# Claim-Level Grounding Verification with Conformal Risk Control for Retrieval-Augmented Generation in Regulated Domains

*A simulation study of hallucination detection, abstention guarantees and audit-ready evidence alignment for legal, medical and compliance workflows*

[1]Sonia Mohapatra, [2]Saswat Swain, [3]Chandrasekhar Panda
[1]Researcher, [2]Ph.D Scholar, [3]Researcher
[1,3]Research and Development Division,
Tekstein Solutions Private Limited, Bhubaneswar, Odisha, India
[2]KSOM — KIIT Deemed to be University,
Bhubaneswar, Odisha, India
[1]smctc98@gmail.com, [2]saswat.mck@gmail.com, [3]contact@chandrasekharpanda.com

***Abstract*—Retrieval-augmented generation (RAG) is now used in legal, clinical and compliance work, where an unsupported sentence is a liability, yet existing hallucination detectors score a response as a whole, give no control over the rate at which ungrounded statements slip through, and leave nothing a reviewer can audit. This paper proposes CGV-RC, a claim-level grounding verification overlay with conformal risk control. A response is decomposed into atomic claims; each claim triggers a claim-conditioned re-retrieval and is scored by claim-level entailment, a numeric and entity consistency check and a provenance check. A logistic model fitted on a calibration split fuses the three, and a split-conformal threshold bounds the claim-level false-negative rate by a chosen tolerance. Three propositions, on the fusion-gain ceiling, the finite-sample conformal miss rate and response-level risk, are derived and validated numerically. Because no public corpus labels the four hallucination types that matter in regulated work, the study uses a synthetic generative model with type-specific detector blind spots calibrated from published ranges, so results characterise the method under stated assumptions rather than measure a deployed system. Across three simulated domains CGV-RC reaches a mean AUROC of 0.887 against 0.681 for passage-level NLI and 0.663 for five-sample self-consistency at about 7 percent of the latter's cost. The gate holds the target miss rate within finite-sample error, but a 1 percent tolerance leaves only 22 percent of grounded claims unreviewed, and extrinsic claims remain the hardest class.**

***Index Terms*—Retrieval-augmented generation, hallucination detection, grounding verification, conformal prediction, selective prediction, natural language inference, regulated domains, auditability.**

## I. Introduction

Retrieval-augmented generation couples a large language model (LLM) with a document retriever so that answers are conditioned on an external corpus rather than on parametric memory alone [1]. The corpus can be curated, versioned and audited, but nothing stops the generator from writing text the evidence does not support: audits of commercial legal research assistants built on RAG report that between one sixth and one third of responses contain at least one unsupported or mis-cited statement [11], [12]. A filed fabricated citation, an invented contraindication or a mis-quoted regulatory clause carries consequences no aggregate accuracy figure captures.

Hallucination detection has received sustained attention [2]. The dominant families are generator-internal signals such as token probability and entropy [10], consistency across repeated samples [4], and natural language inference (NLI) between response and evidence [5], [6], [15]. Each fails where it matters here: token confidence is loosely tied to factuality after instruction tuning; self-consistency misses claims the model repeats confidently but wrongly, at several times the cost; passage-level NLI is unreliable on numbers, dates and entities and cannot see that a correct statement carries the wrong citation. None bounds the rate at which ungrounded claims pass, and none leaves an artefact a reviewer can inspect.

This paper treats grounding verification as selective prediction under a miss-rate constraint. Its contributions are: (1) CGV-RC, an overlay that decomposes a response into atomic claims, re-retrieves evidence per claim, fuses three complementary channels with a calibrated logistic model, gates the result with a split-conformal threshold and writes an audit record per claim (Section IV); (2) three propositions on the fusion ceiling, the conformal miss rate and response-level risk (Section V); (3) a generative simulation model with type-specific detector blind spots that makes every assumption explicit (Section VI); and (4) an evaluation that reports the price of the guarantee, the ablations that failed and the class that remains unsolved (Sections VII and VIII).

## II. Related Work

Table I sets the detector families against the requirements of regulated deployment. FActScore [3] introduced atomic-claim decomposition but targets parametric knowledge rather than

fidelity to supplied evidence; SelfCheckGPT [4] exploits disagreement between stochastic samples at $k$ times the generation cost; SummaC [6], TRUE [5] and MiniCheck [15] established sentence-level NLI as a practical faithfulness signal. RAGTruth [13] annotates hallucination spans in RAG output and separates evident conflict from baseless information, a split adopted here and extended with citation and numeric classes; RAGAS [7] and semantic entropy [10] report averages or model uncertainty rather than agreement with evidence. Conformal prediction [8], [9] has been applied to filtering sub-claims so that retained claims meet a factuality target [14]. The present work places the conformal layer on an evidence-grounded multi-channel score, controls the false-negative rate of the flag rather than the precision of retained claims, and is built around the record-keeping and human-oversight duties the EU AI Act imposes on high-risk systems [16].

TABLE I. DETECTOR FAMILIES AGAINST REGULATED-DEPLOYMENT REQUIREMENTS

| Family (representative) | Signal | Gap for regulated use | AUROC [a] | Cost [b] |
|---|---|---|---|---|
| Token confidence / entropy [10] | Generator-internal | Weak link to factuality; no evidence trail | 0.599 | 0 |
| Self-consistency [4] | Sample agreement | $k$× cost; blind to confident errors | 0.663 | 5.0 |
| Passage-level NLI [5], [6], [15] | Entailment | Fails on numbers, entities, mis-attribution | 0.681 | 0.15 |
| Atomic-claim scoring [3] | Per-claim judge | No rate guarantee; no provenance check | — | — |
| Conformal claim filtering [14] | Self-confidence + conformal | Not evidence-grounded; precision, not FNR | — | — |
| **CGV-RC (this work)** | 3 evidence channels + conformal | FNR bound and audit record; simulation-only validation (Sec. IX) | **0.887** | 0.35 |

[a]Mean over the three simulated domains of Section VII; [b]assumed generator-equivalent passes per claim (Table II).

## III. Problem Formulation

Let $q$ be a query, $E = \{e_1, \ldots, e_K\}$ the retrieved passages and $R$ the response, decomposed into atomic claims $\{c_1, \ldots, c_n\}$, each a declarative statement assessable against one passage. Each claim has a hidden label $y_i \in \{0, 1\}$, with $y_i = 1$ if no passage in the corpus supports it. Ungrounded claims are typed as *intrinsic* (contradicted by the evidence), *extrinsic* (plausible but unsupported), *citation* (supported content attributed to a passage that does not contain it) or *numeric* (a value, date or entity altered from the source). The verifier outputs a flag $f_i \in \{0, 1\}$ and must satisfy, for a tolerance $\alpha$,

$$P(f_i = 0 \mid y_i = 1) \le \alpha, \tag{1}$$

while flagging as few grounded claims as possible. Equation (1) bounds the false-negative rate, the event that causes harm. The verifier must also cost a fraction of the generator, which excludes repeated generation, and every decision must be reconstructible from stored quantities. The grounding score of a claim against an evidence set is

$$g(c_i, E_i) = \max_j s(c_i, e_j), \quad e_j \in E_i, \tag{2}$$

where $s$ is an entailment probability and $E_i \supseteq E$ is the claim-specific evidence set of Section IV-A. One supporting passage suffices, so a retrieval miss propagates directly into a low score, the mechanism examined in Section VII-C.

## IV. Proposed Methodology

Fig. 1 shows the architecture. The generation stage is a conventional RAG pipeline left untouched; CGV-RC receives $(R, E)$ and returns per-claim decisions with audit records.

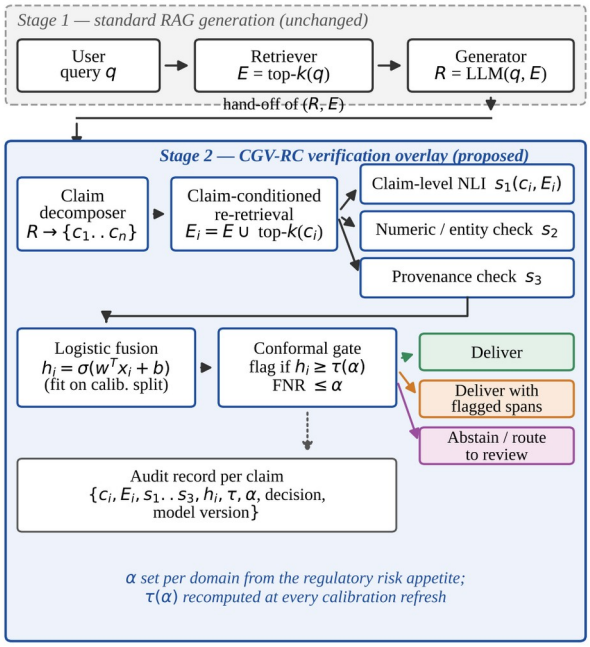



Fig. 1. CGV-RC architecture. Stage 1 is an unmodified RAG pipeline. Stage 2 decomposes the response into atomic claims, re-retrieves evidence per claim, scores three channels, fuses them with a logistic model fitted on a calibration split, and applies a split-conformal threshold $\tau(\alpha)$ that bounds the claim-level false-negative rate; every claim produces an audit record.

### A. Claim Decomposer and Re-Retrieval

The response is split into atomic claims by a constrained prompt that forbids paraphrase and requires each claim to carry its own subject; the simulation assumes this step is correct (Section IX). Because $E$ was retrieved for the query rather than for each claim, a claim derived from background knowledge, or resting on a passage ranked just below the top-$k$ cut-off, has no support in $E$ even when the corpus contains it. A second retrieval with the claim as query forms $E_i = E \cup$ top-$k(c_i)$; in the simulation a missing passage is recovered with probability $\rho = 0.6$.

### B. Verification Channels

Three signals are computed per claim. The *claim-level NLI* channel $s_1$ evaluates (2) on a short claim, removing the dilution that lets passage-level NLI miss one wrong clause. The *numeric and entity* channel $s_2$ compares numbers, dates, units, drug names, section identifiers and party names between the claim and the best-matching passage, because entailment models treat "10 mg" and "100 mg" as near-synonyms. The *provenance* channel $s_3$ checks that any citation attached to the claim resolves

to a passage in $E_i$ that entails it. The three were chosen because their noise is largely independent, the condition under which Proposition 1 says an extra channel earns its cost.

### C. Fusion and Conformal Gate

The feature vector $x_i$ = [logit $s_1$, $s_2$, $s_3$, 1[$c_i$ recovered by re-retrieval]] is mapped to a hallucination probability

$$h_i = \sigma(w^T x_i + b), \tag{3}$$

with ($w$, $b$) fitted by L2-regularised logistic regression on a labelled calibration split never used for evaluation; four inspectable coefficients are preferred to learned deep fusion for audit reasons. The gate applies split-conformal risk control [8], [9]. Let $\{h_j : y_j = 1, j = 1 \dots m\}$ be the fused scores of the ungrounded claims in a second, disjoint calibration set, sorted as $h_{(1)} \le \dots \le h_{(m)}$. Then

$$\tau(\alpha) = h_{(\lfloor (m+1)\alpha \rfloor)}, \;\; f_i = 1[h_i \ge \tau(\alpha)]. \tag{4}$$

A response whose claims all pass is delivered, one with flagged claims is delivered with the spans marked, and one whose flagged fraction exceeds a policy $\theta$ is withheld for human review. Each claim writes the record {$c_i$, $E_i$, $s_{1..3}$, $h_i$, $\tau$, $\alpha$, decision, model version}.

### D. Algorithm and Complexity

Algorithm 1 gives the verification step. For $n$ claims the cost is $n$ re-retrievals, $O(n \log N)$ over $N$ indexed passages, $n \cdot |E_i|$ short-input NLI calls and linear-time extraction; the threshold is one order statistic, $O(m \log m)$ at calibration and $O(1)$ per claim. In generator-equivalent forward passes per claim (Table II) CGV-RC costs 0.35 against 5.0 for five-sample self-consistency.

ALGORITHM 1. CGV-RC VERIFICATION OF ONE RESPONSE

**Input:** response $R$, evidence $E$, weights ($w$, $b$), threshold $\tau(\alpha)$, review policy $\theta$
**Output:** decision ∈ {deliver, deliver-with-flags, abstain}; audit records
1: $\{c_1 \dots c_n\} \leftarrow$ Decompose($R$)
2: **for** $i = 1 \dots n$ **do**
3: $E_i \leftarrow E \cup$ Retrieve($c_i$, $k$); $\rho_i \leftarrow 1[E_i \setminus E$ non-empty]
4: $s_1 \leftarrow \max_{e \in E_i}$ NLI($c_i$, $e$); $s_2 \leftarrow$ NumEntityCheck($c_i$, $E_i$)
5: $s_3 \leftarrow$ ProvenanceCheck($c_i$, $E_i$)
6: $h_i \leftarrow \sigma(w^T[\text{logit } s_1, s_2, s_3, \rho_i] + b)$
7: $f_i \leftarrow 1[h_i \ge \tau(\alpha)]$; Log($c_i$, $E_i$, $s_{1..3}$, $h_i$, $\tau$, $\alpha$, $f_i$)
8: **end for**
9: **if** $\Sigma f_i = 0$ **return** deliver; **elif** $\Sigma f_i / n \le \theta$ **return** deliver-with-flags; **else return** abstain

## V. Analytical Results

*Proposition 1 (fusion ceiling).* Let $L$ = logit $u$ be the latent log-odds of evidence support, with class-conditional means differing by $\Delta$ and common within-class variance $v$, and let $K$ detectors observe $z_k = \beta_k L + \varepsilon_k$ with independent $\varepsilon_k \sim N(0, \sigma_k^2)$. The discriminability of the Fisher-optimal linear fusion is

$$d_f^2 = \Delta^2 / (v + 1/S), \;\; S = \Sigma_k \, \beta_k^2 / \sigma_k^2, \tag{5}$$

with AUROC $\Phi(d_f/\sqrt{2})$ under the Gaussian approximation. *Sketch.* The precision-weighted combination $\Sigma_k (\beta_k/\sigma_k^2) z_k / S = L + \eta$ is sufficient for $L$ with Var($\eta$) = $1/S$; its class separation is $\Delta$ and its within-class variance $v + 1/S$. ∎ Adding detectors drives $1/S$ to zero but never lifts $d_f^2$ above $\Delta^2/v$; channels that read $L$ through a similar lens share noise and gain less than (5) suggests.

*Proposition 2 (miss-rate sandwich).* If calibration and test claims are exchangeable and fused scores are almost surely distinct, rule (4) satisfies

$$\alpha - 1/(m+1) \le \mathrm{P}(f_i = 0 \mid y_i = 1) \le \alpha. \tag{6}$$

This is the split-conformal argument [9] applied to the positive class: the rank of a fresh ungrounded score among the $m$ calibration scores is uniform on $\{1, \dots, m+1\}$. The guarantee is marginal over the calibration draw; one calibration set can realise a miss rate above $\alpha$ (Fig. 4(a)).

*Proposition 3 (response-level risk).* For a response with $k$ ungrounded claims, the probability that at least one is delivered unflagged is at most $k\alpha$, and equals $1 - (1 - \alpha)^k$ under independence; a response-level tolerance $\alpha_R$ is met by $\alpha = \alpha_R / \bar{k}$, with $\bar{k}$ an upper quantile of the ungrounded count per response. The bound is loose when a shared cause, such as a missed cited passage, correlates the misses.

## VI. Simulation Setup

*Epistemic status.* No language model, retriever or NLI model is executed. Public corpora with span-level hallucination labels in regulated domains do not exist at the scale needed to calibrate a conformal gate, and the four-type taxonomy is not annotated anywhere. The study therefore uses a generative model in which each channel observes a latent evidence-support variable through a noisy, type-dependent lens, and the ordering of methods is a consequence of that model. What the simulation can establish is whether the architecture behaves as the propositions predict, what each component contributes under stated assumptions, and what the guarantee costs in review load. Parameters are in Table II and the released code (Python 3.12, NumPy 2.5, SciPy 1.18; seed 20261025).

*Generative model.* For each of three domains, 24,000 claims are drawn; a claim is ungrounded with base rate $\pi_d$ and, if so, typed from a domain mixture. Latent support $u$ is Beta-distributed so that intrinsic contradictions have low support, extrinsic claims intermediate support, and citation and numeric errors high support (the text largely matches; only the attribution or a value is wrong). The query-level retriever returns the supporting passage with probability $r_d$. Each channel observes

$$s_k = \sigma(\beta_k \, \text{logit } \tilde{u} + \varepsilon_k), \;\; \tilde{u} = u + \delta_{k,t}(1 - u), \tag{7}$$

where $\beta_k$ is the channel's fidelity, $\varepsilon_k \sim N(0, \sigma_k^2)$ and $\delta_{k,t} \in [0, 1]$ is the blind spot of channel $k$ for type $t$, the fraction of the gap to full support it wrongly credits. When the passage is absent, evidence-dependent channels see $\tilde{u}$ scaled by 0.45; token confidence is generator-internal and unaffected. Self-consistency draws agreement from Binomial(5, 0.5 + 0.45(2$\tilde{u}$ − 1))/5; the numeric and provenance checks are deterministic detectors with type-specific true-positive and small false-positive rates, active only when evidence is present. The blind spots encode three published observations: NLI is weak on numeric perturbation [6], [15], self-consistency is blind to confidently repeated errors [4], and token probability is loosely tied to factuality after instruction tuning [10]. They are assumptions, not measurements.

*Protocol.* Each domain is split 50/50 into calibration and test for the main comparison; the conformal experiment uses a fit / calibrate / test split repeated over 200 calibration draws. Metrics are AUROC with 100-draw bootstrap intervals, AUPRC, recall at 5% false-positive rate overall and per type, expected calibration error (ECE, 10 bins), and the empirical FNR and coverage of the gate.

*Mid-build correction.* The first configuration gave token-confidence and passage-NLI AUROCs of 0.52 and 0.61, below ranges reported on RAGTruth [13] and SummaC [6]: the citation and numeric blind spots made those claims appear *more* supported than grounded ones, and a retrieval miss collapsed every channel, including the retrieval-agnostic token signal, to a fixed floor. Both were corrected before any result below was produced (the original configuration is kept in the bundle's change log); the correction raised baseline AUROCs by 0.05–0.08 and narrowed the proposed method's margin.

TABLE II. SIMULATION PARAMETERS AND ASSUMED RELATIVE COST

| Quantity | Value / range | Basis |
|---|---|---|
| Base rate $\pi_d$ (legal / med. / compl.); claims per response | 0.22 / 0.18 / 0.15; 5–7 | [11], [12] |
| Retrieval recall $r_d$; re-retrieval recovery $\rho$ | 0.78 / 0.82 / 0.85; 0.60 | assumed |
| Latent $u \sim$ Beta: grounded / intr. / extr. / cit. / num. | (8,2) / (2,6) / (3,5) / (5,4) / (6,3) | design |
| Channel ($\beta$, $\sigma$): token / passage-NLI / claim-NLI | (0.55,1.2) / (1.0,0.85) / (1.2,0.7) | assumed |
| Blind spot $\delta$ (intr., extr., cit., num.): passage-NLI | 0, 0.20, 0.45, 0.55 | [6], [15] |
| Blind spot $\delta$: claim-NLI; self-consistency | 0, 0.15, 0.35, 0.40; 0.10, 0.45, 0.40, 0.30 | [4], [15] |
| Numeric check TPR / FPR; provenance TPR / FPR | 0.85 / 0.03; 0.80 / 0.04 | assumed |
| Cost (generator passes / claim): token / p-NLI / SC / CGV-RC | 0 / 0.15 / 5.0 / 0.35 | assumed |

## VII. Results and Discussion

### A. Detection Performance

Table III reports the main comparison and Fig. 2 the legal domain. Averaged over domains, CGV-RC reaches AUROC 0.887 against 0.681 for passage-level NLI, 0.663 for self-consistency and 0.599 for token confidence, with bootstrap intervals that overlap no baseline in any domain. At a 5% false-positive rate CGV-RC recovers 54.50% of ungrounded claims, self-consistency 34.23% and passage-level NLI 20.49%; in deployment terms (Fig. 3(a)), delivering the 60% lowest-risk claims leaves an ungrounded fraction near 5% under CGV-RC, about 15% under passage-level NLI and 22% with no verification.

Fig. 2(b) shows where the gain comes from. On intrinsic contradictions self-consistency and CGV-RC are close (0.78 versus 0.80), since a contradicted claim is also one the generator is unsure about. The separation opens on the classes the baselines were not built to see: citation errors (0.57 versus at most 0.19) and numeric drift (0.35 versus 0.17). Extrinsic claims are the weakest class for every method including the proposed one (0.41 in legal): nothing in the three channels separates a claim that is true but absent from the corpus from one that is false and absent, though in a compliance setting neither can be defended.

TABLE III. DETECTION PERFORMANCE, TEST SPLIT (AUROC WITH 95% BOOTSTRAP CI; REC@5% = RECALL AT 5% FPR; COST IN GENERATOR PASSES PER CLAIM)

| Domain / method | AUROC [95% CI] | AUPRC | Rec@5% | Cost |
|---|---|---|---|---|
| *Legal* Token conf. | 0.608 [0.595, 0.619] | 0.307 | 0.109 | 0.00 |
| Passage NLI | 0.694 [0.683, 0.703] | 0.430 | 0.199 | 0.15 |
| Self-consist. ($k$=5) | 0.655 [0.645, 0.668] | 0.379 | 0.336 | 5.00 |
| **CGV-RC** | **0.885 [0.878, 0.892]** | **0.712** | **0.524** | **0.35** |
| *Medical* Token conf. | 0.604 [0.590, 0.616] | 0.252 | 0.093 | 0.00 |
| Passage NLI | 0.690 [0.678, 0.701] | 0.397 | 0.229 | 0.15 |
| Self-consist. ($k$=5) | 0.676 [0.664, 0.688] | 0.353 | 0.366 | 5.00 |
| **CGV-RC** | **0.897 [0.890, 0.903]** | **0.688** | **0.580** | **0.35** |
| *Compliance* Token conf. | 0.586 [0.575, 0.599] | 0.198 | 0.091 | 0.00 |
| Passage NLI | 0.658 [0.647, 0.676] | 0.311 | 0.187 | 0.15 |
| Self-consist. ($k$=5) | 0.657 [0.647, 0.667] | 0.289 | 0.324 | 5.00 |
| **CGV-RC** | **0.880 [0.871, 0.887]** | **0.612** | **0.531** | **0.35** |

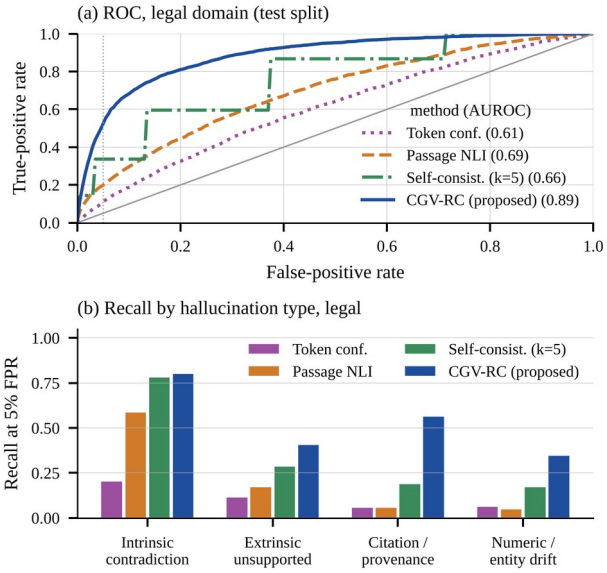



Fig. 2. Detection performance, legal domain. (a) ROC on the test split; the dotted line marks 5% FPR. (b) Recall at 5% FPR by hallucination type.

TABLE IV. ABLATION ON POOLED DATA (THREE DOMAINS, TEST SPLIT; REC@5% = RECALL AT 5% FPR)

| Configuration | AUROC | AUPRC | Rec@5% |
|---|---|---|---|
| **CGV-RC (full)** | **0.886** | **0.663** | **0.529** |
| − without re-retrieval | 0.877 | 0.617 | 0.427 |
| − without numeric/entity check | 0.817 | 0.609 | 0.468 |
| − without provenance check | 0.843 | 0.619 | 0.471 |
| − claim-NLI only | 0.756 | 0.486 | 0.290 |
| + self-consistency (k=5) | 0.887 | 0.665 | 0.527 |
| + passage-NLI + token conf. | 0.889 | 0.671 | 0.543 |

### B. Ablation

Table IV removes or adds one component at a time on pooled data. The numeric and entity check is the most valuable single component (−0.069 AUROC when removed), then the provenance check (−0.043). Re-retrieval moves AUROC by only 0.009 but lifts recall at 5% FPR by 10.14 points, since its

effect falls on grounded claims that would otherwise be false alarms at a strict threshold. Two additions did not pay: five-sample self-consistency raises AUROC by 0.001 at roughly fifteen times the cost, and passage-level NLI plus token confidence adds 0.003; both read the same latent variable through a lens similar to the claim-level NLI channel, as Proposition 1 anticipates.

### *C. Sensitivity to Retrieval Recall*

Fig. 3(b) sweeps query-level retrieval recall from 0.5 to 1.0. Passage-level NLI improves steadily (0.649 to 0.732), confirming that many of its errors are false alarms on grounded claims whose evidence was never retrieved. CGV-RC is comparatively flat (0.875 to 0.896); the open-marker curve attributes this to re-retrieval, worth 0.024 AUROC at $r = 0.5$ and nothing at $r = 1$.

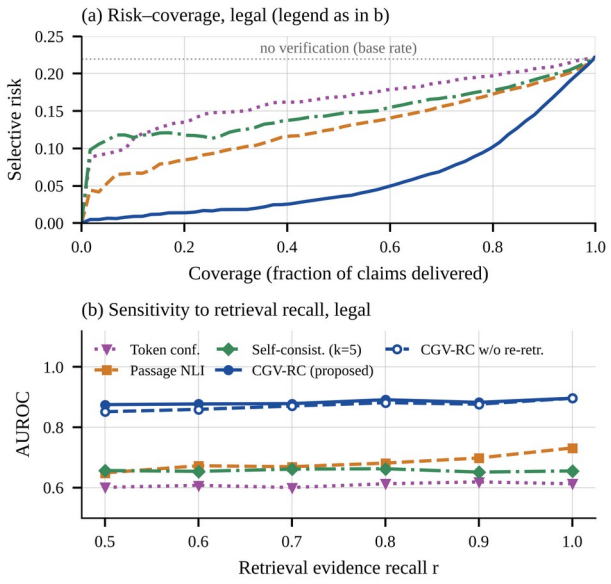



Fig. 3. (a) Selective risk against coverage, legal domain. (b) AUROC against query-level retrieval recall $r$; the open-marker curve removes re-retrieval.

### *D. Validation of the Propositions*

*Proposition 1.* In an idealised configuration without blind spots or retrieval loss, the latent model gives $\Delta = 2.18$ and $v = 0.70$. Equation (5) predicts channel discriminabilities of 0.72, 1.78, 2.13 against measured 0.71, 1.78, 2.13, and fused $d_f = 2.26$ against 2.25; predicted fused AUROC 0.945 compares with 0.950 measured, below the ceiling of 0.967.

*Proposition 2.* Fig. 4(a) plots the empirical claim-level FNR over 200 calibration draws against $\alpha \in \{0.01, 0.02, 0.05, 0.10, 0.20\}$ per domain. The mean lies within 0.001 of $\alpha$ in all fifteen configurations (legal, $\alpha = 0.05$ sits at 0.051, inside Monte Carlo error) and the lower bound is indistinguishable from $\alpha$ at $m \approx 1{,}400$–$1{,}800$. The spread matters: at $\alpha = 0.05$ the 95th percentile of the realised FNR is 0.062 in the medical domain, so a conservative operator should use a smaller $\alpha$ than the policy target or recalibrate on a schedule.

*Proposition 3.* Fig. 5(a) compares the empirical probability that a legal response of $n$ claims contains at least one ungrounded claim delivered unflagged with the union bound $E[k]\alpha$. The bound tracks the curve within 0.01 up to $n = 7$ and is exceeded by at most 0.016 at $\alpha = 0.10$, $n = 1$ (0.116 against 0.100), the single-draw effect of Proposition 2 rather than a failure of the bound.

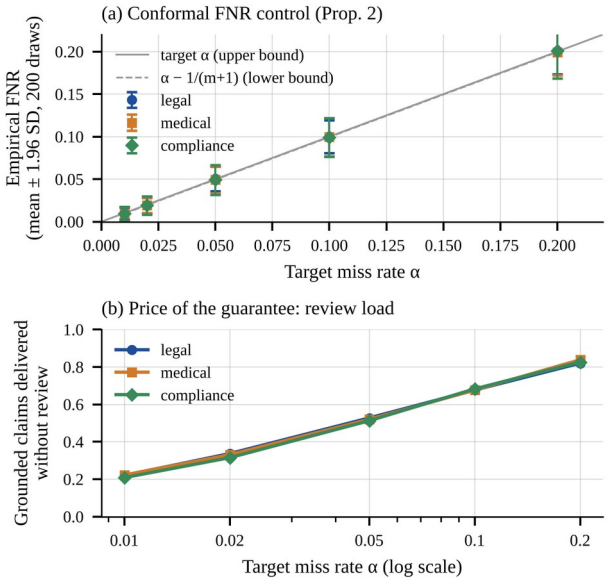



Fig. 4. Conformal gate. (a) Empirical claim-level FNR against target $\alpha$, mean ± 1.96 SD over 200 calibration draws. (b) Grounded claims delivered without review as $\alpha$ tightens.

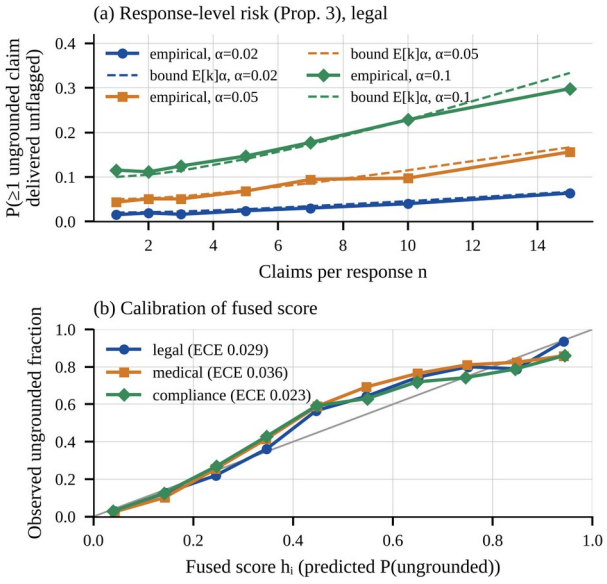



Fig. 5. (a) Response-level failure probability for legal responses of $n$ claims against the union bound of Proposition 3. (b) Reliability diagram of the fused score.

### *E. The Price of the Guarantee*

Fig. 4(b) reports what the gate costs. At $\alpha = 0.05$ about 52% of grounded medical claims pass without review and 57% of all claims are flagged; at $\alpha = 0.01$, closer to what a clinical or litigation setting would demand, coverage falls to 22% and 82% of claims go to a reviewer. That is the honest reading of a distribution-free FNR constraint on a detector with AUROC

near 0.9: the remaining ranking error is paid for in review load. Two mitigations follow: a second-tier verifier with a larger context budget applied to flagged claims only, and per-response control with $\alpha$ set from an upper quantile of $k$ (Proposition 3). The fused score is well calibrated (Fig. 5(b); ECE 0.023–0.036), so $h_i$ can be shown to a reviewer as a probability.

## VIII. Comparative Analysis and Key Outcomes

Tables I and III set the overlay against the baseline families on the axes a procurement review would use. On cost, CGV-RC sits between passage-level NLI and self-consistency, its 0.35 passes per claim dominated by short-input NLI calls that batch well. On complexity, it adds a retrieval round-trip and three scorers but no learned component beyond a four-coefficient logistic regression, which keeps its decisions explicable. On reliability, it is the only configuration with a stated and verified miss-rate bound. An organisation can therefore choose $\alpha$ from its risk appetite and read the review load from Fig. 4(b) before deployment.

## IX. Limitations

Every result is conditional on the channel model of Section VI. The blind-spot parameters are informed by the literature but were set by the authors, and the ordering of methods would change if, say, modern NLI models proved more robust to numeric perturbation than assumed. The simulation also assumes correct claim decomposition, independent claims within a response, exchangeability between calibration and deployment data and a corpus that does not drift; the last two would void the conformal guarantee silently. Cost figures are nominal, extrinsic hallucinations are not solved, and the review load at strict $\alpha$ is high. The next step is to run the overlay on RAGTruth [13] and on domain corpora with span-level labels, with real components in place of the simulated channels.

## X. Conclusion

This paper has cast hallucination detection for RAG in regulated domains as selective prediction under a false-negative-rate constraint and proposed CGV-RC, a claim-level overlay that re-retrieves evidence per claim, fuses an entailment score with numeric and provenance checks, and gates the result with a split-conformal threshold while logging an audit record per claim. Three propositions on the fusion ceiling, the miss-rate sandwich and the response-level union bound are confirmed numerically. Under a stated generative model the overlay raises mean AUROC from 0.681 to 0.887 at about 7% of the cost of five-sample self-consistency and holds the claim-level miss rate at target in every configuration tested. It also reports what did not work: self-consistency adds nothing once claim-level evidence channels are present, extrinsic claims stay largely undetected, and a 1% miss-rate target sends four claims in five to a reviewer. Validation on labelled corpora is the next step.

## References


[1] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in *Proc. NeurIPS*, 2020, pp. 9459–9474.

[2] Z. Ji et al., “Survey of hallucination in natural language generation,” *ACM Comput. Surv.*, vol. 55, no. 12, art. 248, 2023.

[3] S. Min et al., “FActScore: Fine-grained atomic evaluation of factual precision in long form text generation,” in *Proc. EMNLP*, 2023, pp. 12076–12100.

[4] P. Manakul, A. Liusie, and M. J. F. Gales, “SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models,” in *Proc. EMNLP*, 2023, pp. 9004–9017.

[5] O. Honovich et al., “TRUE: Re-evaluating factual consistency evaluation,” in *Proc. NAACL-HLT*, 2022, pp. 3905–3920.

[6] P. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst, “SummaC: Re-visiting NLI-based models for inconsistency detection in summarization,” *Trans. Assoc. Comput. Linguist.*, vol. 10, pp. 163–177, 2022.

[7] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, “RAGAS: Automated evaluation of retrieval augmented generation,” in *Proc. EACL Syst. Demonstrations*, 2024, pp. 150–158.

[8] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. New York, NY, USA: Springer, 2005.

[9] A. N. Angelopoulos and S. Bates, “Conformal prediction: A gentle introduction,” *Found. Trends Mach. Learn.*, vol. 16, no. 4, pp. 494–591, 2023.

[10] L. Kuhn, Y. Gal, and S. Farquhar, “Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation,” in *Proc. ICLR*, 2023.

[11] M. Dahl, V. Magesh, M. Suzgun, and D. E. Ho, “Large legal fictions: Profiling legal hallucinations in large language models,” *J. Legal Anal.*, vol. 16, no. 1, pp. 64–93, 2024.

[12] V. Magesh et al., “Hallucination-free? Assessing the reliability of leading AI legal research tools,” *J. Empir. Legal Stud.*, 2025, early access.

[13] C. Niu et al., “RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models,” in *Proc. ACL*, 2024, pp. 10862–10878.

[14] J. J. Cherian, I. Gibbs, and E. J. Candès, “Large language model validity via enhanced conformal prediction methods,” in *Proc. NeurIPS*, 2024.

[15] L. Tang, P. Laban, and G. Durrett, “MiniCheck: Efficient fact-checking of LLMs on grounding documents,” in *Proc. EMNLP*, 2024, pp. 8818–8847.

[16] European Parliament and Council, “Regulation (EU) 2024/1689 (Artificial Intelligence Act),” *Off. J. Eur. Union*, L 2024/1689, Jul. 2024.