is an asset, such as scientific hypothesis generation, could yield calibrated novelty signals for structured exploration.

References

Yasin Abbasi-Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Taylan Cemgil, and Nenad Tomašev. Mitigating llm hallucinations via conformal abstention. In *NeurIPS Workshop on Statistical Frontiers in LLMs and Foundation Models*, 2024. URL https://arxiv.org/abs/2405.01563.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101. URL https://doi.org/10.1561/2200000101.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022. doi: 10.48550/arXiv.2208.02814. URL https://arxiv.org/abs/2208.02814.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2), 2025. doi: 10.1214/24-AOAS1998. URL https://arxiv.org/abs/2110.01052.

Feliks Bańka and Jarosław A. Chudziak. DeltaHedge: A multi-agent framework for portfolio options optimization. In *Proceedings of the 29th Pacific Asia Conference on Information Systems*, 2025. URL https://aisel.aisnet.org/pacis2025/aiandml/aiandml/25.

Stephen Bates, Anastasios N. Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535. URL https://arxiv.org/abs/2101.02703.

Vidmantas Bentkus. On hoeffding's inequalities. *The Annals of Probability*, 32(2):1650–1673, 2004. doi: 10.1214/009117904000000360. URL https://projecteuclid.org/journals/annals-of-probability/volume-32/issue-2/On-Hoeffdings-inequalities/10.1214/009117904000000360.full.

Gilles Blanchard, Guillermo Durand, Ariane Marandon-Carlhian, and Romain Périer. FDR control and FDP bounds for conformal link prediction. *arXiv preprint arXiv:2404.02542*, 2024. URL https://arxiv.org/abs/2404.02542.

Samuel L. Braunstein, Sibasish Ghosh, and Simone Severini. The laplacian of a graph as a density matrix: A basic combinatorial approach to separability of mixed states. *Annals of Combinatorics*, 10(3):291–317, 2006. doi: 10.1007/s00026-006-0289-3. URL https://arxiv.org/abs/quant-ph/0406165.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763. PMLR, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://www.nature.com/articles/s41586-024-07421-0.

Mahmood Hegazy. Diversity of thought elicits stronger reasoning capabilities in multi-agent debate frameworks. *Journal of Robotics and Automation Research*, 5(3):1–10, 2024. URL https://arxiv.org/abs/2410.12853.

Zirui Hu, Zheng Zhang, Yingjie Wang, Leszek Rutkowski, and Dacheng Tao. CoFact: Conformal factuality guarantees for language models under covariate shift. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=eiBp7rsc3K.

Huipeng Huang, Wenbo Liao, Huajun Xi, Hao Zeng, Mengchen Zhao, and Hongxin Wei. Model-agnostic selective labeling with provable statistical guarantees. *arXiv preprint arXiv:2510.14581*, 2025. URL https://arxiv.org/abs/2510.14581.

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, et al. The facts grounding leaderboard: Benchmarking llms' ability to ground responses to long-form input. *arXiv preprint arXiv:2501.03200*, 2025. URL https://arxiv.org/abs/2501.03200.

Taejong Joo, Shu Ishida, Ivan Sosnovik, Bryan Lim, Sahand Rezaei-Shoshtari, Adam Gaier, and Robert Giaquinto. Graph of agents: Principled long context modeling by emergent multi-agent collaboration. *arXiv preprint arXiv:2509.21848*, 2025. URL https://arxiv.org/abs/2509.21848.

Mintong Kang, Nezihe Merve Gürel, Ning Yu, Dawn Song, and Bo Li. C-RAG: Certified generation risks for retrieval-augmented language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 22963–23000. PMLR, 2024. URL https://proceedings.mlr.press/v235/kang24a.html.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*, 2024. URL https://arxiv.org/abs/2406.15927.

Adam Kostka and Jarosław A. Chudziak. Synergizing logical reasoning, knowledge management and collaboration in multi-agent LLM system. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 203–212. Tokyo University of Foreign Studies, 2024. URL https://aclanthology.org/2024.paclic-1.19/.

Adam Kostka and Jarosław A. Chudziak. Evaluating theory of mind and internal beliefs in LLM-based multi-agent systems. In *Computational Collective Intelligence (ICCCI 2025)*, Lecture Notes in Computer Science, pages 18–32. Springer, 2025. doi: 10.1007/978-3-032-09318-9_2. URL https://link.springer.com/chapter/10.1007/978-3-032-09318-9_2.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2302.09664.

Bracha Laufer-Goldshtein, Adam Fisch, Regina Barzilay, and Tommi S. Jaakkola. Efficiently controlling multiple risks with pareto testing. In *International Conference*

on Learning Representations, 2023. URL https://openreview.net/forum?id=cyg2YXn_BqF.

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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.397. URL https://aclanthology.org/2023.emnlp-main.397/.

Yi Liu. A statistically consistent measure of semantic uncertainty using language models. *arXiv preprint arXiv:2502.00507*, 2025. URL https://arxiv.org/abs/2502.00507.

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*, pages 12076–12100. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Dang Nguyen, Ali Payani, and Baharan Mirzasoleiman. Beyond semantic entropy: Boosting LLM uncertainty quantification with pairwise semantic similarity. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4530–4540. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.234. URL https://aclanthology.org/2025.findings-acl.234/.

Kangjun Noh, Seongchan Lee, Ilmun Kim, and Kyungwoo Song. Multi-LLM adaptive conformal inference for reliable LLM response. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=opuQH9Xyu9.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.847. URL https://aclanthology.org/2023.findings-acl.847/.

Priya Pitre, Naren Ramakrishnan, and Xuan Wang. CONSENSAGENT: Towards efficient and effective consensus in multi-agent LLM interactions through sycophancy mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22112–22133. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1141. URL https://aclanthology.org/2025.findings-acl.1141/.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.13548.

Andries Petrus Smit, Paul Duckworth, Nathan Grinsztajn, Kale-ab Tessera, Thomas D. Barrett, and Arnu Pretorius. Are we going MAD? benchmarking multi-agent debate between language models for medical q&a. In *NeurIPS 2023 Workshop on Deep Generative Models for Health (DGM4H)*, 2023. URL https://openreview.net/forum?id=Bfr0m4Ucl6.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.552. URL https://aclanthology.org/2024.findings-emnlp.552/.

Kamil Szczepanik and Jarosław A. Chudziak. Collaborative LLM agents for C4 software architecture design automation. *arXiv preprint arXiv:2510.22787*, 2025. doi: 10.48550/arXiv.2510.22787. URL https://arxiv.org/abs/2510.22787.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1904.06019.

Nassim Walha, Sebastian G. Gruber, Thomas Decker, Yinchong Yang, Alireza Javanmardi, Eyke Hüllermeier, and Florian Buettner. Fine-grained uncertainty decomposition in large language models: A spectral approach. *arXiv preprint arXiv:2509.22272*, 2025. URL https://arxiv.org/abs/2509.22272.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *arXiv preprint arXiv:2509.05396*, 2025. URL https://arxiv.org/abs/2509.05396. ICML MAS Workshop.

Binwei Yao, Chao Shang, Wanyu Du, Jianfeng He, Ruixue Lian, Yi Zhang, Hang Su, Sandesh Swamy, and Yanjun Qi. Peacemaker or troublemaker: How sycophancy shapes multi-agent debate. *arXiv preprint arXiv:2509.23055*, 2025. URL https://arxiv.org/abs/2509.23055.

Zhiheng Zhang, Yuanzhe Zhang, Daojian Zeng, Kang Liu, and Jun Zhao. Beware of the woozle effect: Exploring and mitigating hallucination propagation in multi-agent debate. *IEEE Transactions on Audio, Speech and Language Processing*, 34:2021–2032, 2026. doi: 10.1109/TASLPRO.2026.3675803. URL https://doi.org/10.1109/TASLPRO.2026.3675803.

Supplementary Material

A THEORETICAL ANALYSIS

This section provides the formal validity guarantee, the union bound justification, and an efficiency argument for why deviation penalties improve coverage.

A.1 VALIDITY GUARANTEE

Theorem 1 (Risk control)**.** *Let $T_1, \ldots, T_n$ be exchangeable calibration tuples, each comprising a query, its extracted atomic facts, and strong-judge labels. Let λ^* be the threshold returned by the LTT calibration procedure with target risk $\alpha \in (0, 1)$ and confidence parameter $\delta \in (0, 1)$. Then*

$$\Pr\big(R(\lambda^*) \leq \alpha\big) \;\geq\; 1 - \delta,$$

where $R(\lambda) = \mathbb{E}[\mathrm{FDP}(\lambda)]$ is the population level false discovery rate defined in Section 3.

Proof sketch. The proof follows the Learn-Then-Test template. For each candidate threshold λ_j, the null hypothesis $H_0^{(j)}\colon R(\lambda_j) > \alpha$ is tested using the combined Hoeffding-Bentkus p-value (9). Under $H_0^{(j)}$, the empirical risk $\widehat{R}(\lambda_j)$ is a mean of n exchangeable $[0, 1]$ bounded random variables, so both the Hoeffding bound (7) and the Bentkus bound (8) yield valid p-values [Bentkus, 2004]. The minimum of two valid p-values, after a union bound correction factor of 2, is itself a valid p-value. The fixed sequence testing procedure tests from λ_m downward and stops at the first non-rejection, controlling the family-wise error rate at level δ without Bonferroni correction because the rejection sets are nested ($\widehat{R}$ is non-increasing in λ). The selected λ^* therefore satisfies $\Pr(R(\lambda^*) > \alpha) \leq \delta$. □

Assumptions. The guarantee requires: (i) exchangeability of calibration tuples, which holds when queries are drawn i.i.d. from the user population, (ii) the scoring function S and ensemble are fixed before calibration (no adaptive selection), (iii) the strong-judge labels are treated as ground truth, and crucially, (iv) the candidate threshold grid $\{\lambda_1, \ldots, \lambda_m\}$ is fixed prior to observing the calibration data. Assumption (iii) is standard in conformal prediction. If the strong judge is imperfect, the guarantee controls risk relative to the strong judge's labels, not absolute factuality.

Failure modes and boundary conditions. For deployment interpretation, four boundary conditions are critical. First, **Shifted query distribution:** if deployment queries are not exchangeable with calibration tuples, finite sample FDR control may fail. Second, **Label quality mismatch:** the certificate is with respect to strong judge labels, so systematic bias transfers to the controlled target. Third, **Adaptive score changes:** modifying the weak judge or encoder post calibration invalidates the certificate. Fourth, **Low risk degeneracy:** when the baseline unfiltered FDR exceeds α, the optimal certified behavior is near total abstention. These conditions delimit the regime in which the guarantee is interpretable.

A.2 UNION BOUND FOR COMBINED P-VALUES

In Section 4.4, we compute the final p-value for each candidate threshold by taking the minimum of the Hoeffding and Bentkus p-values and multiplying by 2:

$$p(\lambda) = \min\big(1, 2 \cdot \min\big(p_{\mathrm{H}}(\lambda),\; p_{\mathrm{B}}(\lambda)\big)\big).$$

This factor of 2 is a standard union bound correction required when combining multiple valid p-values. While both p_{H} and p_{B} are individually valid upper bounds on the probability of observing the empirical risk under the null hypothesis, they are derived from different concentration inequalities and are not perfectly correlated.

If we were to simply take the minimum without correction, the resulting statistic would no longer be uniformly distributed under the null hypothesis, leading to an inflated Type I error rate. Specifically, for any two valid p-values p_1 and p_2, the event $\min(p_1, p_2) \leq \delta$ can occur if either $p_1 \leq \delta$ or $p_2 \leq \delta$. By Boole's inequality (the union bound):

$$\Pr\big(\min(p_1, p_2) \leq \delta\big) \;\leq\; \Pr(p_1 \leq \delta) + \Pr(p_2 \leq \delta) \;\leq\; 2\delta.$$

To ensure the combined test controls the error rate at the target level δ, we must test each individual p-value at level $\delta/2$, which is algebraically equivalent to multiplying the minimum by 2 and testing against δ. This correction guarantees that the finite-sample validity of the Learn-Then-Test procedure remains mathematically sound, preventing anti-conservative threshold selection.

A.3 VARIANCE REDUCTION EFFICIENCY

The practical advantage of deviation penalties over unpenalized scores is that they reduce the variance of the calibration estimator $\widehat{R}(\lambda)$, yielding tighter p-values and a less conservative threshold.

Let S_{dev} and S_{naive} denote scoring functions with and without the deviation penalty. Because S_{dev} suppresses scores where the ensemble exhibits structural collapse or severe disagreement, the resulting empirical risk curve $\widehat{R}(\lambda)$ exhibits lower variance for a fixed calibration set size n. The Hoeffding p-value (7) depends on the gap $\alpha - \widehat{R}(\lambda)$; a scoring function with lower variance produces a more concentrated $\widehat{R}$, increasing the expected gap and decreasing the p-value. Smaller p-values allow LTT to reject H_0 at lower thresholds, accepting more facts. This is confirmed empirically: the γ=0.5 variant achieves recall improvements exceeding 24 percentage points at strict risk budgets over γ=0 (the naive baseline), while eliminating threshold degeneracy.

B EXPERIMENTAL DETAILS

This section provides the information necessary to reproduce the experimental results reported in Section 5.

B.1 HYPERPARAMETERS

Table 4 lists all configurable parameters. No hyperparameter was tuned on the test split.

Table 4: Complete hyperparameter specification. Each certified run fixes the scoring rule and γ before LTT calibration; the γ sweep is a separate robustness analysis.

Parameter	Module	Value
Ensemble size K	Generator	5
Generator model	Generator	Gemini 2.5 Flash
Temperature	Generator	0.7
Max tokens (generation)	Generator	2048
Decomposer model	Decomposer	Gemini 2.5 Flash
Max facts per response	Decomposer	20
Weak judge model	Scorer	Gemini 2.5 Flash
Strong judge model	Calibration	Gemini 3.0 Flash Exp.
Embedding model	Entropy	all-MiniLM-L6-v2
Penalty weight γ	Scorer	0.5 (primary)
Target FDR α	Risk control	0.05 (default)
Confidence δ	Risk control	0.1
Threshold grid size m	Risk control	100
Random seed	Global	42

B.2 SYSTEM PROMPTS

The following prompts are used verbatim in the pipeline. Placeholders `{context}` and `{claim}` (or `{text}`) are filled at runtime.

Atomic Decomposer.

```
Extract atomic facts from the text. Each fact must be: 1. Self-contained
(understandable without context) 2. Verifiable (can be checked as
true/false) 3. Atomic (contains exactly one claim)

Text: {text}

Output format (JSON array): ["fact 1", "fact 2", ...]
```

Weak Judge (NLI Scorer).

```
You are a factual entailment judge. Estimate the probability (0-100) that
the CLAIM is supported by the CONTEXT.

CONTEXT: {context}

CLAIM: {claim}

Guidelines: - 0 = directly contradicted or completely absent - 100 =
explicitly and fully stated in context - Use ANY integer from 0 to 100

Examples: - Claim restates a context sentence verbatim → 95 - Claim is a
reasonable paraphrase of context → 78 - Claim is partially supported, key
detail missing → 42 - Claim is vaguely related but not confirmed → 18 -
Claim contradicts the context → 3

Respond with ONLY a single integer (e.g. 72).
```

Strong Judge (Calibration Labels).

```
You are a strict factual verification judge.

CONTEXT: {context}

CLAIM: {claim}

Rules: - Use ONLY information explicitly stated in the context. - Do NOT
use external knowledge or reasoning beyond what the context states. - If
the context lacks sufficient information to confirm the claim, answer
NOT_SUPPORTED. - Paraphrases and logical entailments count as supported.

Respond with EXACTLY one word: SUPPORTED or NOT_SUPPORTED
```

B.3 ENSEMBLE AGENT PERSONAS

Diversity is injected through five distinct system prompts, each constraining the agent to use only the provided context:

1. *Skeptical scientist*: "You are a skeptical scientist. Answer the question using ONLY information from the provided context. If the context lacks evidence, say so explicitly."
2. *Thorough researcher*: "You are a thorough researcher. Answer the question using ONLY the provided context document. Explore multiple angles present in the text."
3. *Precision assistant*: "You are a precise assistant focused on accuracy. Answer using ONLY the provided context. Quote or paraphrase specific passages to support your answer."
4. *Critical analyst*: "You are a critical analyst. Answer the question using ONLY the provided context. Question any assumptions not directly supported by the text."
5. *Balanced reporter*: "You are a balanced reporter. Answer the question using ONLY the provided context document. Present facts objectively without adding external knowledge."

All five agents share the same base model (Gemini 2.5 Flash) and temperature (T=0.7). The persona variation is the sole source of response diversity.

B.4 SPECTRAL ENTROPY COMPUTATION

Given K response embeddings $\{e_1, \ldots, e_K\}$ from the sentence encoder (all-MiniLM-L6-v2), the spectral entropy is computed as follows:

1. Compute the pairwise cosine similarity matrix $W \in \mathbb{R}^{K \times K}$.
2. Symmetrize: $W \leftarrow (W + W^\top)/2$; clip entries to $[0, 1]$.
3. Compute eigenvalues $\lambda_1, \ldots, \lambda_K = \mathrm{eigvalsh}(W)$.
4. Ensure non-negativity: $\lambda_i \leftarrow \max(\lambda_i, \varepsilon)$ for $\varepsilon = 10^{-10}$.
5. Normalize: $p_i = \lambda_i / \sum_j \lambda_j$.
6. Return $H = -\sum_i p_i \log p_i$.

B.5 CALIBRATION PROTOCOL

The 300 tuples are split into 200 calibration and 100 test tuples using a fixed random seed. Strong-judge labels are collected in a single offline pass before the split is determined. The scoring function and all hyperparameters are fixed prior to calibration. No information from the test split influences threshold selection. For the 20-fold cross-validation reported in Table 1, each fold independently reshuffles the 300 tuples into a 200/100 split and runs the full LTT procedure from scratch.

C DATASET STATISTICS

Table 5 summarizes the calibration dataset derived from the FACTS Grounding benchmark.

Table 5: Dataset summary statistics.

Statistic	Value
Total tuples	300
Total atomic facts	22,155
Facts per tuple (mean / min / max)	73.8 / 5 / 100
Supported facts	21,074 (95.1%)
Unsupported facts	1,081 (4.9%)
Raw weak score u (mean $\pm$ std)	0.942 ± 0.142
Raw weak score u (median)	0.95
Adjusted score S at γ=0.5 (mean $\pm$ std)	0.903 ± 0.141
Ensemble entropy H (mean $\pm$ std)	0.370 ± 0.151
Ensemble entropy H (min / max)	0.010 / 0.937

The high support rate (95.1%) means the baseline unfiltered FDR is approximately 4.9%. At target risk levels $\alpha \geq 0.05$, the LTT procedure can find permissive thresholds with high recall. At stricter budgets ($\alpha \leq 0.03$), the gap between the baseline error rate and the target narrows, making the choice of scoring function critical for maintaining throughput.

D EXTENDED ROBUSTNESS ANALYSIS

D.1 CROSS-VALIDATION DETAILS

Table 6 extends the 20-fold cross-validation results summarized in Table 1. Each fold independently reshuffles the 300 tuples into a 200/100 calibration-test split and runs the full LTT calibration. The Score Deviation penalty wins in all 20 folds at α=0.02 and α=0.04, and in 16 of 20 folds at α=0.03, with zero degenerate thresholds across all runs. The empirical test-set FDR remains well below the target in every fold (mean test FDR of 0.005 at α=0.02, maximum 0.009), confirming that the Hoeffding-Bentkus bounds are conservative but valid.

Table 6: 20-fold cross-validation: Score Deviation (γ=0.5) vs. Naive baseline. Win rate denotes the fraction of folds where Score Deviation achieves strictly higher recall. At α=0.02, the naive baseline suffers threshold degeneracy (λ^*=1.0) in 14 of 20 folds.

α	Strategy	Recall (mean $\pm$ std)	λ^* (mean)	Degen.	Win Rate
0.02	Naive	0.474 ± 0.018	0.995	14/20	—
0.02	Score Dev.	$\mathbf{0.717 \pm 0.097}$	0.907	0/20	20/20
0.03	Naive	0.734 ± 0.210	0.928	0/20	—
0.03	Score Dev.	$\mathbf{0.912 \pm 0.019}$	0.863	0/20	16/20
0.04	Naive	0.905 ± 0.007	0.905	0/20	—
0.04	Score Dev.	$\mathbf{0.961 \pm 0.008}$	0.833	0/20	20/20

D.2 CONVERGENCE TRAJECTORY

Table 7 tracks the calibrated threshold λ^* as the number of calibration tuples increases from 20 to 300 at α=0.05, γ=1.0. Three phases are visible: (1) a conservative phase ($n < 80$) where the concentration bounds are too weak to reject any null hypothesis, resulting in total abstention (λ^*=1.0); (2) a transition at $n \approx 80$ where the first non-trivial threshold emerges; and (3) a refinement phase ($n > 100$) where the threshold decreases monotonically as concentration improves.

Table 7: Threshold convergence as calibration set size increases.

n (tuples)	λ^*	Empirical FDR
20–60	1.000	0.000
80	0.556	0.011
100	0.556	0.009
140	0.535	0.018
200	0.505	0.023
260	0.485	0.026
300	0.394	0.030

D.3 SCORING FORMULA CATALOG

Table 8 lists all 14 scoring formulas evaluated during the ablation study. Each formula was swept over $\gamma \in \{0.25, 0.5, 0.75, 1.0, 1.5, 2.0\}$ and $\alpha \in \{0.02, 0.03, 0.04, 0.05\}$ using the same 200/100 calibration-test split. The Score Deviation penalty (formula 12) is the only formula that consistently beats the Naive baseline at all strict α levels with 95% bootstrap confidence intervals that do not overlap. The ablation separates four hypotheses: unpenalized confidence, low-entropy-only sycophancy penalties, penalties on both entropy extremes, and within-tuple weak-judge score spread.

The entropy-based formulas (2–11, 13–14) target the spectral structure of the ensemble's embedding space. While several of these improve over the Naive baseline at moderate α (e.g., α=0.05), none consistently dominates at strict budgets ($\alpha \leq 0.03$). The Score Deviation penalty operates on a fundamentally different signal: the spread of the weak judge's numerical confidence scores within a tuple, rather than the geometric diversity of the ensemble's responses. This direct connection to the scoring function used by LTT explains its superior calibration efficiency.

D.4 EXTENDED GAMMA SENSITIVITY

Table 9 extends the gamma sensitivity analysis from Table 2 using the full 300-tuple dataset (no train-test split) to show the asymptotic behavior. The FDR guarantee holds across the entire range $\gamma \in [0, 5]$, and recall is essentially constant for $\gamma \geq 0.25$. The threshold λ^* decreases monotonically to compensate for the score compression introduced by larger γ, confirming that the LTT procedure absorbs the penalty strength automatically.

Table 8: All scoring formulas tested. H denotes spectral entropy, $\tilde{H}$ its empirical median, σ_H its standard deviation, and σ_{tuple} the within-tuple score standard deviation.

#	Name	Formula	Intuition
1	Naive	u	No penalty
2	Original	$u/(1+\gamma e^{-H})$	Penalize low H
3	Original β=5	$u/(1+\gamma e^{-5H})$	Sharper low-H
4	U-shape	$u/(1+\gamma\lvert H-\tilde{H}\rvert)$	Both extremes
5	U-shape (norm.)	$u/(1+\gamma\lvert H-\tilde{H}\rvert/\sigma_H)$	Z-scored
6	Gaussian bell	$u/(1+\gamma e^{-(H-\tilde{H})^2})$	Penalize middle
7	Gaussian bell k=5	$u/(1+\gamma e^{-5(H-\tilde{H})^2})$	Sharper bell
8	Entropy boost	$u(1+\gamma H_{\text{norm}})$	Reward diversity
9	Sigmoid (s=3)	$u/(1+\gamma(1-\sigma_s))$	Smooth low-H
10	Sigmoid (s=10)	$u/(1+\gamma(1-\sigma_s))$	Sharp sigmoid
11	Log U-shape	$u/(1+\gamma\log(1+\lvert H-\tilde{H}\rvert))$	Soft U-shape
12	**Score Deviation**	$\mathbf{u}/(\mathbf{1}+\gamma\sigma_{\textbf{tuple}})$	**Score spread**
13	Entropy boost (clip)	$u(1+\gamma H_{\text{clip}})$	Boost high H
14	Composite	$u(1+\gamma H)/(1+\gamma)$	Diversity blend

Table 9: Gamma sweep on the full dataset at α=0.05, δ=0.1.

γ	λ^*	FDR	Recall
0.00	0.707	0.029	0.992
0.25	0.586	0.030	0.993
0.50	0.505	0.030	0.993
1.00	0.394	0.030	0.993
2.00	0.283	0.030	0.993
5.00	0.152	0.030	0.993

E NOTATION REFERENCE

Table 10: Summary of notation used throughout the paper.

Symbol	Definition
$T_i = (x_i, \mathcal{Z}_i, c_i)$	Query-response tuple (exchangeable unit)
K	Number of ensemble agents
u_m	Raw weak-judge entailment score for fact z_m
H_{spec}	Spectral entropy of ensemble similarity matrix
σ_{tuple}	Within-tuple score standard deviation
$S(z_m)$	Adjusted (penalized) score
γ	Deviation penalty weight
λ, λ^*	Filtering threshold; calibrated threshold
α	Target false discovery rate
δ	Confidence parameter ($1-\delta$ guarantee)
$R(\lambda)$	Population FDR: $\mathbb{E}[\text{FDP}(\lambda)]$
$\widehat{R}(\lambda)$	Empirical FDR on calibration set
p_H, p_B	Hoeffding and Bentkus p-values
$\text{FDP}(\lambda)$	Per-tuple false discovery proportion