International Journal of Research and Development in Engineering Sciences
ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/
Regular Issue, Vol. 8, Issue. 6 (November – December) , 2026 , Pages: 97 – 105
Received: 19 March 2026 ; *Accepted:* 10 November 2026 ; *Published:* 21 November 2026
https://doi.org/10.63328/IJRDES-V8RI6P10
https://www.youtube.com/watch?v=CARX33CC1Co
https://archive.org/details/jsp-ijrdes-v8ri6p10
Original Paper : Applied Research – Experimental Research Paper

Trustworthy Retrieval-Augmented Generation Through Confidence Calibration, Signal Fusion, and Conformal Risk Control

Ujval Mutyala

Department of Computer Technology, Eastern Illinois University , Illinois , USA

* *Corresponding Author : Ujval Mutyala ; ujvalroy@gmail.com*

Abstract: A retrieval-augmented system that answers every question with the same assurance is dangerous exactly when its retriever has returned the wrong passages. We study how to attach a trustworthy confidence to an answer, and how to use it to decide when to answer and when to hand over to a person. We built a simulated retrieval-augmented pipeline in which answer quality depends on how much correct and misleading evidence is retrieved, and in which the system exposes three families of signals: the generator's own token confidence, the agreement among ten sampled answers, and retrieval scores. We prove that recalibration cannot change how well confidence ranks right and wrong answers, that a calibrated score predicts its own selective accuracy, that calibration under a shift in evidence quality is bounded by how much the shift changes the mix of latent states, and that conformal risk control guarantees a target error rate under exchangeability. In simulation, the raw confidence was badly overconfident (expected calibration error 0.075). Temperature scaling did not help at all (0.075), because the miscalibration was a bias and not a slope; Platt scaling, with an intercept, reduced it to 0.008. Stacking all signals gave the best ranking (AUROC 0.971 against 0.951) and Brier score (0.057). The calibrated score predicted selective accuracy to within 0.008. Conformal risk control met its target exactly in-domain (risk 0.050 for a target of 0.05) and failed under shift (0.078 after a retrieval shift, 0.290 after a retrieval and generator shift), and re-tuning the threshold on 300 labelled target questions restored it (0.047). The aggregate calibration error also hid a serious problem: when the retrieved evidence was dominated by distractors, the system's confidence averaged 0.42 after temperature scaling, against an actual accuracy of 0.05. All results are simulated.
Keywords: Retrieval-Augmented Generation, Calibration, Selective Prediction, Conformal Risk Control.

1. Introduction

Retrieval-augmented generation (RAG) gives a language model a short stack of retrieved passages and asks it to answer from them [1]. When the retriever finds the right passage, answers are usually good. When it finds near-misses, the model can produce a fluent, confident and wrong answer, and nothing in the text of the answer reveals the difference. For any use where a wrong answer costs something, what matters is a number that says how far to trust each answer, and a rule for when to say "I am not sure". Two ideas are standard here. Calibration asks that an answer given with 80% confidence be right 80% of the time [3]. Selective prediction asks the system to abstain on the cases it is least sure of, trading coverage for accuracy [13]. Both are well studied for classifiers. A RAG system is different in an important way: its errors have two sources that look the same from the outside. The generator can misread good evidence, and the retriever can supply bad evidence. A confidence signal that tracks one may not track the other, and a shift in the data usually changes one of them.

We study this with a simulation that makes the two sources explicit and lets us check every claim against known ground truth. We make five contributions.

- A simulated RAG pipeline with a latent notion of evidence quality, and three families of observable confidence signals (Section 3). Four results (Section 4): recalibration does not change ranking; calibrated scores predict their own selective accuracy; calibration under shift is bounded by the change in the mix of latent states; and conformal risk control guarantees a target risk under exchangeability.
- A comparison of seven calibration methods, in-domain

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P10

and under two kinds of shift, including the finding that temperature scaling can fail completely (Section 5).

- An experiment on selective answering with a conformal guarantee, showing that the guarantee holds in-domain, fails under shift, and can be restored with a small labelled sample from the new domain.
- A conditional analysis showing that overall calibration error can hide large overconfidence in the cases where retrieval failed.

Everything is simulated, with generator behaviour modelled and not run. We did not use a real language model or a real corpus.

2. Background and Related Work

Retrieval-augmented generation. Lewis and colleagues combined a retriever with a sequence-to-sequence generator for knowledge-intensive tasks [1]. Position matters: models use information at the start and end of a long context better than in the middle [2], a feature our simulation includes.

Calibration. Modern neural networks tend to be over-confident, and a single temperature parameter fixes much of this for classifiers [3]. Platt scaling fits a slope and an intercept [4], and isotonic regression fits a monotone step function [5]. The expected calibration error summarises miscalibration over confidence bins [6]. The Brier score [7] is a proper scoring rule that decomposes into calibration and refinement terms [8, 9]. Under dataset shift, calibration degrades [10].

Uncertainty for language models. Sampling several answers and taking the most common one improves reasoning accuracy [11], and agreement among samples is a usable confidence signal. Semantic uncertainty groups samples by meaning before computing entropy [12]. Our agreement signal is the simplest version of this idea.

Selective prediction and conformal guarantees. Geifman and El-Yaniv trade coverage for risk with a reject option [13]. Conformal risk control chooses a threshold so that the expected value of a monotone loss is below a target, with a finite-sample guarantee under exchangeability [14].

What we add. Each tool is known. We examine them together in a setting with two sources of error, show where each one breaks, and quantify what is hidden by an aggregate metric.

3. Research Methodology and Gap

3.1. A Simulated RAG Pipeline

Each question has a retrievability r drawn from Beta(4, 2) (mean 0.67) in-domain. The retriever returns five passages, each *supporting* the right answer with probability $0.8r$, a *distractor* that supports a wrong answer with probability $0.5(1 - r)$, and otherwise *irrelevant*. Retrieval scores are Gaussian with means 2.0 (support), 1.4 (distractor) and 0 (irrelevant) and unit variance, so distractors look almost as relevant as support. Passages are presented in score order and weighted by position, w = (0.30, 0.18, 0.12, 0.15, 0.25), which is high at the start and end. Let E_c and E_w be the weighted amounts of supporting and distracting evidence. A single sampled answer is correct with probability $\sigma(6(E_c - E_w)-0.8+h)$, where $h \sim N(0, 1)$ stands for what the generator knows without retrieval. The system draws ten samples (wrong answers are spread over three alternatives) and returns the most common answer, so correctness is that of the modal answer. In-domain accuracy is 79%. Figure 1 shows the pipeline.

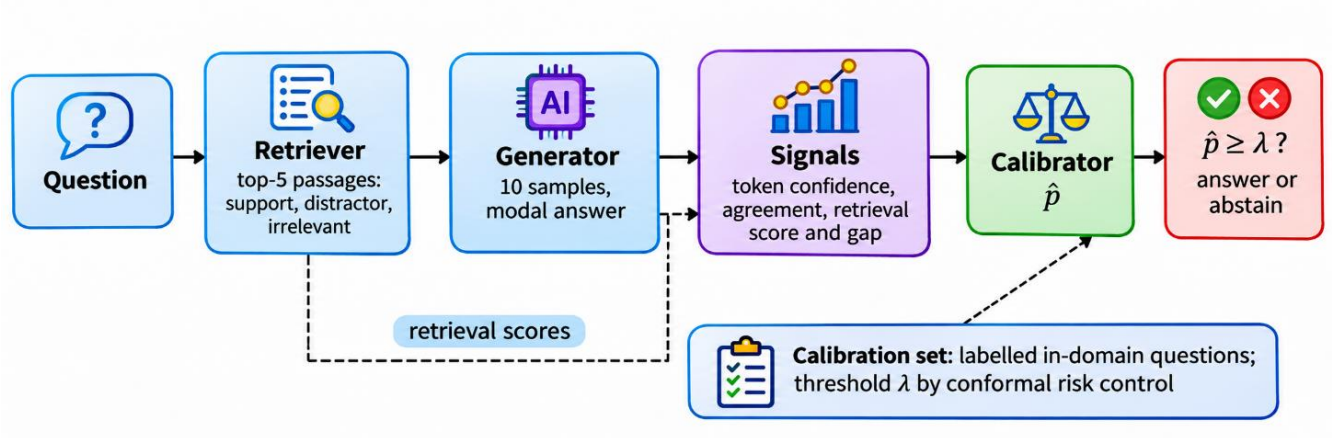

Figure. 1 Pipeline and calibration. The system exposes three families of signals. A calibrator maps them to a probability of being correct, and a threshold chosen on labelled calibration data decides whether to answer or abstain.

3.2. Signals

Three families of signals are observable. *Token confidence*: a logit equal to $2.0 + 0.8\,\ell + \varepsilon$, where ℓ is the logit of a single sample being correct and $\varepsilon \sim N(0, 1)$, which is monotone in true quality but biased upward, meaning overconfident. *Agreement*: the fraction of the ten samples that give the modal answer. *Retrieval*: the top score, the gap between the first and second scores, and the mean of the top three.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P10

3.3. Shifts

Retrieval shift: retrievability is drawn from Beta(2, 3) (mean 0.4), so retrieval returns fewer supporting passages and more distractors, and accuracy falls to 51%. **Generator shift**: retrieval is as before, but the token-confidence bias grows (intercept 3.2) and its slope falls to 0.5, as would happen with a new model or prompt; accuracy is unchanged at 79% but the confidence is less informative. **Both**: the two together, with accuracy 51%.

3.4. Calibrators and metrics

All calibrators are fitted on 2,000 labelled in-domain questions. *Raw*: the sigmoid of the token logit. *Temperature scaling*: divide the logit by a fitted T. *Platt*: logistic regression on the token logit (slope and intercept). *Isotonic*: monotone regression on the raw confidence. Three *stacks* are logistic regressions on groups of signals: retrieval only, agreement only, agreement with retrieval (no token confidence), and all signals. Metrics: expected calibration error (ECE, 15 equal-width bins), Brier score, and AUROC for separating right from wrong answers. We use 30,000 test questions per condition and 20 independent calibration sets, and report means.

4. Proposed System

Let $y \in \{0, 1\}$ indicate a correct answer and $s \in [0, 1]$ a confidence score.

Proposition 1 (Recalibration and ranking)**.** *The Brier score decomposes as* $\mathrm{BS} = U - \mathrm{Res}(s) + \mathrm{Rel}(s)$*, with* $U = y\text{✓}(1 - y\text{✓})$ *a constant, resolution* $\mathrm{Res} = \mathrm{E}[(g(s) - y\text{✓})^2]$ *and reliability* $\mathrm{Rel} = \mathrm{E}[(s - g(s))^2]$*, where* $g(s) = \mathrm{E}[y \mid s]$*. Replacing* s *by the injective map* $g(s)$ *removes the reliability term and leaves resolution unchanged, so it lowers the Brier score by exactly* $\mathrm{Rel}(s)$*. Any strictly increasing map of* s *leaves the AUROC unchanged.*

Proof. This is Murphy's decomposition [8]. If g is injective, conditioning on $g(s)$ is the same as conditioning on s, so Res is unchanged, and the new reliability is $\mathrm{E}[(g(s) - g(g(s)))^2] = 0$. The AUROC depends only on the ranking of scores, which a strictly increasing map preserves.

So recalibration can fix the numbers but cannot change which answers are ranked as more likely to be right. Only a better *signal* can do that.

Proposition 2 (Selective accuracy of a calibrated score)**.** *If* $\mathrm{E}[y \mid s] = s$ *then for every threshold* τ *with* $\mathrm{P}(s \geq \tau) > 0$*,*

$$E[y \mid s \geq \tau] = E[s \mid s \geq \tau]. \quad (1)$$

Proof. By the tower property, $E[y \mid s \geq \tau] = E[\, E[y \mid s] \mid s \geq \tau \,] = E[s \mid s \geq \tau]$.

A calibrated system can therefore forecast its own accuracy at any coverage by averaging its scores over the questions it answers.

Proposition 3 (Calibration under a change in evidence quality). *Let* Z *be a latent state (such as the quality of the evidence), with* $a_z = P(y = 1 \mid Z = z)$ *unchanged by the shift, while the mix of states given the score changes from* $P_S(z \mid s)$ *to* $P_T(z \mid s)$*. If the score is calibrated at the source, then at the target*

$$|E_T[y \mid s] - s| \leq (\max_z a_z - \min_z a_z) \cdot \mathrm{TV}(P_T(\cdot \mid s), P_S(\cdot \mid s)). \quad (2)$$

Proof. $E_T[y \mid s] - s = \Sigma_z a_z (P_T(z \mid s) - P_S(z \mid s))$, since $s = \Sigma_z a_z P_S(z \mid s)$. The weights difference sums to zero, so subtracting min a_z from every a_z changes nothing, and the sum is at most (max a_z − min a_z) times the total-variation distance.

The bound is small when the score is nearly a sufficient statistic for the state, so that the shift cannot change the posterior over states given the score. It is large when the score does not separate the states, which is the case for a signal that does not see the retriever.

Proposition 4 (Conformal risk control). *Let* (s_i, y_i)*,* $i \leq n + 1$*, be exchangeable, and let* $L_i(\lambda) = \text{✓}[s_i \geq \lambda,\ y_i = 0]$ *be the loss of answering when the score is at least* λ *and the answer is wrong. This loss is non-increasing in* λ*. Choose*

$$\lambda\text{✓} = \min\{\, \lambda : (1/(n+1))\, (\, \Sigma_{i \leq n} L_i(\lambda) + 1\,) \leq \alpha \,\}. \quad (3)$$

Then $E[L_{n+1}(\lambda\text{✓})] \leq \alpha$*.*

Proof. Let λ' be the minimal λ with $(1/(n+1))\, \Sigma_{i \leq n+1} L_i(\lambda) \leq \alpha$, which depends on all $n + 1$ points symmetrically. Since $L_{n+1} \leq 1$, we have $\Sigma_{i \leq n} L_i(\lambda) + 1 \geq \Sigma_{i \leq n+1} L_i(\lambda)$, so the condition in Eq. (3) is stronger and $\lambda\text{✓} \geq \lambda'$. Because L is non-increasing, $L_{n+1}(\lambda\text{✓}) \leq L_{n+1}(\lambda')$. By exchangeability, $E\, L_{n+1}(\lambda') = E[(1/(n+1))\, \Sigma_{i \leq n+1} L_i(\lambda')] \leq \alpha$.

This is the finite-sample argument of Angelopoulos and colleagues [14] for this particular loss. Note what it controls: the expected rate of *answering and being wrong*, measured over all questions, so abstentions count as zero loss. It is not the error rate among the answered questions. It also needs exchangeability, which a shift breaks.

5. Results and Analysis

5.1. Calibration Methods

Table 1 gives the metrics and Figure 2 shows reliability diagrams. In-domain, the raw token confidence is badly overconfident: its ECE is 0.075, and the curve lies far below the diagonal. **Temperature scaling does not help** (0.075). The reason is that the overconfidence is a *bias*. The confidence logit is $2.0 + 0.8\ell$ while the true log-odds of the system's answer being right are roughly ℓ plus an offset in the other direction, and a single scalar T can only shrink the logit toward zero, which pulls high and low confidences together but cannot shift them down. Platt scaling, which fits an intercept as well as a slope, reduces the ECE to 0.008 and the Brier score from 0.088 to 0.069. Isotonic regression reaches 0.013 with a similar Brier score, but needs more data (below). None of

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P10

these three changes the AUROC (0.951, 0.951 and 0.951; isotonic is 0.949, a tiny drop from ties), as Proposition 1 says. Ranking comes from the signals. Retrieval signals alone rank poorly (AUROC 0.669), because a distractor scores almost as high as a supporting passage. Agreement alone ranks well (0.928), and stacking all signals ranks best (0.971), with the best Brier score (0.057). The stack has slightly higher ECE than Platt (0.015 against 0.008) because it fits more parameters from the same 2,000 questions.

Table. 1 Calibration metrics for each method in-domain, under retrieval shift and under generator shift (means over 20 calibration sets; 30,000 test questions per condition). ECE and Brier: lower is better. AUROC: higher is better.

Method	In-domain			Retrieval shift			Generator shift		
	ECE	Brier	AUROC	ECE	Brier	AUROC	ECE	Brier	AUROC
Raw	0.075	0.088	0.951	0.182	0.152	0.944	0.138	0.150	0.834
Temperature	0.075	0.086	0.951	0.176	0.149	0.944	0.124	0.145	0.834
Platt	0.008	0.069	0.951	0.045	0.100	0.944	0.100	0.134	0.834
Isotonic	0.013	0.070	0.949	0.045	0.100	0.942	0.100	0.134	0.821
Retrieval only	0.009	0.156	0.669	0.239	0.291	0.649	0.010	0.154	0.671
Consistency only	0.034	0.089	0.928	0.142	0.178	0.850	0.035	0.090	0.925
Cons. + retrieval	0.033	0.087	0.932	0.129	0.170	0.858	0.035	0.088	0.929
All signals	0.015	0.057	0.971	0.037	0.085	0.956	0.091	0.108	0.925

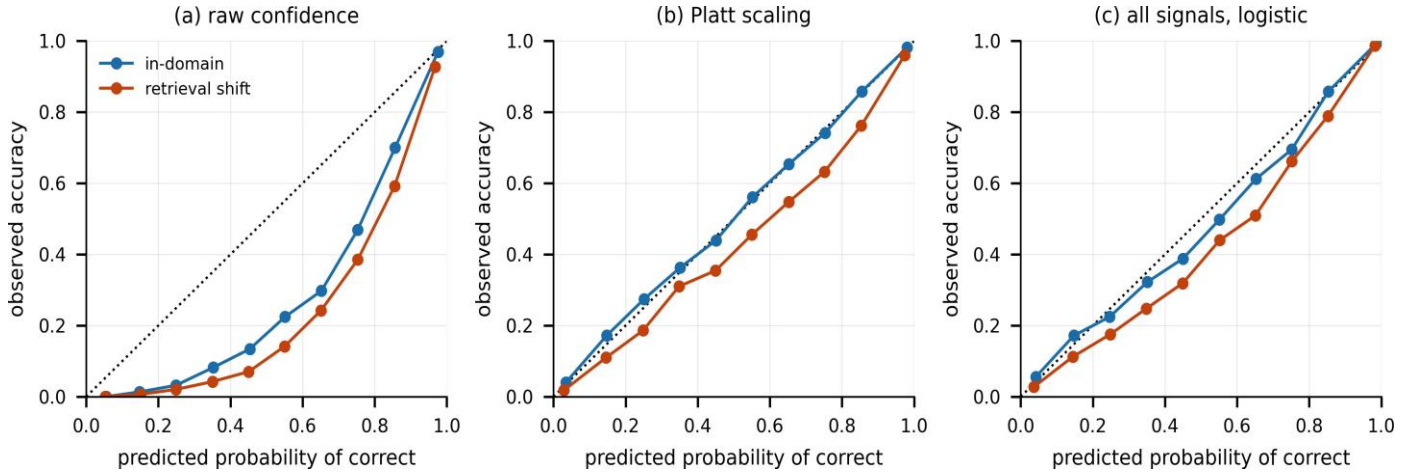

Figure. 2 Reliability diagrams (accuracy against predicted probability) in-domain and under retrieval shift. The raw confidence is far below the diagonal. Platt scaling and the stack fix the in-domain curve but sit below the diagonal again under retrieval shift.

5.2. How much Calibration Data is needed

Figure 4 varies the number of labelled calibration questions from 50 to 5,000. Temperature scaling stays at an ECE of about 0.075 for every sample size, as expected from its structural failure. The other methods improve with data. With only 50 questions, Platt scaling has an ECE of 0.042 and the stack 0.040, and isotonic regression is worse at 0.058. With 5,000 questions they reach 0.007, 0.015 and 0.009. The Brier score tells a different story: the stack beats Platt from 100 questions on, and by 5,000 its Brier score is 0.056 against 0.069, because it uses better signals. So the simple Platt calibrator is the best choice for the number only, and the stack for the ranking. A few hundred labelled questions suffice for either.

5.3. Calibration under Shift

Under *retrieval shift* every method gets worse (Table 1, Figure 2). Platt and isotonic go from about 0.01 to 0.045 and 0.045, and the stack from 0.015 to 0.037, the best of the lot. Raw confidence and temperature scaling reach 0.182 and 0.176. The reliability diagrams show the direction of the error: the shifted curves lie below the diagonal, so the system is more overconfident than before, as Proposition 3 predicts for a score that does not see all of the latent state. Figure 5 sweeps the share of shifted questions in the test set. The ECE of every method rises roughly linearly with the share, from 0.014 for the stack at 0% to 0.041 at 100%, with the stack and isotonic staying well below the raw

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P10

confidence and temperature scaling (0.184 and 0.176).

Under *generator shift* the picture reverses. The stack that includes token confidence has an ECE of 0.091 and its AUROC drops from 0.971 to 0.925, since it leans on the signal that changed. Platt (0.100) and isotonic (0.100) are no better. The calibrators that use only retrieval scores (0.010) or agreement and retrieval without token confidence (0.035) are barely affected. But those same calibrators do worse under retrieval shift (0.239 for retrieval alone and 0.129 for agreement plus retrieval).

No signal set is robust to both kinds of shift: signals that come from the generator are fragile when the generator changes, and signals that come from the retriever are fragile when the corpus changes.

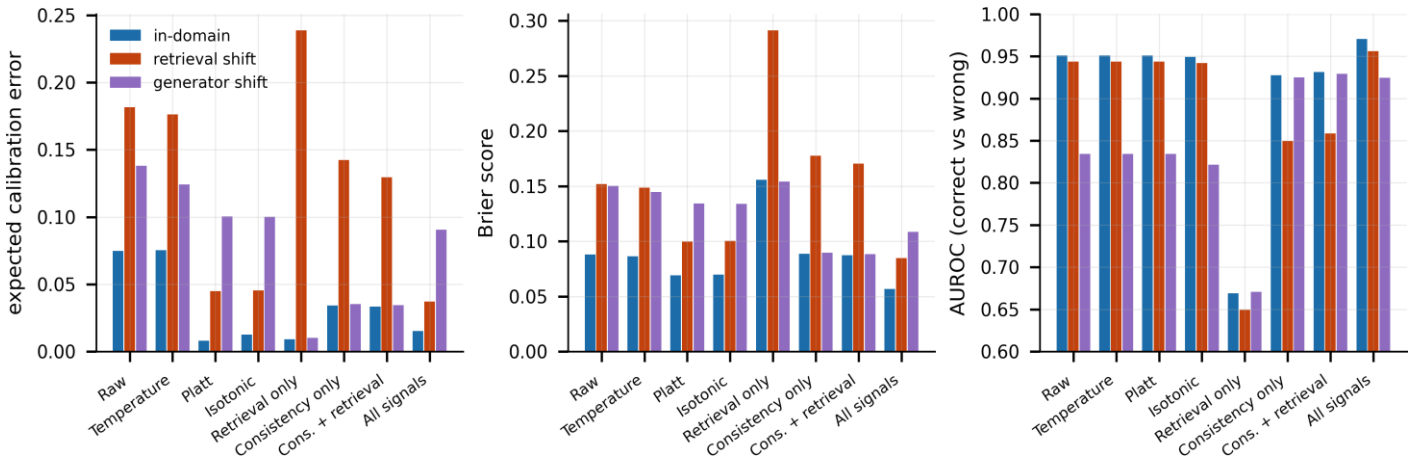

Figure. 3 Calibration error, Brier score and AUROC by method and condition. Generator-side signals (token confidence) fail under generator shift. Retrieval-side signals do not, but they rank poorly.

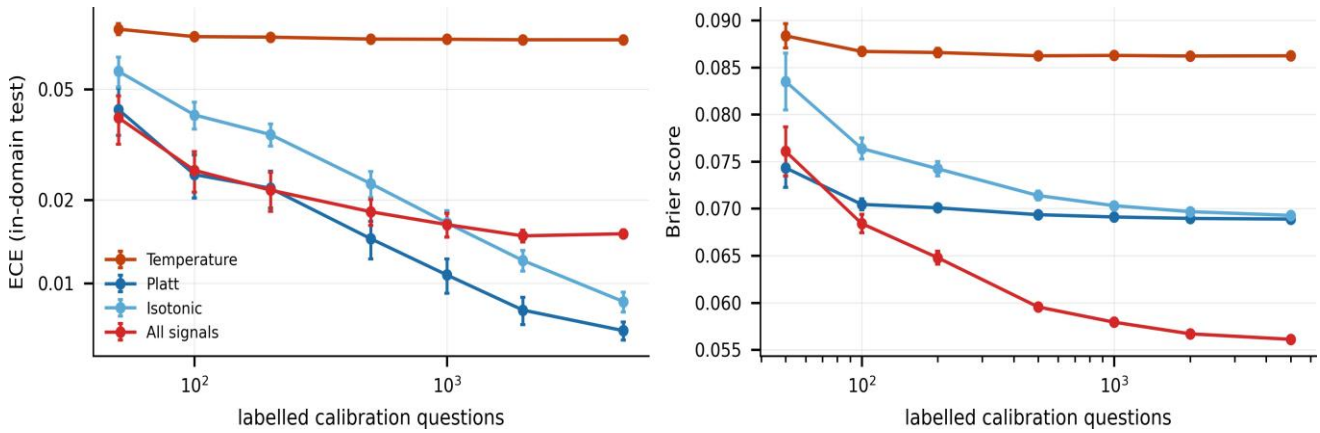

Figure. 4 Effect of the number of labelled calibration questions on in-domain ECE (left, log scale) and Brier score (right). Error bars are 95% CIs over 30 calibration sets.

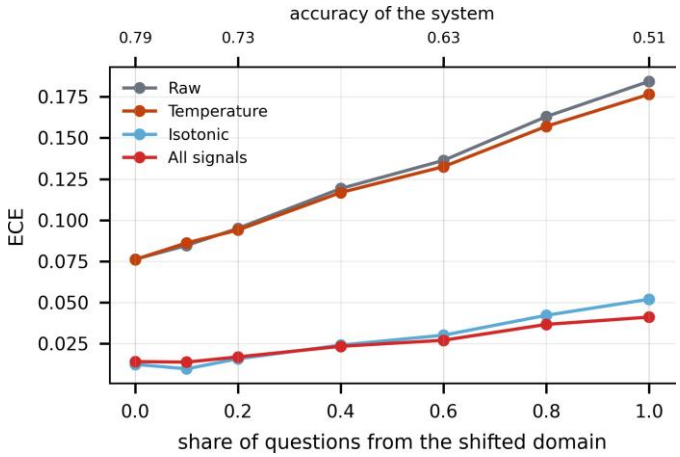

Figure. 5 ECE as the share of questions from the shifted (retrieval) domain grows. The top axis shows the system's accuracy.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P10

5.4. Selective answering

Figure 6 gives risk–coverage curves: the error rate among the answered questions as the system abstains on more of them. At full coverage the system is wrong on 21.0% of questions. Answering only the 80% most confident questions by the stack cuts the error to 5.7%, against 9.1% with agreement alone, 6.6% with the raw confidence and 17.0% with retrieval alone. At 60% coverage the stack, agreement and raw confidence are all below 2%, since the questions they rank highest are nearly all right. The stack's advantage is concentrated at coverage between 70% and 90%, where it separates the questions with mixed evidence from the supported ones.

Proposition 2 says a calibrated system can predict this. Table 2 compares the mean calibrated score of the answered questions with their true accuracy at six coverage levels, and the largest gap is 0.008. This matters operationally. A team can quote a selective accuracy from the scores alone, before labelling any test questions.

Table. 2 Predicted selective accuracy (mean calibrated score of answered questions) against actual, in-domain, for the stack.

Coverage	Predicted	Actual	Gap
50%	0.996	1.000	0.003
60%	0.992	0.999	0.006
70%	0.979	0.988	0.008
80%	0.942	0.943	0.001
90%	0.874	0.869	0.005
100%	0.794	0.790	0.004

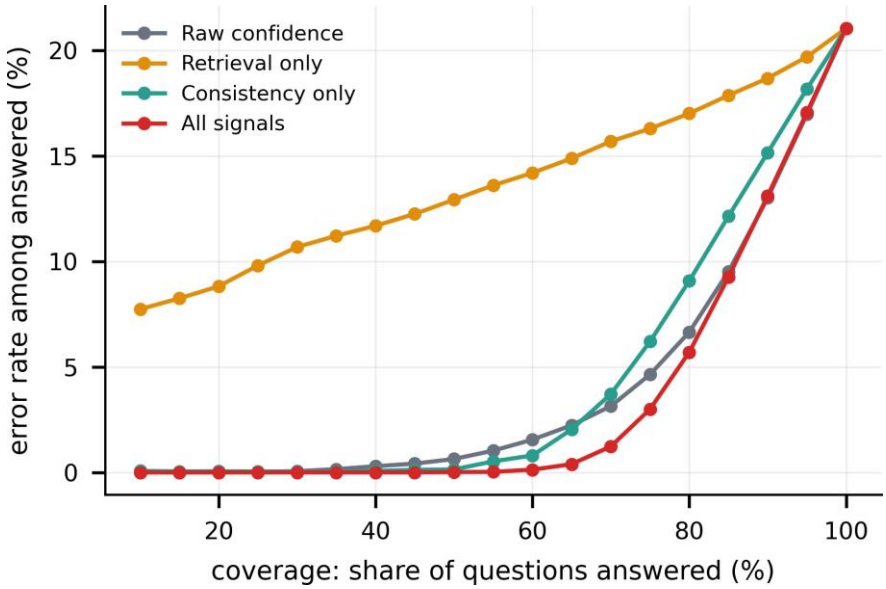

Figure. 6 Error rate among answered questions against coverage, for four confidence signals (in-domain).

5.5. A guarantee, and where it breaks

We chose the answer threshold by conformal risk control (Proposition 4) on 1,000 labelled in-domain questions, for targets of 0.02, 0.05 and 0.10, and measured the achieved risk on 4,000 fresh questions, over 300 independent calibration sets (Table 3, Figure 7a). In-domain the guarantee holds with almost no slack: the mean achieved risk is 0.021, 0.050 and 0.100 for the three targets, with coverage 0.74, 0.80 and 0.88. Under retrieval shift the same thresholds give risks of 0.030, 0.078 and 0.173, which exceed the targets, by as much as 70%. With a retrieval and generator shift the guarantee fails by a wide margin: the risk is 0.290 for a target of 0.05. The system gives up some coverage under retrieval shift (0.54 at target 0.05, down from 0.80), since fewer questions are confidently right, but not enough.

The remedy is cheap. Re-tuning only the *threshold* on labelled questions from the shifted domain restores the guarantee (Table 4, Figure 7b). With 100 shifted-domain labels the mean risk is 0.044, with 300 it is 0.047 and with 1,000 it is 0.050, against the target 0.05. Small samples are conservative on average (0.034 with 50 labels) but have high variance: the 90th-percentile risk is 0.066 with 50 labels and 0.065 with 300. In practice this means a team should keep a small, current labelled sample from each deployment domain and recompute the threshold when the domain changes.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P10

Table. 3 Conformal risk control with thresholds fitted in-domain (1,000 questions, 300 repeats, 4,000 test questions each). Risk is the share of questions answered wrongly. Coverage is the share answered.

Target	Test Domain	Risk (Mean)	Risk (90th pct.)	Coverage
0.02	In-Domain	0.021	0.028	0.74
0.02	Retrieval Shift	0.030	0.039	0.45
0.02	Both Shifts	0.197	0.226	0.66
0.05	In-Domain	0.050	0.059	0.80
0.05	Retrieval Shift	0.078	0.094	0.54
0.05	Both Shifts	0.290	0.312	0.78
0.1	In-Domain	0.100	0.113	0.88
0.1	Retrieval Shift	0.173	0.198	0.67
0.1	Both Shifts	0.384	0.404	0.89

Table. 4 Re-tuning the threshold on labelled retrieval-shift questions (target risk 0.05; 300 repeats).

Shifted labels	Risk (mean)	Risk (90th pct.)	Coverage
50	0.034	0.066	0.45
100	0.044	0.067	0.48
300	0.047	0.065	0.49
1000	0.050	0.060	0.49

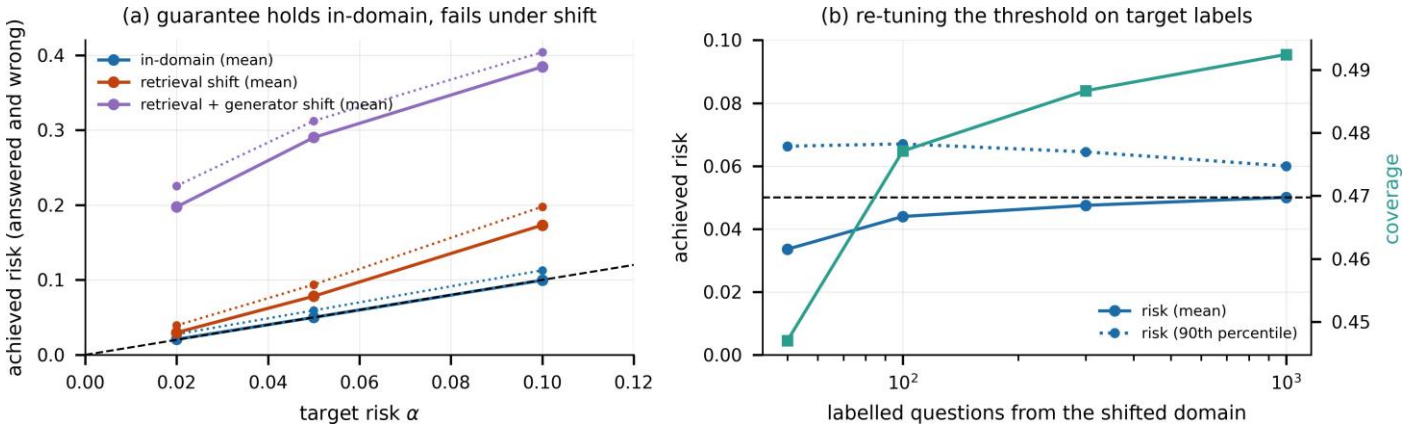

Figure. 7 (a) Achieved risk against target in-domain and under shift (solid: mean over 300 calibrations; dotted: 90th percentile). The dashed line is the target. (b) Re-tuning the threshold on labelled shifted-domain questions restores the target.

5.6. What the average hides

All the metrics above are averages over the question mix. They can conceal systematic overconfidence in the cases that matter most. We split the in-domain test questions by the latent evidence state: *support dominant* (67% of questions), *mixed* (24%) and *distractor dominant* (10%), where the misleading passages outweigh the supporting ones. Figure 8 compares predicted confidence with observed accuracy in each state.

In the support-dominant state, all calibrators are accurate: predictions near 0.94 against an accuracy of 0.97. In the distractor-dominant state, accuracy is 0.05, but the temperature-scaled confidence is 0.42 on average, Platt gives 0.20 and even the stack gives 0.19, about three times too high. These are the cases where the retriever has failed and the answer is most likely to be confidently wrong. They are a small share of questions (10%), so their miscalibration barely shows in the overall ECE. This reflects the information available: none of the signals sees directly whether retrieved passages are distractors, since a distractor looks almost as relevant as support. Reporting calibration by evidence state, where it can be known, would be a worthwhile diagnostic.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P10

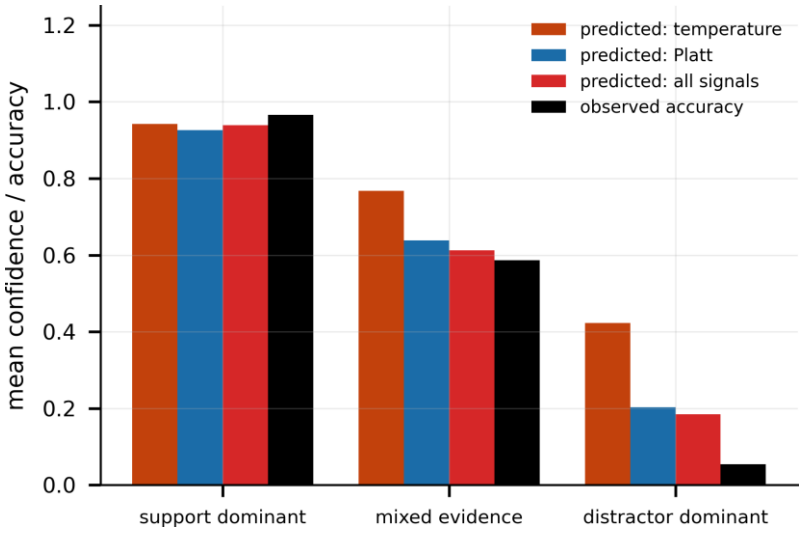

Figure. 8 Predicted confidence and observed accuracy by latent evidence state. In the rare distractor-dominated state all calibrators remain overconfident.

Discussion and Limitations

Simulated pipeline. The generator is a formula and not a language model. The relationship between evidence and accuracy, the form of the overconfidence and the independence of signals given the latent state are all assumptions. Real systems have richer failure modes, such as answers that are right for the wrong reason, partially correct answers and unanswerable questions. The qualitative lessons should be tested on a real system before they are relied on.

Binary correctness. We scored each answer as right or wrong. Open-ended generation has degrees of correctness, and agreement among samples needs a notion of when two answers are the same, which is itself a hard problem [12].

Temperature scaling. Our finding that it fails depends on the bias in our simulated overconfidence. For classifiers whose logits are over-scaled and not shifted, a temperature works well [3]. The lesson is to check which kind of miscalibration is present and not that temperature scaling is poor.

Stacking and the cost of labels. The best-ranked signal set needed labelled data from the deployment domain. We estimated ECE in 15 equal-width bins, which has known biases, particularly for small samples.

Exchangeability. The conformal guarantee needs the calibration and test questions to be exchangeable. Real traffic drifts, and we showed that even a moderate retrieval shift breaks the guarantee. Monitoring (see Paper 3 of this series on drift detection) and periodic re-tuning of the threshold are part of making the guarantee usable.

Costs of abstaining. Abstention sends the question to a person or to another system, which has a cost and a capacity (see Paper 2 on reviewer fatigue). Setting α is a business choice and not a statistical one.

6. Conclusion

A RAG system needs two kinds of care to know when it is wrong. The numbers have to be calibrated, and in our simulation that needed a bias term, not just a temperature. The signals have to be good enough to rank right and wrong answers, and no recalibration can add that. A conformal threshold turns the confidence into a guarantee on the rate of wrong answers, which held exactly in-domain and failed under shift until the threshold was re-tuned on a few hundred labelled questions from the new domain. The aggregate error also hid the worst overconfidence, in the few cases where retrieval returned misleading passages. The next steps are to test these findings on real RAG systems, to find signals that detect distractor-dominated retrieval, and to monitor for the shifts that invalidate the guarantee.

References

[1]. P. Lewis et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020.

[2]. N. F. Liu et al. Lost in the middle: how language models use long contexts. *Transactions of the ACL*, 12:157–173, 2024.

[3]. C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

[4]. J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, 1999.

[5]. B. Zadrozny, C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *KDD*, 2002.

[6]. M. P. Naeini, G. F. Cooper, M. Hauskrecht. Obtaining well

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P10

calibrated probabilities using Bayesian binning. In *AAAI*, 2015.

[7]. G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[8]. A. H. Murphy. A new vector partition of the probability score. *J. Applied Meteorology*, 12(4):595–600, 1973.

[9]. M. H. DeGroot, S. E. Fienberg. The comparison and evaluation of forecasters. *The Statistician*, 32:12–22, 1983.

[10]. Y. Ovadia et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *NeurIPS*, 2019.

[11]. X. Wang et al. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

[12]. L. Kuhn, Y. Gal, S. Farquhar. Semantic uncertainty: linguistic invariances for uncertainty estimation in natural language generation. In ICLR, 2023.

[13]. Y. Geifman, R. El-Yaniv. Selective classification for deep neural networks. In NeurIPS, 2017.

[14]. A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, T. Schuster. Conformal risk control. In ICLR, 2024.

A. Parameters

Questions: retrievability Beta(4, 2) in-domain, Beta(2, 3) under retrieval shift. Five passages; support probability $0.8r$, distractor $0.5(1 - r)$. Retrieval score means 2.0, 1.4, 0; unit variance. Position weights (0.30, 0.18, 0.12, 0.15, 0.25). Per- sample correctness logit $6(Ec - Ew) - 0.8 + h$, $h \sim N(0, 1)$. Ten samples; wrong answers split (0.5, 0.3, 0.2). Token confidence logit: intercept 2.0, slope 0.8, noise sd 1.0 in-domain; intercept 3.2, slope 0.5, noise sd 1.5 under generator shift. Calibration sets 2,000 (main), 50–5,000 (sweep, 30 repeats); test sets 30,000. ECE: 15 equal-width bins. Conformal risk control: 1,000 calibration questions, 4,000 test, 300 repeats. Code and figures are in the supplementary archive. All figures were generated from code at 600 dpi and printed at their stored width or smaller.

How To Cite:

Ujval Mutyala , " Trustworthy Retrieval-Augmented Generation Through Confidence Calibration, Signal Fusion, and Conformal Risk Control " , International Journal of Research and Development in Engineering Sciences (IJRDES), Volume 8 , Issue 6 , pp 97 – 105 , November 2026.
CrossRef DOI: https://doi.org/10.63328/ijrdes-v8ri6p10

Declaration

Conflicts of Interest: The authors declare no conflict of interest.

Author Contribution: All authors wrote the main manuscript text and also consent to the submission.

Ethical approval: Not applicable.

Consent to Participate: All authors consent to participate.

Funding: Not applicable, and No funding was received

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Personal Statement: We declare with our best of knowledge that this research work is purely Original Work and No third party material used in this article drafting. If any such kind material found in further online publication, we are responsible only for any judicial and copyright issues.

Generative AI Statement: The authors declare that generative AI was not used in the preparation or creation of this manuscript. The alternative text (alt text) accompanying the figures in this article was generated by Frontiers with the assistance of artificial intelligence. Reasonable efforts have been made to ensure the accuracy and appropriateness of the generated alt text, including review by the authors wherever possible. If you identify any inaccuracies or issues, please contact the editorial team.

Publisher's Note: The statements, opinions, and claims presented in this article are exclusively those of the authors and do not necessarily reflect the views of their affiliated institutions, the publisher, editors, or reviewers. Any products discussed or evaluated, as well as any claims made by their manufacturers, are not warranted, approved, or endorsed by the publisher.

Acknowledgements

We thank everyone who inspired our work

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com