When Abstention Gates Overtrust Noisy Evidence: An Empirical Audit of Calibration and Thresholding in Lightweight RAG Evidence-Sufficiency Models

Kushaagra Patel

AI Explainability and Accountability (AIEA) Lab, University of California, Santa Cruz

Sierra High School; Diablo Valley College; San Joaquin Delta College, Manteca, CA, USA | Research mentorship: Professor Leilani H. Gilpin

Abstract

Retrieval-augmented generation (RAG) systems retrieve passages that could overlap with the query, but lack the evidence necessary to fully address it. In this work, I examine whether a simple evidence sufficiency gate can discern fully informative contexts from one of three regimes that I construct of insufficient evidence contexts. Using 1,000 prefiltered HotpotQA questions, I construct four different contexts per question (one sufficient, three insufficient; 4,000 contexts in all) and compare dense cosine similarity, TF-IDF logistic regression, and a four-feature logistic regression. On the final 200-question set, unmodified cosine similarity produces false sufficiency rates (FSRs) of 44.5% and 19.0% for the partial evidence and lexical trap regimes. Platt scaling lowers cosine ECE from 0.1619 to 0.0267; the chosen operating point produces FSRs of 54.5% and 25.5% for those regimes. The four-feature model obtains AUROC 0.8583, with 12.5% and 8.5% FSRs, respectively, at a Y=1 threshold of 0.50. Applying the estimated Y=1 threshold over 600 precalibrated insufficient contexts yields 15.5% and 10.5% FSRs over the test set; applying only the partial evidence and lexical trap contexts produces 8.5% and 7.5%. These results are descriptive and apply to the constructed question-clustered contexts. Since there are several contexts per question that may be interrelated, and since the procedure for estimating the decision threshold at this level has not been validated, these results do not provide any distribution free finite sample guarantee. The study examines evidence sufficiency, not LLM hallucination rates downstream.

Keywords: Retrieval-Augmented Generation, Evidence Sufficiency, Selective Prediction, Abstention, Probability Calibration, Conformal Methods, Explainable AI

1. Introduction

In retrieval-augmented generation (RAG) systems, the language model responses are grounded using retrieved documents [1]. The retrieval passage may be irrelevant, insufficient, or even misleading [3,4]. If the evidence available is insufficient to answer the question, a downstream language model may generate a spurious response. In an attempt to mitigate the risk without using an LLM verifier for each query, a classifier or thresholding can predict the presence (Y=1) or absence (Y=0) of sufficient evidence in the context C for question Q [10,11].

A prior RAG-filtering prototype, called Sieve, led to the following tighter look at lightweight gates for multi-hop questions. These questions imply the necessity to retrieve facts, which could be split among several different passages. The two medical scenarios below illustrate this issue and serve as retrieval mistakes. These examples do not belong to any clinical tests or evaluations performed on a medical database.

1. Near miss retrieval (2_Partial): A hypothetical clinical decision support system has a two-part question regarding renal dosing adjustment of an antibiotic used to treat Lyme carditis. The retrieved context mentions the antibiotic but lacks the passage, where one can find the dosing part. A similarity-based gate would find the context good enough because it relates to the topic of the question even if there was no reasoning step made.

2. Lexical Trap (3_LexicalTrap): A hypothetical retriever provides the context that contains the terms of "renal dosing," "antibiotic" and "Lyme" yet the context belongs to another patient cohort and use case. A lexical or density similarity gate might get confused with the topicality overlap for answer-retrieval evidence. This case represents the illustration of a potential downstream risk; this paper does not investigate clinical mistakes.

I measure two complementary metrics for each gate:

$$\text{Normal Sufficiency (TPR)} = P(\hat{Y} = 1 \mid Y = 1) = \frac{\sum_{i=1}^{N} \mathbb{I}(\hat{Y}_i = 1 \wedge Y_i = 1)}{\sum_{i=1}^{N} \mathbb{I}(Y_i = 1)}$$

$$\text{False Sufficiency Rate (FSR)} = P(\hat{Y} = 1 \mid Y = 0) = \frac{\sum_{i=1}^{N} \mathbb{I}(\hat{Y}_i = 1 \wedge Y_i = 0)}{\sum_{i=1}^{N} \mathbb{I}(Y_i = 0)}$$

- Normal Sufficiency: true positive rate on complete contexts, P(Ŷ=1 | Y=1)

- False-Sufficiency Rate (FSR): the ratio of insufficient contexts that are incorrectly marked as sufficient, P(Ŷ=1 | Y=0). FSR can be regarded as the gate's level estimation of false acceptance and a proxy for downstream risks, though FSR does not show the hallucination rate in generated answers.

Calibration performance is evaluated by means of the expected calibration error (ECE) and the Brier score, and inter-calibrations at different operating points are done via negative-class score quantiles. Formally speaking, to achieve finite sample guarantees, the exchangeability assumption between the calibration and test units must be met. However, since the benchmark contains several contexts per question, row-level guarantees are not considered formally.

Three research questions define the scope of the analysis:

RQ1 (calibration vs. operating-points behavior): How does calibration affect Platt scaling and Brier score? How does examination of various operating points influence Normal Sufficiency and FSRs for regime-specific cases of 2_Partial and 3_LexicalTrap?

RQ2 (thresholds in noise mixtures): What is the difference between negative-class quantile thresholds calculated over all insufficient contexts and thresholds estimated over more difficult insufficient regimes when applying it for the test partition evaluation?

RQ3 (interpretable features): What is the relationship between interpretability features and the output of a white-box model? How does normal-sufficiency/FSR trade-off manifest at standard and quantile thresholds?

2. Related work

Abstention and selective prediction. The selective prediction paradigm ensures abstention when the classifier is not confident enough [10,11]. It is essential to emphasize that in RAG models, the retrieval relevance and answer sufficiency are related but different aspects. While retrieval relevancy of a particular passage is a guarantee of its topical relevance, topical relevance does not guarantee completeness because there might be incomplete passage with some facts necessary for building multi-hop answer. Dense embeddings and lexical retrieval scores are common components in the retrieval process pipeline [12,13]; however, they cannot guarantee that facts chains are complete.

Probability calibration and operating points behavior. Probability calibration refers to the transformation of classifiers' outputs to probabilities reflecting actual probabilities of events. Platt Scaling uses a sigmoid transformation of the out-of-the-bag probabilities [5]. Model's calibration is usually measured through expected calibration error (ECE), Brier score, and some other metrics; these two metrics represent an orthogonal pair [6]. Aggregating calibration metrics leads to hiding subgroup errors; in my case, cosine-based ECE drops from 0.1619 to 0.0267 when applying Platt scaling. The FSRs for the tested operating points remain relatively high in spite of that.

Conformal Methods and Distribution Shift. Conformal methods and conformal risk control provide finite-sample results conditional on distributional assumptions that hold under particular sampling schemes [7,8]. However, distributional guarantees under covariate shift may not hold without other assumptions [9]. In the current study, I consider each source question with multiple contexts, leading to partitioning into queries-disjoint sets that do not make the context rows conditionally independent. Thus, the following results for negative-class quantile thresholds are based on the operationally motivated evaluation of operating points. The finite-sample guarantees cannot be provided due to a need in an appropriate construction of the calibrator and corresponding analysis.

3. Methodology and Datasets Construction

3.1 Controlled Four Regimes Benchmark (HotpotQA, N = 4,000)

I build the controlled benchmark using the distractor split of HotpotQA [2] dataset, which is a multihop question-answering dataset, where bridge questions require an integrated answer from several Wikipedia documents. My benchmark includes 1,000 filtered questions. For each question four versions of the context will be generated (1,000 × 4 = 4,000 rows). Each of those rows will be annotated with the binary label $Y \in \{0,1\}$. For each question, four rows will be generated, so that I have four dependent observations related to the same question but not four different independent questions.

1_Sufficient (Y = 1, N = 1,000): generates a single text obtained by concatenation of the two gold-supported paragraphs, containing all the information needed to build an answer. This setting tests for the normal sufficiency $P(\hat{Y} = 1|Y = 1)$.

2_Partial (Y = 0, N = 1,000): forms pairs that include one selected gold setup paragraph and one distractor paragraph. Setup paragraphs were chosen not to contain the answer string by themselves, forming the almost correct context, i.e., the partial information necessary to answer the question independently. This approach corresponds to an operationally motivated

generation rule. Lack of an answer string does not prohibit some semantic leakage through aliases/paraphrasing.

3_LexicalTrap (Y = 0, N = 1,000): In this context regimen, the two most appropriate distractor paragraphs relative to each of the question are chosen using BM25Okapi [13]. Hence, the context has the same keywords but does not include the gold supporting paragraph. In this case, the context is built due to having the common keywords, instead of being the proof of non-related semantic context of the paragraph.

4_RandomNoise (Y = 0, N = 1,000): This context regimen connects two distractor paragraphs taken from another question, by offsetting it by q(i + 137) mod 1,000. Thus, a control unrelated context baseline is set.

3.2 Two-stage Pilot Audit and Contamination Fixes

A 200-example pilot is generated (50 questions × 4 contexts), and then is tested on cross-encoder/nli-deberta-v3-base NLI model. The automatic judge agreed with construction rule labels in 193 out of 200 rows (96.5%): 44/50 for 1_Sufficient, 49/50 for 2_Partial, and 50/50 for both 3_LexicalTrap and 4_RandomNoise. In addition, a manual audit was conducted by the author. The 96.5% represents the agreement between the construction rule and automatic judge labels and not inter-annotator agreement. It should be noted that the pilot does not validate all the rows in the final 4,000-context benchmark.

Two contamination types were found during manual review of the pilot and resolved before the final benchmark construction:

1. Yes/no answer-string contamination: 6 disagreements in 1_Sufficient occurred while comparing questions with "yes" or "no" answer. String matching check worked by the rule that any occurrence of the answer string in the paragraph meant that the particular answer is present. Thus, the benchmark became limited to questions of bridge type and no more yes/no questions were included in it.

2. Partial-context answer-string leakage: The pilot included questions where the chosen initially gold paragraph contained an answer string. The construction rule 2_Partial is fixed by choosing a setup paragraph without this answer string and using it together with distractor paragraph. This addresses the known problem of surface-form leakage, but not the problem of semantic insufficiency for every final example.

3.3 Strict Four-Way Query-Disjoint Split Architecture

Each question_id along with its four variants in context is assigned randomly and uniquely to one partition. This makes sure that the same query never shows up in the train, probability calibration, threshold calibration, or test split. However, the correlation between the four variants per partition stays intact: more precisely, each question provides 3 relevant contexts with Y=0 for the calibration and test partitions.

1_Train (400 questions / 1,600 contexts): Model parameter tuning; optimizing parameters of the TF-IDF feature set; StandardScaler parameters.

2_ProbCal (200 questions / 800 contexts): Parameter selection for Platt scaling based on the separate held-out partition; model scores on the held-out partition are used for that; this partition differs from the one used for threshold calibration.

$$\hat{p}(x) = \sigma(a \cdot z(x) + b) = \frac{1}{1 + \exp(-(a \cdot z(x) + b))}$$

3_ConfCal (200 questions / 800 contexts): 200 sufficient-context items + 600 insufficient-context items (200 per each of the two negative regimes), where the negative-class quantiles are computed over the insufficient contexts.

4_FinalTest (200 questions / 800 contexts): Held-out partition used for evaluating test metrics and drawing graphs; includes 200 contexts of each of the four regimes.

3.4 Models, Features, Calibration Metrics, and Thresholds Estimation

Comparison is done across two model families plus one four-feature logistic regression model.

Baseline 1. Dense cosine similarity. The All-MiniLM-L6-v2 encoder [12] creates 384-dimensional vectors for the question Q and context C. Cosine similarity between them constitutes a gate score; the simple baseline uses the 1_Train Macro-F1 threshold selected (τ=0.5700), and the second variant is Platt-scaled.

Baseline 2. TF-IDF logistic regression. The model has a feature set composed of 5,000 unigrams and bigrams in TF-IDF representation of the question vector X_Q, the context vector X_C, and the elementwise multiplication of the two vectors X_Q ⊙ X_C. Logistic regression has been trained using L2 regularization (C=2.0) and balanced class weights. Both baselines are considered in raw and Platt-scaled versions.

Four-feature logistic regression (white-box model): (1) feature_entity_recall, i.e., fraction of capitalized query tokens in the context (used as an entity recall proxy); (2) feature_token_overlap, i.e., fraction of query content tokens of at least three letters found in the context; (3) feature_negation_count, i.e., number of negation or hedge tokens in the context (not, no, never, none, neither, nor, cannot, n't, failed, lack); and (4)feature_min_hop_sim, minimum cosine similarity of the question with each of retrieved paragraphs with the all-MiniLM-L6-v2. Entity recall stands for the proxy feature here and does not refer to the results of any named-entity-recognition model.

$$\text{min_hop_sim} = \min\left(\mathbf{e}_q^\top \mathbf{e}_{p_1},\ \mathbf{e}_q^\top \mathbf{e}_{p_2}\right)$$

Minimum Hop is explicitly built to assess the minimal level of alignment between queries and contexts at paragraph level. Thus, some cases of partial contexts may still exhibit relatively high query–context alignment, while Minimum Hop is not capable of recognizing that reasoning is needed with respect to the corresponding paragraph.

Metrics for evaluation of calibration accuracy. Expected Calibration Error (ECE) was calculated with 10 probability bins. Brier Score was used for the assessment of probability calibration in 4_FinalTest (800 contexts). These numbers indicate overall statistics over test sets; Forecasted Structural Reliability (FSR) in subgroups will be presented later, since probability calibration gives no information about specific regimes.

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|, \quad \mathrm{Brier} = \frac{1}{N} \sum_{i=1}^{N} (\hat{p}(x_i) - Y_i)^2$$

Empirical quantile for the threshold determination. In order to define a proper cutoff value at reference level α = 0.10, an empirical (1−α)-quantile was calculated for probabilities of a Platt-calibrated model for 3_ConfCal rows, where Y = 0. There were two options for threshold estimation: a pooled estimate across all three insufficiency regimes and an estimate limited to 2_Partial and 3_LexicalTrap contexts only. The reason is that the table rows correspond to questions, thus the row-level thresholds mean empirical analysis of the threshold.

$$\hat{\tau}_\alpha = \mathrm{Quantile}\left(\{\hat{p}_j : Y_j = 0\}_{j=1}^{n_0}, \frac{\lceil (n_0+1)(1-\alpha) \rceil}{n_0} \right)$$

Context from 4_FinalTest (800 contexts based on 200 questions) is predicted as a sufficient one ($\hat{Y}$ = 1), if its calibrated probability p(x) > τ ; otherwise, the gate predicts insufficiency.

4. Results: Performance and Calibration Accuracy (RQ1)

The results of discrimination power, probability calibration and Normal Sufficiency are shown in Table 1A. Table 1B provides FSR values for particular working points of 4_FinalTest (800 contexts from 200 questions).

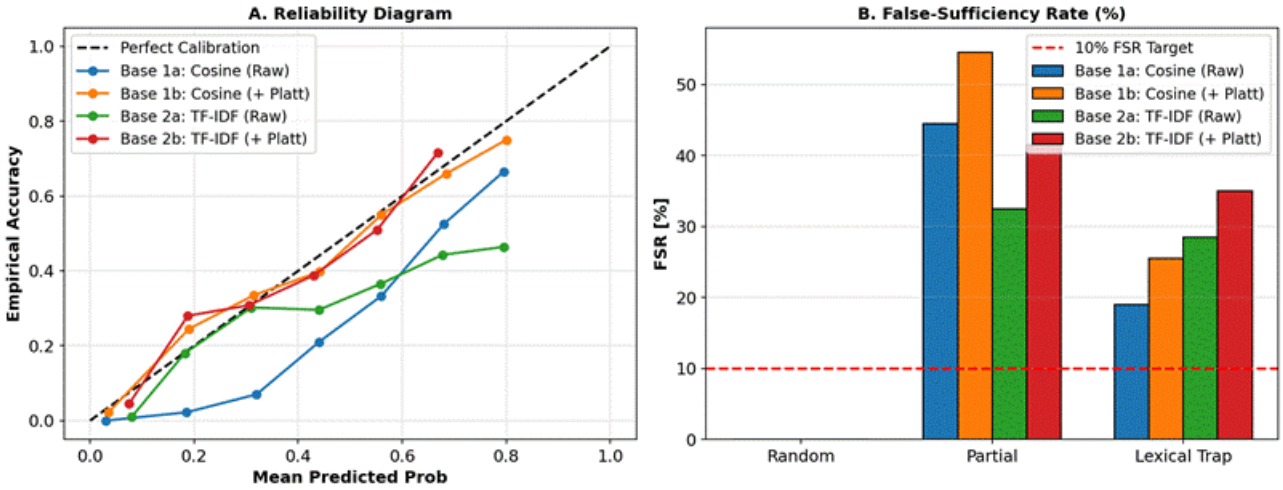
Table 1A. Model discrimination, calibration, and Normal Sufficiency on 4_FinalTest

Model	AUROC	F1	Brier	ECE	Norm. Suff.
Cosine (raw)	0.8298	0.7002	0.1725	0.1619	65.50%
Cosine (+ Platt)	0.8298	0.6943	0.1394	0.0267	74.00%
TF-IDF (raw)	0.7192	0.6118	0.1842	0.1049	54.00%
TF-IDF (+ Platt)	0.7192	0.6105	0.1685	0.0491	64.00%
WB (raw)	0.8583	0.7009	0.1617	0.1503	80.50%
WB (+ Platt), τ=0.500	0.8583	0.7384	0.1266	0.0292	50.50%
WB (+ pooled τ)	0.8583	0.7350	0.1266	0.0292	52.50%
WB (+ hard-neg. τ)	0.8583	0.7248	0.1266	0.0292	44.00%

Table 1B. Overall and regime-specific False-Sufficiency Rates on 4_FinalTest

Model	Overall FSR	Random	Partial	Lexical trap
Cosine (raw)	21.17%	0.00%	44.50%	19.00%
Cosine (+ Platt)	26.67%	0.00%	54.50%	25.50%
TF-IDF (raw)	20.33%	0.00%	32.50%	28.50%
TF-IDF (+ Platt)	25.50%	0.00%	41.50%	35.00%
WB (raw)	29.00%	0.00%	52.50%	34.50%
WB (+ Platt), τ=0.500	7.00%	0.00%	12.50%	8.50%
WB (+ pooled τ)	8.67%	0.00%	15.50%	10.50%
WB (+ hard-neg. τ)	5.33%	0.00%	8.50%	7.50%

Note. F1 denotes Macro-F1; WB = white-box model. Each negative regime contains 200 test contexts; overall FSR is calculated across the 600 insufficient test contexts. Values are point estimates, and context rows from the same question are clustered.



Reliability plots and FSR statistics by regime of the cosine and TF-IDF baselines are shown in Figure 1. In both cases, applying Platt scaling helps reduce the general value of the ECE, although the reported regime-specific FSR values for partial evidence and lexical traps are higher at the tested operating points.

4.1 Baseline performance in near-misses and lexical-traps settings

At all the operating points given in the paper, 0.0% FSR for 4_RandomNoise suggests that it is not too difficult to distinguish irrelevant contexts within the proposed framework. Raw cosine similarity reports 44.5% FSR for 2_Partial contexts. Its similarity score for 2_Partial (0.5536 ± 0.1274) is somewhat close to the similarity score for 1_Sufficient (0.6225 ± 0.1323), which may indicate that similarity to just one paragraph can overshadow the lack of similarity to the second one. The raw TF-IDF baseline gives 32.5% FSR for 2_Partial and 28.5% FSR for 3_LexicalTrap contexts.

4.2 Calibration correction and regime-specific FSR

In case of the raw cosine similarity, its test-set ECE equals 0.1619 and the Brier Score is 0.1725. Application of the Platt scaling to 2_ProbCal leads to a reduction of the cosine ECE to 0.0267 (83.5%). Normal Sufficiency for cosine similarity increases at the chosen operating point from 65.5% to 74.0%. Similar to cosine, the TF-IDF baseline also shows reduction in ECE value (from 0.1049 to 0.0491).

Nevertheless, it is possible to conclude that the general improvement in the ECE is not always accompanied by positive results in terms of FSR. In case of the cosine-score variants, 2_Partial FSR increases from 44.5% to 54.5%, while 3_LexicalTrap FSR goes from 19.0% to 25.5%. In the case of the TF-IDF-score variant, the FSR increases from 32.5% to 41.5% and from 28.5% to 35.0%, correspondingly. As one can see from the above data, there is no consistency between ECE improvements and FSRs. It is critical to take into account the operating points; otherwise, the differences can only be due to the Platt correction.

5. White-Box Features and Quantile thresholding (RQ2 & RQ3)

5.1 White-box feature distribution and logistic regression coefficients (RQ3)

My four-feature logistic regression model demonstrates AUROC = 0.8583 and Macro-F1 = 0.7384 at the considered operating point for the Platt-scaled variant. Standardized coefficients and characteristics of the distributions of the four features in 1_Train are provided in Table 2.

Table 2. White-box feature coefficients and feature distribution characteristics

Feature	β (train)	Sufficient	Partial	Lexical trap	Random noise
Token overlap	+1.2122	0.7538 ± 0.1377	0.6420 ± 0.1381	0.6323 ± 0.1359	0.1824 ± 0.1018
Minimum-hop sim.	+0.8236	0.4145 ± 0.1519	0.3269 ± 0.1439	0.3083 ± 0.1469	−0.0047 ± 0.0788
Entity-proxy recall	+0.7640	0.8282 ± 0.2033	0.7553 ± 0.2327	0.5269 ± 0.3008	0.0690 ± 0.1403
Negation count	−0.3109	0.1400 ± 0.3828	0.2240 ± 0.5021	0.3320 ± 0.5900	0.2770 ± 0.5390

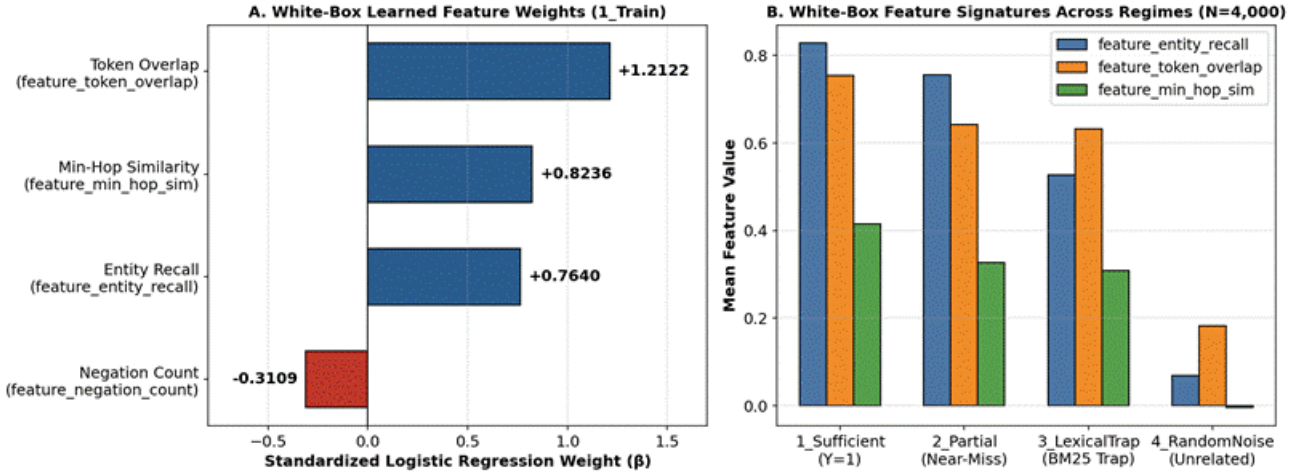


Figure 2. White-box features analysis. Panel A shows standardized logistic regression coefficients calculated from 1_Train. Panel B demonstrates mean values of the features across all four fabricated regimes; this plot is merely descriptive, and causality among the features and the predicted value is not implied here.

These features help comprehend the work of the model: mean min-hop similarity in 2_Partial is smaller compared to the one in 1_Sufficient; correspondingly, mean capital-token recall is lower in 3_LexicalTrap than in 1_Sufficient. Coefficients of the aforementioned features are positive, whereas there is also a negative coefficient assigned to the negation-count feature along with the token-overlap feature.

In respect to the minimum-hop similarity feature, it represents structure not included in the mere context-query similarity measure: the context can match the query in general, but the selected paragraph is not as similar to the query. Mean of feature_min_hop_sim is 0.4145 ± 0.1519 for 1_Sufficient and 0.3269 ± 0.1439 for 2_Partial.

Negation-count feature receives a negative coefficient (-0.3109) fitted in this model, however, the mean of this feature is larger in 3_LexicalTrap than in 1_Sufficient. It is clear that the dependency captured by the model in respect to this specific dataset, since there might be negation terms even in supporting sentences.

5.2 Pooled and hard-negative-restricted quantile thresholds (RQ2)

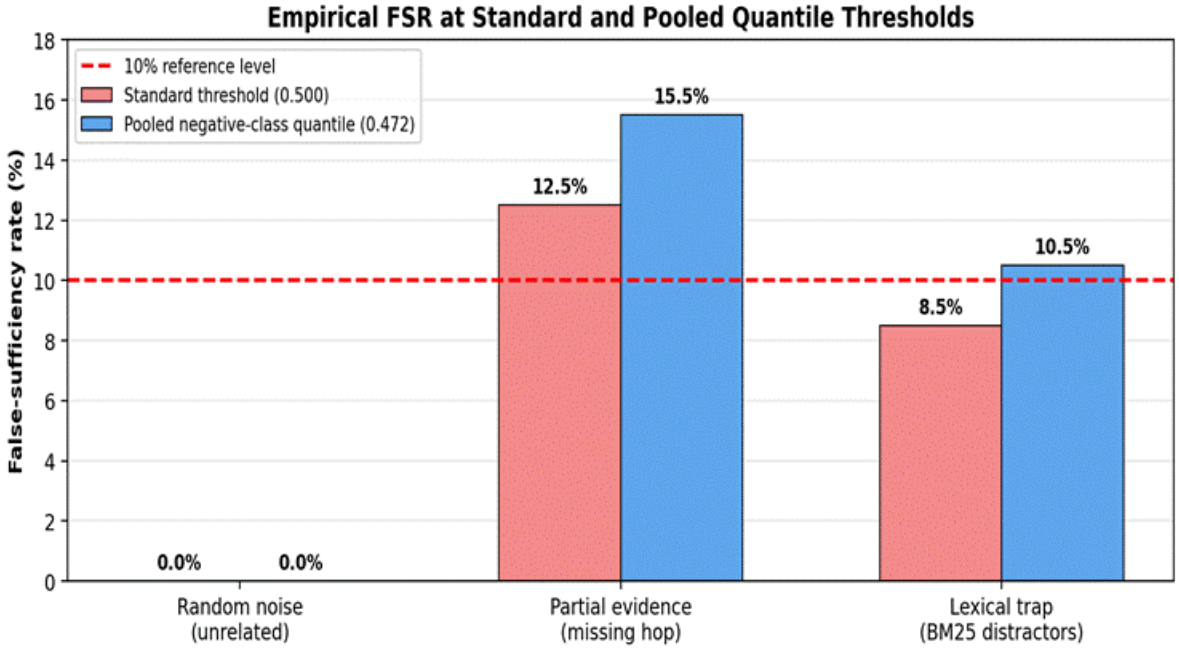


Figure 3. Test FSR at the standard 0.500 threshold and at the pooled negative-class quantile threshold (τ = 0.4722) for the Platt-scaled white-box model. The lower pooled threshold results in higher FSR for 2_Partial (15.5%) and 3_LexicalTrap (10.5%) contexts in this particular test partition.

FSRs for 2_Partial and 3_LexicalTrap contexts are 12.5% and 8.5% at the 0.500 threshold and the overall FSR for all 600 contexts in this test partition is 7.0%. At the pooled quantile threshold of 0.4722, these FSRs become 15.5% and 10.5% correspondingly, and the overall FSR becomes 8.67%.

Pooled threshold is lower than 0.500 because calibration set contains 200 4_RandomNoise rows with low sufficiency values; these easy negatives impact the pooled threshold.

Table 3. Quantiles of calibration sets and corresponding test FSRs

The table 3 below shows the thresholds calculated for different calibration sets and corresponding FSRs on the same test partition. This comparison is performed just to demonstrate sensitivity of calibration sets composition empirically.

Calibration subset	n_0	90th pct.	τ cutoff	Overall test FSR	Partial FSR	Lexical FSR
Random noise only	200	0.0077	0.0080	65.17%	99.00%	96.50%
Lexical trap only	200	0.4579	0.4660	8.83%	16.00%	10.50%
Partial only	200	0.5834	0.5889	3.50%	5.50%	5.00%

Pooled (all negatives)	600	0.4715	0.4722	8.67%	15.50%	10.50%
Hard-negative subset	400	0.5441	0.5453	5.33%	8.50%	7.50%

This particular row set 200 4_RandomNoise within 3_ConfCal provides the calibration probability with the maximum value 0.0685 and the 90th percentile value 0.0077. It means that the pooled calibration subset includes many very low-scoring negatives along with difficult cases, namely partial evidence, and lexical trap contexts. For my experiment, the 90th percentile threshold in this case is 0.4722 that is less than the traditional reference threshold of 0.500.

The threshold under consideration corresponds to the overall false-sense rate (FSR) of 8.67% on 4_FinalTest. Still, the FSR values depending on the context regimes are 15.5% on 2_Partial and 10.5% on 3_LexicalTrap. Thus, the results from the subgroups show that the averaged value in terms of pooling is not enough for describing results if the proportions in regimes are different. The lack of guarantees of query-level calibration makes it impossible to interpret the mentioned value as the guaranteed future upper error rate.

5.3 Hard-negative-restricted quantile threshold

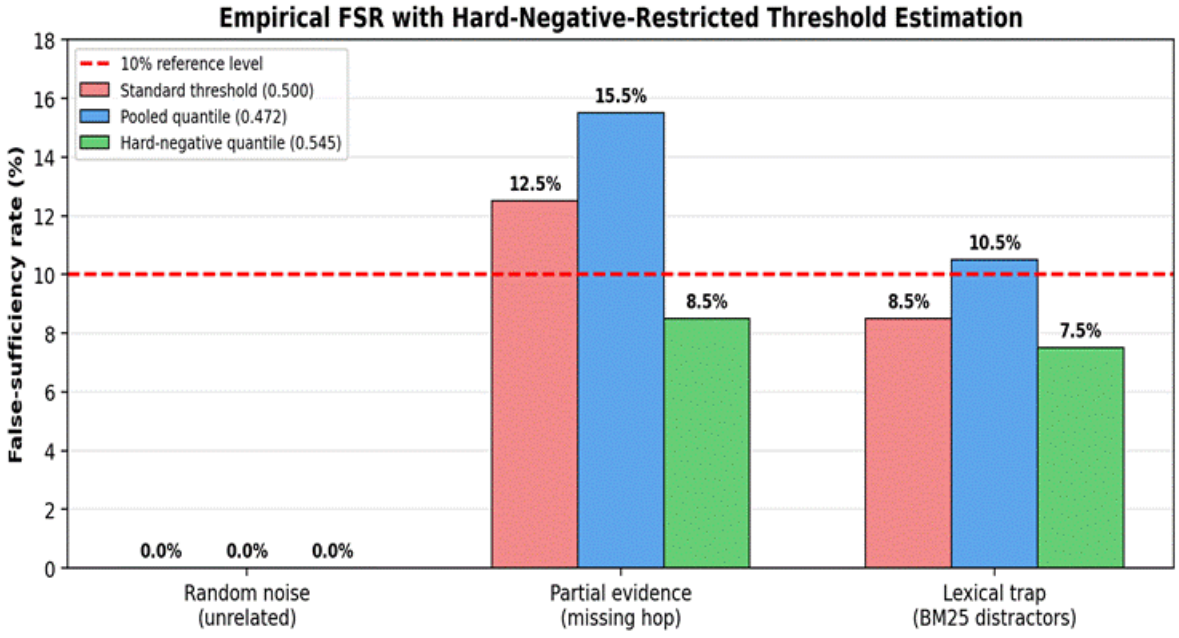


Fig. 4. Test FSR at the standard, pooled, and hard-negative-restricted thresholds. The hard negative-restricted threshold (τ = 0.5453) calculated based on calibration rows 2_Partial and 3_LexicalTrap gives me the test FSR of 8.5% and 7.5% correspondingly.

The limitation of the calibration on 400 contexts of 2_Partial and 3_LexicalTraps raises the threshold from 0.4722 to 0.5453. The FSR on 4_FinalTest equals 8.5% for 2_Partial and 7.5%

for 3_LexicalTrap as well as 5.33% in total for all 600 insufficient contexts. The Normal Sufficiency rate is 44.0% versus 50.5% of the threshold of 0.500. Thus, I may note the shift in the tradeoff to reject a greater number of insufficient contexts than keep a greater proportion of complete contexts. conclusion is specific to the constructed benchmark and fixed test partition, and no guarantee of finite-sample kind is given here.

6. Discussion, Limitations, and Conclusion

Based on the presented above results, the following suggestions are proposed for performing future evaluation of lightweight RAG abstinence gates:

1) Compute Normal Sufficiency together with aggregated and regime-specific false sub-acceptance rate (FSR) metrics. The ECE or even regime-specific accuracy will show higher than the actual false-acceptance rate due to the fact that inputs may lack partial evidence and have lexical traps;

2) Tune the cutoff value depending on the deployment distribution and observe the behavior of subgroup rates with changes in the distribution. The conducted experiments show that pooled quantile produces dissimilar results depending on the difficulty of the subset and this doesn't provide reasons to generally prohibit random negatives usage.

Consider not only the overlap feature but also structural paragraph minimum similarity and analyze its effect on different sets and retrievals settings.

Avenues for future work and limitations. The experimental settings are artificially constructed using a specific rule set based on HotpotQA and aren't extracted from a production retriever. Choosing four regimes could be too limited and won't cover all the challenges of the production-level retriever. The task uses Wikipedia-based dataset and therefore no performance on clinical, financial and other domains datasets is provided. Furthermore, the experimental pipeline doesn't include language model execution and, thus, the measurements don't include halluncination rates. The audit is self-made and the 96.5% accuracy value was computed by comparing rules with automatic NLI judge. The answer-string check isn't capable of capturing aliases and paraphrasing. Three negative contexts per question have been used in both splits calibration and testing. No query-level clustering as well as corresponding techniques of calibration and uncertainty estimation weren't applied to produce the reported quantiles and rates. The current report includes only the point estimates of metrics and no confidence intervals obtained via query-cluster boostrapping have been provided. Detailed description of the processes of features extraction, scoring, quantile interpolation and tiebreaking is needed in order to fully reproduce the corresponding part of the research. Latency metrics aren't available and the deployment speed can't be discussed.

Conclusion. This work compares dense cosine similarity, TF-IDF logistic regression, and the four-feature logistic-regression gate in terms of their ability to discriminate between contexts of 4,000 synthetic bridge questions built on the basis of 1,000 HotpotQA questions. As far as the held-out set is concerned, Platt scaling improves the cosine Expected Calibration Error (ECE) from 0.1619 down to 0.0267. Nevertheless, the selected operating point leads to greater failure-sensitive rates (FSR) for the partial-evidence and lexical-trap contexts. In case of the four-feature classifier, the achieved AUROC is 0.8583. FSRs in this scenario make up 12.5% and 8.5% correspondingly for each context regime. For the pooled FSRs of the negative class

quantile, the FSRs are larger (15.5% and 10.5%). FSRs for calibration rows, on the other hand, equal 8.5% and 7.5%; Normal Sufficiency values are smaller. All in all, aggregation-based calibration and error measures seem to mask model performances in the particular regimes.

References

[1] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[2] Z. Yang et al., “HotpotQA: A dataset for diverse, explainable multi-hop question answering,” in *Proceedings of EMNLP*, 2018, pp. 2369–2380. https://doi.org/10.18653/v1/D18-1259

[3] O. Yoran, T. Wolfson, O. Ram, and J. Berant, “Making retrieval-augmented language models robust to irrelevant context,” in *Proceedings of ICLR*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/8011b23e1dc3f57e1b6211ccad498919-Abstract-Conference.html

[4] H. Joren et al., “Sufficient context: A new lens on retrieval augmented generation systems,” in *Proceedings of ICLR*, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/33dffa2e3d2ab74a783d1a8c292f66d9-Abstract-Conference.html

[5] J. C. Platt, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” in *Advances in Large Margin Classifiers*, A. J. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA, USA: MIT Press, 1999, pp. 61–74.

[6] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 1321–1330. https://proceedings.mlr.press/v70/guo17a.html

[7] A. N. Angelopoulos, S. Bates, E. Fisch, L. Lei, and T. Schuster, “Conformal risk control,” in *Proceedings of ICLR*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html

[8] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. New York, NY, USA: Springer, 2005. https://doi.org/10.1007/b106715

[9] R. J. Tibshirani, R. F. Barber, E. Candes, and A. Ramdas, “Conformal prediction under covariate shift,” in *Advances in Neural Information Processing Systems*, vol. 32, 2019. https://proceedings.neurips.cc/paper_files/paper/2019/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html

[10] C. K. Chow, “On optimum recognition error and reject tradeoff,” *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 41–46, 1970. https://doi.org/10.1109/TIT.1970.1054406

[11] Y. Geifman and R. El-Yaniv, “Selective classification for deep neural networks,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017. https://papers.nips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html

[12] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in *Proceedings of EMNLP-IJCNLP*, 2019, pp. 3982–3992. https://doi.org/10.18653/v1/D19-1410

[13] S. Robertson and H. Zaragoza, “The probabilistic relevance framework: BM25 and beyond,” *Foundations and Trends in Information Retrieval*, vol. 4, nos. 1–2, pp. 1–174, 2009. https://doi.org/10.1561/1500000019