

Uncertainty-Aware Fact Checking with LLMs

Yulin Chen, Shubhalaxmi Mukherjee, Pradeep K. Murukannaiah

Delft University of Technology, Delft, The Netherlands

cyl03@outlook.com, s.mukherjee-2@tudelft.nl, p.k.murukannaiah@tudelft.nl

Abstract

Large Language Models (LLMs) offer a scalable approach to fact-checking, but their parametric knowledge remains unreliable for complex, specialized claims. Existing solutions unconditionally perform verification using Retrieval-Augmented Generation (RAG), which is computationally expensive and susceptible to performance-limiting noise. We propose an adaptive fact-checking pipeline that invokes RAG only when the LLM’s uncertainty surpasses a calibrated threshold. We introduce two UQ methods: (1) Adversarial Consistency Evaluation (ACE), a novel approach that detects model uncertainty by assessing its logical consistency, and (2) a trained ensemble that combines the strengths of several white-box UQ methods and ACE. When evaluated on the SciFact and QuanTemp benchmarks, our adaptive pipeline preserves at least 88% of the top-performing baseline’s accuracy. Simultaneously, our approach reduces retrieval frequency and token consumption by up to 55.3% and 37.1%, respectively, consistently outperforming decomposition-based RAG baselines at a fraction of the cost.

1 Introduction

Large Language Models (LLMs) are becoming increasingly popular for automated fact-checking because they offer a more scalable and cost-effective alternative to human verifiers. While LLMs enable us to fact-check a large number of claims quickly, they frequently suffer from hallucinations, generating explanations that appear plausible but are factually incorrect due to insufficient parametric knowledge (Bang et al., 2023; Dhuliawala et al., 2024; Maynez et al., 2020; McKenna et al., 2023). This limitation is especially problematic when responses concern high-stakes domains such as healthcare, science, and politics (Pal et al., 2023; Tian et al., 2024; DeVerna et al., 2024), where misinformation can lead to serious consequences. Verifying

claims in these fields often demands specialized knowledge and careful reasoning over strict medical guidelines, empirical research data, or precise political quotes and statistics. Consequently, despite their clear efficiency gains, LLMs’ limited domain knowledge and reliability remain major obstacles to their widespread deployment. To mitigate hallucinations, Retrieval-Augmented Generation (RAG) has become the standard solution, grounding model-generated responses in external knowledge (Guu et al., 2020; Izacard et al., 2023; Lewis et al., 2020; Sahoo et al., 2024).

Consider the claim, "The Netherlands is a member of the EU and has the highest GDP per capita amongst its members in 2025". It contains two distinct atomic assertions: One, the Netherlands is a member of the EU, and two, it has the highest GDP per capita among member states in 2025. While the first assertion is factually correct, the second is false, as the Netherlands placed only fourth in 2025 (Statistics Netherlands (CBS), 2026). If evaluated holistically as a single unit, an LLM might erroneously validate the entire sentence, as the true subclaim acts as a strong distracting signal that diverts the model from scrutinizing the false one. Fine-grained fact-checking addresses this by decomposing the claim into sub-claims and fact-checking each separately, allowing for a more precise and nuanced evaluation of overall factuality.

However, this fine-grained approach also introduces significant efficiency challenges (Hu et al., 2025; Tang et al., 2024; Xie et al., 2025). By indiscriminately decomposing and verifying every extracted sub-claim, standard pipelines incur excessive retrieval and inference costs, often proving computationally wasteful. As demonstrated by Hu et al. (2025), applying well-known decomposition methods like VeriScore (Song et al., 2024), and FactScore (Min et al., 2023) splits a single claim into an average of 1.96 and 4.02 sub-claims, respectively. Since the pipeline must execute both

document retrieval and LLM verification for each sub-claim, the total computational overhead is expected to increase by 96% - 302% compared to single-pass RAG. Furthermore, [Hu et al.](#) note that the excessive fragmentation of claims can actively harm fact-checking accuracy, which is driven by the build-up of cumulative retrieval noise from irrelevant or distracting evidence. It is also observed that for the task of political fact-checking, the performance of their RAG pipeline fell below that of zero-shot baselines as the ratio of noisy retrievals increased. Furthermore, [Zeng et al. \(2025\)](#) and [Cucunasu et al. \(2024\)](#) show that semantically related yet factually irrelevant documents are especially harmful, steering models away from the correct answer and degrading accuracy.

To optimize the trade-off between accuracy and efficiency, a fact-checking system must reliably recognize its own knowledge boundaries. Recent literature demonstrates that Uncertainty Quantification (UQ) techniques can effectively measure an LLM’s internal uncertainty, potentially allowing systems to identify fabricated claims and corresponding knowledge gaps ([Fadeeva et al., 2026](#); [Huang et al., 2024](#); [Kadavath et al., 2022](#); [Kang et al., 2025](#); [Kuhn et al., 2023](#)).

Building on this premise, we address the aforementioned efficiency challenges with an adaptive, automated fact-checking pipeline (Figure 1) that evaluates each claim holistically rather than decomposing it into atomic sub-claims. To quantify the model’s uncertainty in its initial verdict, the pipeline draws on a set of interchangeable uncertainty methods: information-theoretic estimators from LM-Polygraph ([Fadeeva et al., 2023](#)), a new logic-based method (ACE), and a logistic regression ensemble that combines these methods as features. Whenever the corresponding uncertainty score exceeds a pre-calibrated threshold, the system deems the model’s parametric verdict untrustworthy and re-verifies the claim using retrieved evidence. Unlike exhaustive frameworks (e.g., FActScore ([Min et al., 2023](#)), VeriScore ([Song et al., 2024](#))) that retrieve evidence indiscriminately for claim verification, our pipeline invokes RAG only for the claims the model is most uncertain about. In summary, we investigate the following broad research question:

RQ : Does bypassing RAG based on the uncertainty of an LLM’s parametric verdict adequately balance efficiency and accuracy?

The contributions of this work are as follows:

1. We introduce a fact-checking framework that improves efficiency by triggering RAG-based re-verification only when the LLM uncertainty is above a calibrated threshold.
2. We propose Adversarial Consistency Evaluation, a method that measures model uncertainty by evaluating the logical consistency of its verdicts between a claim and its’ negation.
3. We introduce a logistic regression ensemble that uses multiple UQ methods and ACE as input features that serves as a new UQ measure.

2 Related Work

Adaptive Retrieve-then-verify To improve fact-checking efficiency, recent work systematically assesses the need for external evidence before retrieval. These methods employ a combination of confidence scores, semantic consistency, token probabilities and measures of complexity of a claim as signals to bypass RAG in the verdict-generation process. **FIRE** ([Xie et al., 2025](#)) employs an agent-based framework that integrates retrieval and verification, dynamically deciding whether to output a final verdict or issue a subsequent search query based on the model’s confidence in its current judgment. **FaStFact** ([Wan et al., 2025](#)) extracts a zero-shot parametric verdict and uses its normalized token probability to bypass RAG if it exceeds a predefined threshold. Alternatively, **Rowen** ([Ding et al., 2025](#)) measures model confidence via semantic consistency across diverse prompts, while **Adaptive-RAG** ([Jeong et al., 2024](#)) trains a lightweight classifier to separate queries by complexity, allowing their RAG pipeline to skip retrieval entirely for easy queries while proceeding with single/multi-hop retrieval for complex ones. Despite these advances, existing adaptive frameworks typically rely on single measures of uncertainty and lack a systematic comparison to justify their choice relative to other UQ methods. This paper addresses this gap by independently evaluating multiple UQ methods to serve as triggers for RAG.

Uncertainty Quantification in LLMs Predictive uncertainty in LLMs, which reflects the variability of the model’s output for a given input, is typically attributed to two categories. **Aleatoric uncertainty**, which stems from inherent generation

randomness, and **epistemic uncertainty**, which arises from parametric knowledge gaps and can be mitigated with additional training data, in-context learning, and RAG (Kiureghian and Ditlevsen, 2009; Huang et al., 2024; Shorinwa et al., 2025; Kang et al., 2025; Liu et al., 2025). To measure model uncertainty, past work has introduced methods that can be broadly divided into black-box (sampling-based) and white-box (logit-based) categories (Fadeeva et al., 2023). In this work, we focus on **white-box, information-theoretic UQ methods** from *LM-Polygraph* (Fadeeva et al., 2023). These methods derive uncertainty scores directly from the model’s output token probabilities, making them more computationally efficient than black-box techniques that require repeated sampling or auxiliary NLI models at inference (Kuhn et al., 2023).

3 UQ-Aware RAG for Fact-checking

We define automated fact-checking as a classification task where, given an input claim c , the objective is to produce a verdict $\hat{v} \in \mathcal{V}$ on the claim’s veracity, where $\mathcal{V}$ represents a predefined set of labels.

Building on the structural foundations of FaStFact’s selective RAG fact-checking framework (Wan et al., 2025), we propose a pipeline (Figure 1) that adapts its verification strategy based on the model’s uncertainty in verifying each claim c . In the first step, the LLM generates an initial verdict $\hat{v}_{\text{param}}$ relying strictly on its parametric memory. While the original FaStFact architecture derives its confidence exclusively from the normalized log probability of the label token, our pipeline expands this evaluation by using one of the available UQ methods in Table 1 to quantify the model’s uncertainty of its parametric verdict (U_c).

In the next step, the system evaluates the uncertainty score U_c against a calibrated threshold τ^* to dynamically determine the strategy for obtaining the final verdict $\hat{v}_{\text{final}}$.

When $U_c \leq \tau^*$, we deem the model to be confident enough to verify the claim without external knowledge retrieval. The pipeline bypasses RAG entirely and returns $\hat{v}_{\text{param}} = \hat{v}_{\text{final}}$. Otherwise, we proceed with RAG. The evidence-grounded verdict $\hat{v}_{\text{rag}}$ is then used as the final prediction $\hat{v}_{\text{final}}$. Prompt details can be found under Appendix E.1.

Threshold Calibration After parametric evaluation and uncertainty estimation, the pipeline uses

an uncertainty threshold τ to determine whether claims require external evidence. Following FaStFact (Wan et al., 2025), claims with uncertainty scores below τ^* are accepted using parametric verdicts, while uncertain claims are re-verified through RAG. Since uncertainty scores vary across UQ methods, models, and claim distributions, fixed heuristic thresholds are unreliable. Therefore, the threshold is dynamically calibrated for each UQ method using a held-out set of claims.

To determine the optimal threshold for each UQ method, we optimize for **Youden’s Index** (Youden, 1950), a standard measure from diagnostic testing that quantifies a binary classifier’s ability to discriminate between two classes. Formally, for a given threshold $\tau \in \mathbb{R}$, Youden’s Index is defined as:

$$J(\tau) = \text{Sensitivity}(\tau) + \text{Specificity}(\tau) - 1 \quad (1)$$

In the context of our fact-checking pipeline, *sensitivity* (True Positive Rate) is the proportion of claims with incorrect parametric verdicts $\hat{v}_{\text{param}}$ that the threshold successfully flags for retrieval. In contrast, *specificity* (True Negative Rate) is the proportion of claims with correct verdicts $\hat{v}_{\text{param}}$ that successfully bypass retrieval-based re-verification. To calibrate the threshold for each UQ method, we instruct the LLM to verify each claim from a held-out calibration set and calculate the uncertainty score of its response. Each verdict $\hat{v}_{\text{param}}$ is then compared against its ground truth v^* , partitioning the calibration set into claims the model verified correctly on its own and those that failed. For any chosen threshold, this partition is split again into four groups that form a confusion matrix (Figure 4 in Appendix), from which sensitivity, specificity, and Youden’s index can be computed directly. To identify the optimal threshold, we treat each unique uncertainty score observed in the calibration set as a candidate threshold, denoted as $\hat{\tau}$. By sweeping over this set of candidates, we select the final threshold τ^* that maximizes $J(\tau)$.

Adversarial Consistency Evaluation (ACE) To achieve a more complete and diverse evaluation of the model’s true uncertainty, we propose introducing a complementary logical perspective, in addition to the information-theoretic UQ methods listed in Table 1. Prior research Fluri et al. (2024) demonstrate that the reliability of LLM predictions can be evaluated by testing their logical consistency when the ground truth is unknown. The principle

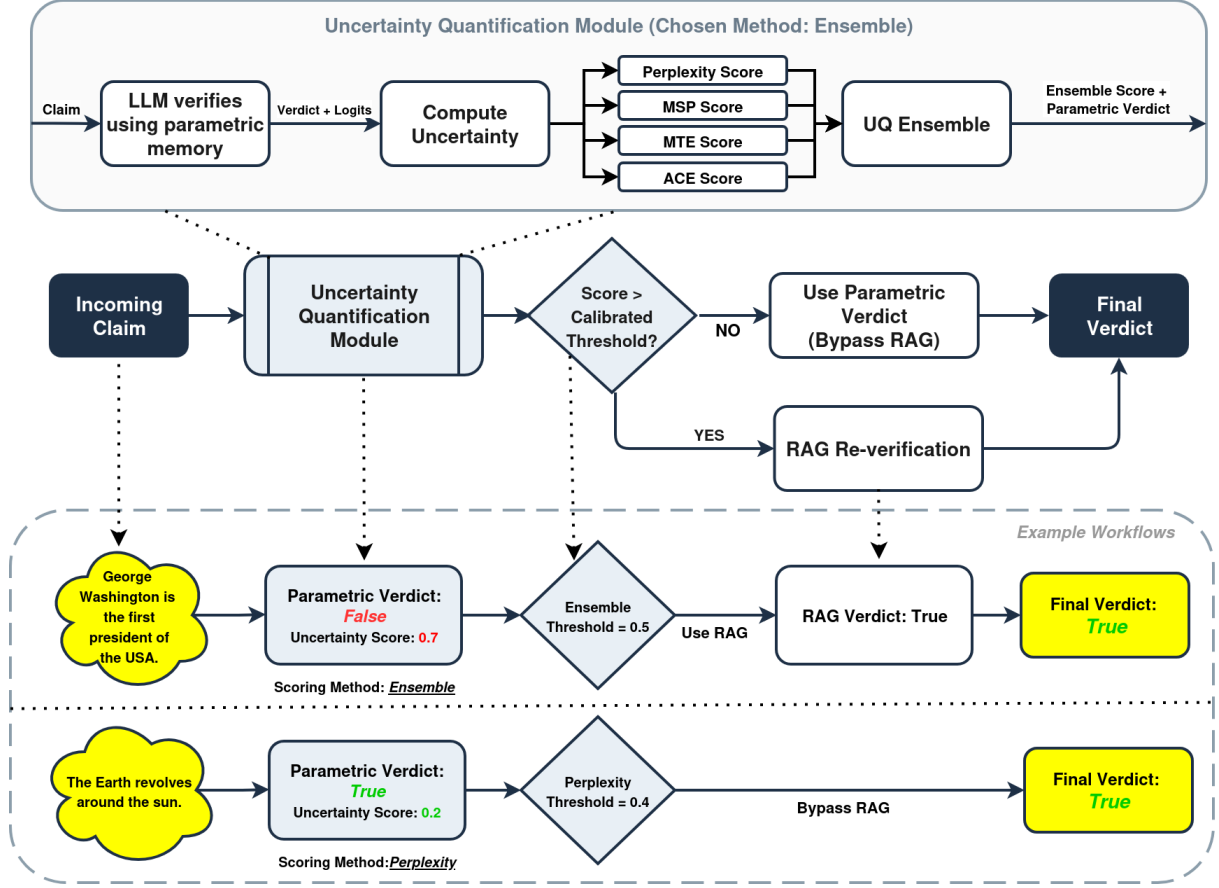


Figure 1: Brief overview of our Uncertainty Quantification (UQ)-guided fact-checking pipeline.

of negation consistency dictates that the probabilities of an event and its negation must be complementary. We adapt this concept to the domain of factual verification, operating on the principle of non-contradiction, which holds that a claim and its exact negation are mutually exclusive.

Given an input claim c , ACE prompts the LLM (Appendix E.2) to generate a negated adversarial claim $\neg c$ (Figure 2 in Appendix). The LLM then evaluates the pair of claims to produce verdicts $\hat{v}_{\text{param}}$ and $\hat{v}_{\text{adv}}$. For such a pair, Fluri et al. (2024) quantify the model’s violation of their consistency principle using a metric, defined as $\epsilon = |P(c) - (1 - P(\neg c))|$, where $P(c)$ represents the probability assigned by the model to the original claim being true. However, this approach requires us to compute the probability that each claim is assigned a "true" label (or equivalent) during evaluation. Alternatively, we propose the following discrete score:

Defining $\mathcal{V}_{\text{def}}$ as the subset of definitive, non-neutral verdicts (e.g., {True, False}), we formalize

the ACE uncertainty score U_{ACE} as:

$$U_{\text{ACE}} = \begin{cases} 1.0, & \text{if } \hat{v}_{\text{param}} = \hat{v}_{\text{adv}} \text{ and } \hat{v}_{\text{param}} \in \mathcal{V}_{\text{def}} \\ 0.0, & \text{otherwise} \end{cases} \quad (2)$$

It evaluates logical consistency at the verdict level without computing token probabilities. We assign a score of 1.0 if the model outputs identical, non-neutral labels for both the claim and its negation, and 0.0 otherwise. Thus, violating mutual exclusivity between a claim and its negation provides a sufficiently strong signal that the model’s judgement is not reliable.

Since U_{ACE} yields discrete rather than continuous scores, we do not calibrate an operating threshold or evaluate it individually in the end-to-end pipeline. Instead, it serves strictly as a feature for our UQ ensemble. We do, however, evaluate its standalone ability to detect model uncertainty (AU-ROC) on both the calibration and test sets.

UQ Ensemble Xiong et al. (2024) observe that many UQ methods exhibit weak correlations with one another, indicating that they capture comple-

Method	Type	Category	Training Data?
Maximum Sequence Probability	White-box	Information-based	No
Perplexity (Jelinek et al., 1977)			No
Mean Token Entropy (Fomicheva et al., 2020)			No
Adversarial Consistency Evaluation (ACE) (Ours)	Black-box	Logic consistency	No
Ensemble w/o ACE (Ours)	Hybrid	Ensembling	Yes
Ensemble w/ ACE (Ours)			

Table 1: Model uncertainty methods evaluated in our fact-checking pipeline. For more details about the White-box methods see Appendix A.

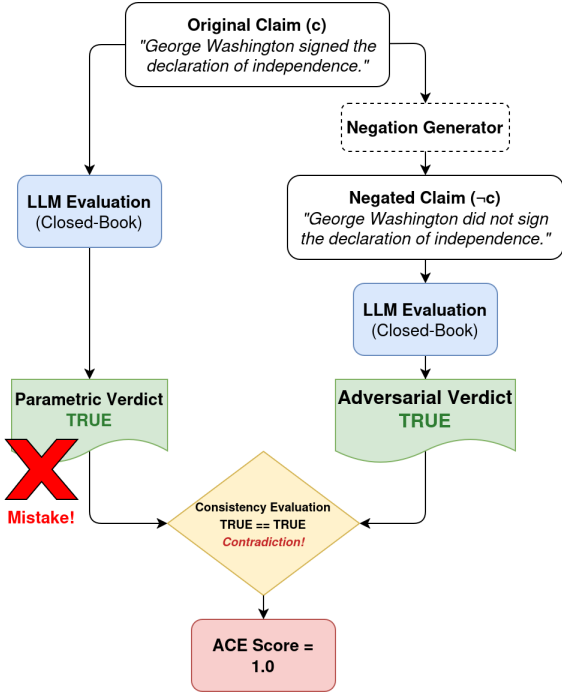


Figure 2: Example of the proposed Adversarial Consistency Evaluation (ACE) mechanism.

mentary aspects of model uncertainty and that multi-sample methods such as Semantic Entropy yield only marginal performance improvements over single-pass alternatives, despite incurring 5-10x higher inference cost. They also demonstrate that aggregating single-pass metrics into a lightweight regression model consistently matches or outperforms the best single and multi-sample methods across multiple QA benchmarks.

Driven by these insights, we propose a supervised ensemble of UQ methods. First, we instruct the LLM to produce a parametric verdict $\hat{v}_{\text{param}}$ for every claim and simultaneously compute a set of uncertainty scores, aggregating them into a feature vector $\mathbf{f} \in \mathbb{R}^K$, where each dimen-

sion corresponds to an independent UQ measure: $\mathbf{f} = [U_1, U_2, \dots, U_K]$.

We approach label assignment by framing the ensemble as a *model-error predictor*: rather than directly measuring uncertainty, the ensemble learns to predict the probability that $\hat{v}_{\text{param}}$ is incorrect. Thus, LLM prediction errors serve as a reliable proxy for the need for RAG-based re-verification. Each label y_{train} is therefore assigned by comparing the parametric verdict $\hat{v}_{\text{param}}$ against its ground truth v^* , consistent with the convention that higher scores indicate a greater likelihood of error:

$$y_{\text{train}} = \begin{cases} 1.0 & \text{if } \hat{v}_{\text{param}} \neq v^* \\ 0.0 & \text{if } \hat{v}_{\text{param}} = v^* \end{cases} \quad (3)$$

Next, we train a logistic regression model $g : \mathbb{R}^K \rightarrow [0, 1]$ on these feature-label pairs, mapping the feature vector $\mathbf{f}$ of UQ scores to the correctness of $\hat{v}_{\text{param}}$. Formally, the output of g given an input vector of uncertainty scores $\mathbf{f}$ is defined as:

$$g(\mathbf{f}) = \frac{1}{1 + \exp(-(\mathbf{w}^T \mathbf{f} + b))} \quad (4)$$

where $\mathbf{w}$ represents the learned weight vector and b is the bias term. This resulting probability serves as our ensemble uncertainty score, defined as $U_{\text{Ensemble}}(\mathbf{f}) = g(\mathbf{f})$.

As with all UQ methods evaluated in this work, a threshold τ_{Ensemble} is needed to establish the boundary between confident and uncertain LLM parametric verdicts. For this, we compute U_{Ensemble} scores over the same data split used for training the ensemble. To prevent data leakage, we compute the scores via 5-fold stratified cross-validation. These scores are subsequently used to calibrate an optimal threshold τ_{Ensemble}^* , as described in the following section.

3.1 RAG Re-verification

When a verdict’s uncertainty score meets or exceeds the calibrated threshold ($U(\hat{v}_{\text{param}}) \geq \tau^*$), the pipeline re-verifies the claim against externally retrieved evidence. We frame retrieval as a closed task, limiting the knowledge base to the evidence corpora provided within our datasets. For each knowledge base, corpus documents are first divided into smaller chunks with a small overlap between consecutive chunks. This overlap corresponds to approximately 2–3 sentences, ensuring that chunk boundaries do not discard relevant context from neighboring passages. To further preserve the broader document context, we follow the contextual retrieval approach of Anthropic (2024) by prepending the source document’s abstract to each chunk before indexing. Each enriched chunk is then encoded into dense and sparse representations and subsequently stored.

To extract the most relevant evidence, we follow Nogueira and Cho (2020) and implement a two-stage retrieve-and-rerank pipeline. In the first stage, the raw claim c is simultaneously encoded by both dense and sparse encoders, producing two ranked lists of initial candidates. The rankings are then combined using Reciprocal Rank Fusion (RRF) (Cormack et al., 2009), which calculates a unified score for each candidate chunk d across n retrieval methods: $RRF(d) = \sum_{i=1}^n \frac{1}{k + \text{rank}_i(d)}$ where $\text{rank}_i(d)$ is the chunk’s position in the i -th ranked list, and k is a smoothing constant. In the second stage, the claim and top candidates ranked by RRF are passed through a re-ranker, which scores each candidate chunk by directly comparing it against the claim to produce a final relevance score. The top-scoring chunks are then concatenated together as the grounding evidence D_c .

Following the RAG framework introduced by Lewis et al. (2020), we provide the verifier LLM with retrieved evidence D_c alongside original claim c . The model is then prompted (see full prompt in Appendix E.1) to re-evaluate the claim c and produce an evidence-grounded verdict $\hat{v}_{\text{rag}}$, which replaces $\hat{v}_{\text{param}}$ as our pipeline’s final verdict.

4 Experimental Setup

To evaluate our framework, we focus on three key aspects: the effectiveness of our UQ setup, fact-checking accuracy, and computational efficiency versus baselines. Below, we detail the configurations used across these evaluations.

Datasets We evaluate our pipeline on two challenging benchmark datasets: **SciFact** (Wadden et al., 2020) and **QuanTemp** (V et al., 2024). SciFact requires verifying scientific and biomedical claims (in English) against the research literature, a task that requires extensive domain expertise. QuanTemp targets numerical and temporal reasoning over real-world political and socio-economic claims. Both datasets are 3-class classification tasks. For more details about the splits used for threshold calibration and ensemble training see Appendix F.

Models We use Qwen3-4B-Instruct-2507, Mistral-7B-Instruct-v0.3, and Llama-3.1-8B-Instruct served on vLLM. All generations use a sampling temperature of 0.3, following the setting used by V et al. (2024) for claim verification on QuanTemp, and we report results over three seeded runs to account for run-to-run variability.

Retrieval Setup We employ a hybrid setup consisting of Qwen3-4B-Embedding (Zhang et al., 2025) and splade-cocondenser-ensembledistil (110M) (Formal et al., 2022), ranking chunks by semantic similarity and lexical overlap with the claim, respectively. The top 100 chunks from each ranking are fused via RRF (Equation 3.1), yielding a pool of 50 chunks that rank highly across both retrievers, which is subsequently reranked by Qwen3-4B-Reranker (Zhang et al., 2025). The top n chunks are then provided as grounding evidence, where $n = 1$ for SciFact and $n = 2$ for QuanTemp, since QuanTemp documents are substantially longer and require more retrieved chunks to adequately cover the relevant evidence (see Appendix F).

Baselines We compare our adaptive pipeline against three fact-checking approaches of increasing complexity. **Parametric Only** is a baseline in which the LLM verifies each claim using only its parametric memory, with no access to external evidence. **Always RAG** verifies using RAG for every claim unconditionally, which represents the performance ceiling our adaptive pipeline aims to match. **Fine-grained (Atomic) RAG** adopts the core concept of FActScore (Min et al., 2023) by prompting an LLM to break down claims into atomic sub-claims, retrieving and verifying for each. Unlike FActScore, which computes a precision score over atomic facts, our implementation aggregates sub-claim verdicts into one final verdict to align with

our classification-based evaluation setting.

Ensemble We train the UQ ensemble using scikit-learn’s (Pedregosa et al., 2011) *LogisticRegression* with `class_weights='balanced'` to counter class imbalance in the training data. As an ablation, we evaluate the ensemble both with (Denoted as *Ensemble+ACE*) and without ACE (Denoted as *Ensemble*) as a feature to isolate its contribution.

Evaluation Metrics To evaluate the discriminative power of UQ methods, we report **AUROC**. We report **Balanced Accuracy** (Brodersen et al., 2010), computed as the unweighted mean of per-class recall across all verdict classes, to prevent the majority class from dominating the evaluation score given the class imbalances inherent in our datasets (see Appendix D). To quantify the efficiency gains of our adaptive framework, we report the average **token usage** per claim and the **Flag Rate**, the percentage of claims re-verified using retrieved evidence.

5 Results and Discussion

In this section, we present the experimental results of our adaptive fact-checking framework. We evaluate our UQ-aware RAG framework across three key dimensions: (1) the predictive power of our UQ methods in distinguishing between correct & incorrect model parametric verdicts, (2) the end-to-end verification accuracy of our UQ-based pipeline compared to baselines, and (3) the computational efficiency gains achieved by selectively bypassing RAG-based re-verification.

5.1 Evaluating UQ Methods

Table 2 reports the individual performances of our UQ methods across all evaluated configurations. We discuss the key findings below.

Effectiveness of information-theoretic methods is task- and model-dependent On QuanTemp, information-theoretic methods perform consistently across all three models, with AUROC scores well above 0.5. On SciFact, however, their behavior is more model-dependent. For Qwen 4B, discriminative power collapses to well below random (AUROC 0.400–0.414), which is not observed for Mistral 7B and Llama 8B configurations. This inconsistent cross-dataset and cross-model behavior indicates that standalone information-theoretic

signals are heavily task- and model-dependent, limiting their generalizability in uncertainty-based fact-checking workflows.

ACE is effective only when claims have a clear logical opposite. On SciFact, ACE provides a decent uncertainty signal when paired with Qwen 4B and Mistral 7B, with AUROC above 0.6, outperforming information-theoretic methods, but it is worse for Llama 8B (0.510), which essentially resembles a random signal. On QuanTemp, however, ACE underperforms the random baseline across all models (0.406–0.475), a pattern that also holds during threshold calibration (see Appendix D). We hypothesize that this is because QuanTemp focuses on numerical and temporal claims, which are more challenging to negate into clear logical opposites. While many scientific claims feature clearly mutually exclusive relationships (e.g., ‘increases’ vs. ‘decreases’), numerical claims such as ‘decreased by 1.2%’ and ‘did not decrease by 1.2%’ are opposites but not mutually exclusive. Therefore, negating precise statistics yields ambiguous counterclaims that degrade ACE into a noisy, uninformative uncertainty signal.

The ensemble is consistent across domains and models. On QuanTemp, the UQ ensemble outperforms the top-performing standalone methods across all three models. On SciFact, it avoids the performance degradation observed with information-theoretic methods for Qwen 4B. Furthermore, adding ACE as a feature improves the ensemble on SciFact for both Qwen 4B and Mistral 7B, where the standalone ACE method also performed well. In contrast, ACE underperformed with Llama 8B, causing the ensemble’s performance to deteriorate after incorporating the method. On QuanTemp, where ACE is consistently worse than random, its inclusion has little effect. These results demonstrate that the ensemble remains robust across tasks and models, overcoming the dependence we observed with standalone UQ methods.

5.2 Fact-Checking Performance and Efficiency

Table 3 presents the end-to-end fact-checking performance and per-claim computational cost of our best approaches versus baselines. We present accuracy and efficiency jointly, as they together form an important trade-off. Full results of evaluated configurations are reported in Appendix C.

Category	Method	SciFact			QuanTemp		
		Qwen 4B	Mistral 7B	Llama 8B	Qwen 4B	Mistral 7B	Llama 8B
Information-Theoretic (LM-Polygraph)	Maximum Sequence Probability	0.414	0.575	0.616	0.647	0.621	0.640
	Perplexity	0.410	0.574	0.616	0.647	0.599	0.640
	Mean Token Entropy	0.400	0.592	0.631	0.647	0.572	0.635
Logical Consistency	ACE (Ours)	0.658	0.611	0.510	0.468	0.406	0.475
UQ Ensemble	Ensemble (Ours)	0.583	0.583	0.687	0.650	0.674	0.646
	Ensemble+ACE (Ours)	0.676	0.654	0.680	0.646	0.680	0.645

Table 2: AUROC for parametric verdicts across SciFact and QuanTemp evaluation sets. All standard deviations < 0.01.

(a) SciFact						
Model	Method	Acc	FR	Tok	Δ Acc	Δ Tok
Qwen 4B	Parametric Only	0.440	—	179	−42.0	−67.6
	Always RAG	0.760	—	553 $\pm$ 1	—	—
	Fine-grained RAG	0.739	—	1272 $\pm$ 1	−2.7	+130.0
	Ensemble Perplexity[†]	0.718$\pm$0.006	76.3$\pm$0.9	600$\pm$5	−5.5	+8.6
	MTE Perplexity[‡]	0.669	70.0	559	−11.9	+1.1
Mistral 7B	Parametric Only	0.438	—	253	−37.4	−64.1
	Always RAG	0.699	—	705 $\pm$ 1	—	—
	Fine-grained RAG	0.625	—	2130 $\pm$ 1	−10.7	+202.3
	MTE MTE[†]	0.672	58.0	672	−3.9	−4.6
	Ensemble MSP[‡]	0.616$\pm$0.033	46.0$\pm$7.5	584$\pm$57	−11.9	−17.2
Llama 8B	Parametric Only	0.297	—	205	−28.2	−64.0
	Always RAG	0.414 $\pm$ 0.002	—	569	—	—
	Fine-grained RAG	0.433 $\pm$ 0.001	—	1899 $\pm$ 1	+4.4	+233.7
	Perplexity Ensemble+ACE[†]	0.379$\pm$0.001	66.7	585	−8.5	+2.9
	Ensemble MSP[‡]	0.354$\pm$0.010	44.7$\pm$9.9	455$\pm$58	−14.5	−20.0
(b) QuanTemp						
Model	Method	Acc	FR	Tok	Δ Acc	Δ Tok
Qwen 4B	Parametric Only	0.368	—	175	−38.4	−86.2
	Always RAG	0.597 $\pm$ 0.002	—	1264	—	—
	Fine-grained RAG	0.513 $\pm$ 0.003	—	3652 $\pm$ 8	−14.1	+188.9
	Ensemble Perplexity[†]	0.514$\pm$0.002	49.5	811$\pm$1	−13.8	−35.9
	MTE Perplexity[‡]	0.514$\pm$0.002	49.5	811$\pm$1	−13.8	−35.9
Mistral 7B	Parametric Only	0.450	—	246	−34.4	−83.2
	Always RAG	0.685 $\pm$ 0.002	—	1468	—	—
	Fine-grained RAG	0.596 $\pm$ 0.006	—	5341 $\pm$ 47	−13.0	+263.9
	MTE MTE[†]	0.668$\pm$0.001	77.7$\pm$1.1	1394$\pm$16	−2.5	−5.0
	Ensemble MSP[‡]	0.649	65.4	1214	−5.3	−17.3
Llama 8B	Parametric Only	0.402	—	202	−33.9	−83.8
	Always RAG	0.607	—	1245	—	—
	Fine-grained RAG	0.547 $\pm$ 0.001	—	4512 $\pm$ 1	−10.0	+262.4
	Perplexity Ensemble+ACE[†]	0.554$\pm$0.004	51.6$\pm$1.0	1187$\pm$13	−8.8	−4.7
	Ensemble MSP[‡]	0.544	46.5	783	−10.4	−37.1

Table 3: Performance and cost comparison of baselines vs. our UQ-based pipeline (bold) on (a) **SciFact** and (b) **QuanTemp**, averaged over 3 runs (subscripts = non-zero std. dev.). For each model, we show the two best UQ configurations: [†] (Blue rows) = highest Balanced Accuracy, [‡] (Green rows) = best accuracy–efficiency trade-off (most tokens saved per % accuracy lost). **Acc**: Balanced Accuracy; **FR**: Flag Rate (%); **Tok**: avg. total token usage per sample; Δ **Acc**, Δ **Tok**: % change vs. Always RAG.

Across both benchmarks, the pipeline preserves accuracy while reducing retrieval frequency. The goal of an efficient fact-checking pipeline is to preserve the accuracy of *Always RAG* and *Fine-grained RAG* while minimizing retrieval frequency. On SciFact, Qwen’s accuracy-optimal

configuration (Ensemble) achieved 0.718 Balanced Accuracy, only 5.5% lower than the *Always RAG* baseline while skipping retrieval for 23.7% of claims, and Mistral’s accuracy-optimal configuration (Mean Token Entropy) lost only 3.9% compared to the baseline while skipping RAG for

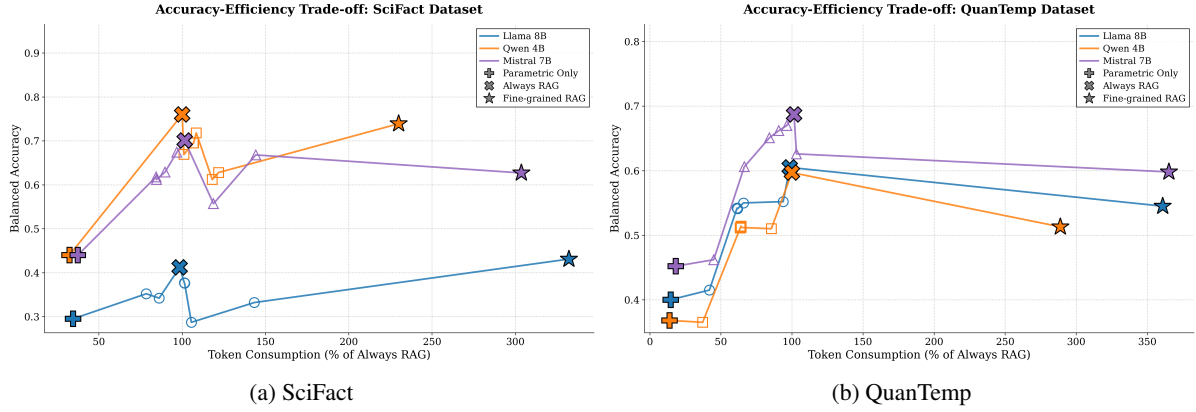


Figure 3: Accuracy-efficiency trade-off curves demonstrating the relationship between balanced accuracy and average tokens consumed per claim across the SciFact and QuanTemp datasets.

42.0% of claims. The same pattern holds for QuanTemp, where the accuracy-optimal configurations lose between 2.5% and 13.8% relative to their *Always RAG* baselines while skipping RAG for 22.3% to 50.5% of claims. These results indicate that the UQ methods can distinguish uncertain parametric verdicts well enough to maintain competitive accuracy without retrieving for every claim. Llama 8b on SciFact does not achieve the same performance. The *Parametric Only* baseline’s accuracy of 0.297 indicates that its parametric memory is mostly unreliable for these claims. *Always RAG* achieves only 0.414 accuracy. Notably, this is the only configuration where *Fine-grained RAG* surpasses *Always RAG*, by 4.4%.

Token savings depend on the length of the retrieved evidence. The token savings of our pipeline depend on the combined length of evidence retrieved to re-verify a flagged claim. On SciFact, because we retrieve only a single short chunk, the *Always RAG* baseline is inexpensive (553–705 tokens), leaving the overhead of parametric verification and UQ largely unrecovered. Larger savings on SciFact must instead be bought at the cost of accuracy, as the trade-off-optimal configurations reduce tokens by 17.2% to 20.0% while incurring an 11.9% to 14.5% drop in accuracy. This overhead is most pronounced for the ACE-based ensembles, which can increase token usage by up to 44.8% (see Appendix C), as the aggregate cost of generating a parametric verdict and running ACE’s extra verification on the negated claim exceeds the savings from skipping RAG re-verification.

On QuanTemp, the trend reverses. Because its claims require substantially more evidence to be verified (see Appendix F), the cost of RAG is far

higher (*Always RAG* uses 1,245–1,468 tokens per claim). The savings from skipping it easily absorb the overhead of parametric verification and uncertainty measurement. These results show that adaptively bypassing RAG re-verification can yield meaningful token savings, but only when verifying a claim requires longer retrieved evidence. Otherwise, the overhead of measuring uncertainty could easily negate those gains.

6 Conclusion

In this paper, we proposed an adaptive pipeline that leverages uncertainty quantification to mitigate the computational costs of RAG-based fact-checking. We propose ACE, a UQ method based on the logical consistency of verdicts for negated claims. ACE proves effective apart from claims that lack clear logical opposites. We also introduce an ensemble of ACE and white-box UQ methods that consistently outperforms its constituent methods and remains robust across models and datasets. Our pipeline preserves accuracy with top-performing baselines on SciFact and QuanTemp while substantially reducing retrieval frequency and, in many cases, overall token cost. A calibrated UQ threshold incurs a one-time computational cost, but it could deliver significant savings for a real-world fact-checking system. This work can be further extended to LLM responses that also include evidence-based explanations for the verdicts - quantifying long texts remains a difficult task, but the model’s reasoning becomes more auditable.

Limitations

Explainability & Compatibility To prevent uncertainty scores from being diluted by generation

noise, our pipeline restricts outputs to a single verdict label. This is based on the assumption that label token probabilities accurately encapsulate the model’s true uncertainty regarding complex claims. However, we sacrifice interpretability by omitting any model-generated explanations for its verdicts. Furthermore, the UQ methods (except ACE) presented in this thesis require access to the model’s internals, thereby rendering them inapplicable to black-box LLMs that do not expose token-level probabilities.

The Costs of Measuring Uncertainty Generating the parametric verdict and uncertainty scores, as well as running our ACE method, cumulatively incur token and other computational costs such as latency. While this can be easily offset by frequently bypassing RAG for complex datasets, computational cost savings diminish for tasks with inherently low retrieval costs (e.g., when the retrieved evidence is very short). Therefore, a primary limitation of our approach is that the upfront cost of measuring model uncertainty can outweigh the savings of bypassing RAG for fact-checking tasks that rely on short-context retrieval.

Scope and Evaluated Models Due to resource constraints, our evaluations were limited to models with 4–8B parameters. Furthermore, the scope of our findings is currently restricted to predominantly English-language claims, and the generalizability of these efficiency gains to multilingual settings remains largely unexplored. We hypothesize that larger models will exhibit stronger parametric verification capabilities across languages, thereby further reducing the frequency of RAG re-verification.

Generalizability of Thresholds and Ensemble The effectiveness of our UQ method thresholds and ensemble relies on in-domain calibration. Because optimal threshold values and logistic regression weights vary significantly across LLMs and target tasks, the pipeline may degrade in performance when exposed to out-of-distribution data. Consequently, applying our method to a new model or task requires explicit recalibration, which imposes a strict dependence on the availability of task-specific labeled data.

Sensitivity to Retrieval Noise When a claim is routed for external verification, the pipeline’s accuracy depends on the quality of the downstream retrieval and verification stages. In scenarios where irrelevant evidence is retrieved, the system risks

either failing to correct an initial parametric error or erroneously overriding a correct parametric verdict due to uninformative context (see Appendix G). In this work, we retrieve from a fixed corpus of curated evidence, whereas real-world fact-checking would be exposed to open, unverified sources, where retrieval is substantially noisier. Our reported accuracies should therefore be interpreted as an optimistic bound on what the pipeline would achieve under realistic retrieval conditions.

Compute Resources

Per dataset, we run 6 calibration and 27 evaluation jobs, with most completing within 2 hours, except *Fine-grained RAG* taking up to ~8 hours on QuanTemp. This totals to approximately 80–120 GPU-hours. Each job requires a single NVIDIA A40 GPU (48GB VRAM) and 32GB of RAM.

References

- Anthropic. 2024. Contextual retrieval. <https://www.anthropic.com/engineering/contextual-retrieval>. Accessed: 2026-05-10.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. *A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity*. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.
- Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. 2010. *The balanced accuracy and its posterior distribution*. In *2010 20th International Conference on Pattern Recognition*, pages 3121–3124.
- Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. *Reciprocal rank fusion outperforms condorcet and individual rank learning methods*. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’09*, page 758–759, New York, NY, USA. Association for Computing Machinery.
- Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. *The power of noise: Redefining retrieval for rag systems*. In *SIGIR 2024 - Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*,

- pages 719–729. Association for Computing Machinery, Inc.
- Matthew R. DeVerna, Harry Yaojun Yan, Kai-Cheng Yang, and Filippo Menczer. 2024. [Fact-checking information from large language models can decrease headline discernment](#). *Proceedings of the National Academy of Sciences*, 121(50):e2322823121.
- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. [Chain-of-verification reduces hallucination in large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.
- Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2025. [Rowen: Adaptive retrieval-augmented generation for hallucination mitigation in llms](#). In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP 2025*, page 12–21, New York, NY, USA. Association for Computing Machinery.
- Ekaterina Fadeeva, Aleksandr Rubashevskii, Dzhannis Piatrashyn, Roman Vashurin, Shehzaad Dhuliawala, Artem Shelmanov, Timothy Baldwin, Preslav Nakov, Mrinmaya Sachan, and Maxim Panov. 2026. [Faithfulness-aware uncertainty quantification for fact-checking the output of retrieval-augmented generation](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 6814–6836, San Diego, California, United States. Association for Computational Linguistics.
- Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. [LM-polygraph: Uncertainty estimation for language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461, Singapore. Association for Computational Linguistics.
- Lukas Fluri, Daniel Paleka, and Florian Tramèr. 2024. [Evaluating superhuman models with consistency checks](#). In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 194–232.
- Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. [Unsupervised quality estimation for neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 8:539–555.
- Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. [From distillation to hard negative sampling: Making sparse neural ir models more effective](#). In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’22*, page 2353–2359, New York, NY, USA. Association for Computing Machinery.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. [Retrieval augmented language model pre-training](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.
- Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, and 7 others. 2020. [Array programming with NumPy](#). *Nature*, 585(7825):357–362.
- Qisheng Hu, Quanyu Long, and Wenya Wang. 2025. [Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance?](#) In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6313–6336, Albuquerque, New Mexico. Association for Computational Linguistics.
- Hsiu-Yuan Huang, Yutong Yang, Zhaoxi Zhang, Sanwoo Lee, and Yunfang Wu. 2024. [A survey of uncertainty estimation in llms: Theory meets practice](#). *Preprint*, arXiv:2410.15326.
- Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. [Atlas: Few-shot learning with retrieval augmented language models](#). *Journal of Machine Learning Research*, 24:1–43.
- Frederick Jelinek, Robert L. Mercer, Lalit R. Bahl, and Janet M. Baker. 1977. [Perplexity—a measure of the difficulty of speech recognition tasks](#). *Journal of the Acoustical Society of America*, 62.
- Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. [Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7036–7050, Mexico City, Mexico. Association for Computational Linguistics.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and

- 17 others. 2022. [Language models \(mostly\) know what they know](#). *Preprint*, arXiv:2207.05221.
- Sungmin Kang, Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, and Amir Salman Avestimehr. 2025. [Uncertainty quantification for hallucination detection in large language models: Foundations, methodology, and future directions](#). *ArXiv*, abs/2510.12040.
- Armen Der Kiureghian and Ove Ditlevsen. 2009. [Aleatory or epistemic? does it matter?](#) *Structural Safety*, 31(2):105–112. Risk Acceptance and Risk Communication.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. [Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation](#). In *The Eleventh International Conference on Learning Representations*.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS ’20*, Red Hook, NY, USA. Curran Associates Inc.
- Xiaoou Liu, Tiejun Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. [Uncertainty quantification and confidence calibration in large language models: A survey](#). In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, KDD ’25*, page 6107–6117, New York, NY, USA. Association for Computing Machinery.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. [On faithfulness and factuality in abstractive summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.
- Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Hosseini, Mark Johnson, and Mark Steedman. 2023. [Sources of hallucination by large language models on inference tasks](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2758–2774, Singapore. Association for Computational Linguistics.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [FactScore: Fine-grained atomic evaluation of factual precision in long form text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.
- Rodrigo Nogueira and Kyunghyun Cho. 2020. [Passage re-ranking with bert](#). *Preprint*, arXiv:1901.04085.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. [Med-HALT: Medical domain hallucination test for large language models](#). In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334, Singapore. Association for Computational Linguistics.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. [A comprehensive survey of hallucination in large language, image, video and audio foundation models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724, Miami, Florida, USA. Association for Computational Linguistics.
- C. E. Shannon. 1948. [A mathematical theory of communication](#). *The Bell System Technical Journal*, 27(3):379–423.
- Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2025. [A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions](#). *ACM Comput. Surv.*, 58(3).
- Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. [VeriScore: Evaluating the factuality of verifiable claims in long-form text generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474, Miami, Florida, USA. Association for Computational Linguistics.
- Statistics Netherlands (CBS). 2026. [GDP per capita: Netherlands ranks fourth in EU in 2025](#).
- Liyan Tang, Philippe Laban, and Greg Durrett. 2024. [MiniCheck: Efficient fact-checking of LLMs on grounding documents](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.
- Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, Rezarta Islamaj, Aadit Kapoor, Xin Gao, and Zhiyong Lu. 2024. [Opportunities and challenges for chatgpt and large](#)

- language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493.
- Venkatesh V, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. [Quantemp: A real-world open-domain benchmark for fact-checking numerical claims](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 650–660, New York, NY, USA. Association for Computing Machinery.
- David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. [Fact or fiction: Verifying scientific claims](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.
- Yingjia Wan, Haochen Tan, Xiao Zhu, Xinyu Zhou, Zhiwei Li, Qingsong Lv, Changxuan Sun, Jiaqi Zeng, Yi Xu, Jianqiao Lu, Yinhong Liu, and Zhijiang Guo. 2025. [FaStFact: Faster, stronger long-form factuality evaluations in LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23814–23854, Suzhou, China. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Zhuohan Xie, Rui Xing, Yuxia Wang, Jiahui Geng, Hasan Iqbal, Dhruv Sahnan, Iryna Gurevych, and Preslav Nakov. 2025. [FIRE: Fact-checking with iterative retrieval and verification](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2901–2914, Albuquerque, New Mexico. Association for Computational Linguistics.
- Miao Xiong, Andrea Santilli, Michael Kirchhof, Adam Golinski, and Sinead Williamson. 2024. [Efficient and effective uncertainty quantification for LLMs](#). In *Neurips Safe Generative AI Workshop 2024*.
- William J Youden. 1950. Index for rating diagnostic tests. *Cancer*, 3(1):32–35.
- Linda Zeng, Rithwik Gupta, Divij Motwani, Yi Zhang, and Diji Yang. 2025. [Worse than zero-shot? a fact-checking dataset for evaluating the robustness of RAG against misleading retrievals](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *Preprint*, arXiv:2506.05176.

A Whitebox UQ methods

Maximum Sequence Probability

For an input prompt $\mathbf{x}$ and a generated output sequence $\mathbf{y} = (y_1, \dots, y_L)$, the *Maximum Sequence Probability* (MSP) measures the joint probability of generating the entire sequence. LM-Polygraph estimates this uncertainty by calculating the negated sum of log-likelihoods across all tokens:

$$U_{MSP}(\mathbf{y}) = - \sum_{l=1}^L \log P(y_l | y_{<l}, \mathbf{x}, \theta) \quad (5)$$

This negation ensures that lower generation probabilities translate to higher uncertainty scores.

Perplexity

Building upon MSP, *perplexity* normalizes the sequence probability for varying output lengths by averaging the negated log-likelihoods over length L and exponentiating the result:

$$U_{Perplexity}(\mathbf{y}) = \exp \left(- \frac{1}{L} \sum_{l=1}^L \log P(y_l | y_{<l}, \mathbf{x}, \theta) \right) \quad (6)$$

This length-normalization prevents longer generated responses from being disproportionately penalized, with higher scores indicating greater model uncertainty.

Mean Token Entropy

Unlike MSP and perplexity, which evaluate the likelihood of the final generated sequence, token entropy quantifies uncertainty across the model’s entire vocabulary distribution V at each generation step. For each candidate token v with conditional probability p_v , the entropy of the generated token at step l , $\mathcal{H}(y_l)$ is defined as (Shannon, 1948):

$$\mathcal{H}(y_l) = - \sum_{v \in V} p_v \log p_v \quad (7)$$

The *Mean Token Entropy* (MTE) is then computed by averaging these step-wise entropies across the sequence length L :

$$U_{MTE}(\mathbf{y}) = \frac{1}{L} \sum_{l=1}^L \mathcal{H}(y_l) \quad (8)$$

Higher U_{MTE} scores indicate greater model uncertainty during token selection.

B Adversarial Consistency Evaluation

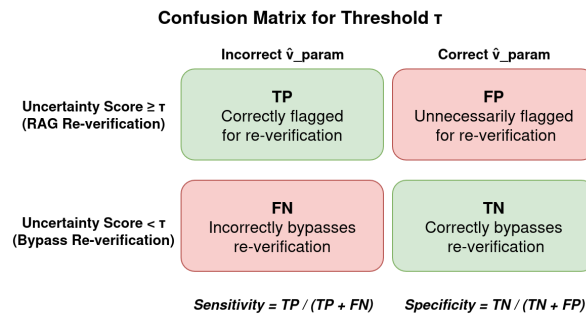


Figure 4: Confusion matrix for routing threshold τ

Method	SciFact					QuanTemp				
	Acc	FR (%)	Tok	ΔAcc (%)	ΔTok (%)	Acc	FR (%)	Tok	ΔAcc (%)	ΔTok (%)
Parametric Only	0.440	0.0	179	-42.0	-67.6	0.368	0.0	175	-38.4	-86.2
Always RAG	0.760	100.0	553 $\pm$ 1	-	-	0.597 $\pm$ 0.002	100.0	1264	-	-
Fine-grained RAG	0.739	0.0	1272 $\pm$ 1	-2.7	+130.0	0.513 $\pm$ 0.003	0.0	3652 $\pm$ 8	-14.1	+188.9
MaxSeqProb	0.695	76.3	591	-8.5	+6.8	0.512	49.6	813	-14.2	-35.7
MeanTokEnt	0.669	70.0	559	-11.9	+1.1	0.513	49.6	812	-14.0	-35.8
Perplexity	0.695	76.3	590	-8.5	+6.8	0.514$\pm$0.002	49.5	811$\pm$1	-13.8	-35.9
Ensemble	0.718$\pm$0.006	76.3$\pm$0.9	600$\pm$5	-5.5	+8.6	0.511 $\pm$ 0.002	49.0 $\pm$ 1.0	804 $\pm$ 12	-14.4	-36.4
Ensemble+ACE	0.628 $\pm$ 0.005	39.3 $\pm$ 1.1	674 $\pm$ 6	-17.4	+21.9	0.510 $\pm$ 0.006	47.6 $\pm$ 2.2	1081 $\pm$ 27	-14.6	-14.5

Table 4: End-to-End Efficiency results for the **Qwen 4B** model across datasets, averaged over 3 runs. **Acc** = Balanced Accuracy, **FR** = Flag Rate, **Tok** = Total Tokens, ΔAcc and ΔTok = relative % difference vs. Always RAG. **Blue** rows indicate the UQ method that achieves the highest Balanced Accuracy, **green** rows indicate the UQ method that achieves the best accuracy-efficiency trade-off, measured by $\Delta\text{Tok}/\Delta\text{Acc}$. Standard deviations are shown as subscripts where non-zero. Note that overhead associated with ACE is included in **Ensemble+ACE**.

Method	SciFact					QuanTemp				
	Acc	FR (%)	Tok	ΔAcc (%)	ΔTok (%)	Acc	FR (%)	Tok	ΔAcc (%)	ΔTok (%)
Parametric Only	0.438	0.0	253	-37.4	-64.1	0.450	0.0	246	-34.4	-83.2
Always RAG	0.699	100.0	705 $\pm$ 1	-	-	0.685 $\pm$ 0.002	100.0	1468	-	-
Fine-grained RAG	0.625	0.0	2130 $\pm$ 1	-10.7	+202.3	0.596 $\pm$ 0.006	0.0	5341 $\pm$ 47	-13.0	+263.9
MaxSeqProb	0.627 $\pm$ 0.041	51.0 $\pm$ 13.7	622 $\pm$ 99	-10.4	-11.7	0.649	65.4	1214	-5.3	-17.3
MeanTokEnt	0.672	58.0	672	-3.9	-4.6	0.668$\pm$0.001	77.7$\pm$1.1	1394$\pm$16	-2.5	-5.0
Perplexity	0.610 $\pm$ 0.017	46.7 $\pm$ 3.3	586 $\pm$ 25	-12.8	-16.8	0.660 $\pm$ 0.002	72.0 $\pm$ 0.9	1311 $\pm$ 13	-3.7	-10.7
Ensemble	0.616$\pm$0.033	46.0$\pm$7.5	584$\pm$57	-11.9	-17.2	0.604 $\pm$ 0.028	48.1 $\pm$ 9.4	954 $\pm$ 140	-11.8	-35.0
Ensemble+ACE	0.666 $\pm$ 0.009	52.4 $\pm$ 2.7	1006 $\pm$ 19	-4.8	+42.8	0.624 $\pm$ 0.005	57.8 $\pm$ 1.0	1493 $\pm$ 13	-9.0	+1.7

Table 5: End-to-End Efficiency results for the **Mistral 7B** model across datasets, averaged over 3 runs. **Acc** = Balanced Accuracy, **FR** = Flag Rate, **Tok** = Total Tokens, ΔAcc and ΔTok = relative % difference vs. Always RAG. **Blue** rows indicate the UQ method that achieves the highest Balanced Accuracy, **green** rows indicate the UQ method that achieves the best accuracy-efficiency trade-off, measured by $\Delta\text{Tok}/\Delta\text{Acc}$. Standard deviations are shown as subscripts where non-zero. Note that overhead associated with ACE is included in **Ensemble+ACE**.

Method	SciFact					QuanTemp				
	Acc	FR (%)	Tok	ΔAcc (%)	ΔTok (%)	Acc	FR (%)	Tok	ΔAcc (%)	ΔTok (%)
Parametric Only	0.297	0.0	205	-28.2	-64.0	0.402	0.0	202	-33.9	-83.8
Always RAG	0.414 $\pm$ 0.002	100.0	569	-	-	0.607	100.0	1245	-	-
Fine-grained RAG	0.433 $\pm$ 0.001	0.0	1899 $\pm$ 1	+4.4	+233.7	0.547 $\pm$ 0.001	0.0	4512 $\pm$ 1	-10.0	+262.4
MaxSeqProb	0.378 $\pm$ 0.001	66.7	586	-8.7	+3.0	0.544	46.5	783	-10.4	-37.1
MeanTokEnt	0.344 $\pm$ 0.001	51.3	499	-17.0	-12.4	0.552	50.9	839	-9.2	-32.6
Perplexity	0.379$\pm$0.001	66.7	585	-8.5	+2.9	0.544	46.5	783	-10.4	-37.1
Ensemble	0.354$\pm$0.010	44.7$\pm$9.9	455$\pm$58	-14.5	-20.0	0.543 $\pm$ 0.001	46.9 $\pm$ 0.2	788 $\pm$ 2	-10.7	-36.7
Ensemble+ACE	0.334 $\pm$ 0.008	50.6 $\pm$ 3.9	824 $\pm$ 21	-19.3	+44.8	0.554$\pm$0.004	51.6$\pm$1.0	1187$\pm$13	-8.8	-4.7

Table 6: End-to-End Efficiency results for the **Llama 8B** model across datasets, averaged over 3 runs. **Acc** = Balanced Accuracy, **FR** = Flag Rate, **Tok** = Total Tokens, ΔAcc and ΔTok = relative % difference vs. Always RAG. **Blue** rows indicate the UQ method that achieves the highest Balanced Accuracy, **green** rows indicate the UQ method that achieves the best accuracy-efficiency trade-off, measured by $\Delta\text{Tok}/\Delta\text{Acc}$. Standard deviations are shown as subscripts where non-zero. Note that overhead associated with ACE is included in **Ensemble+ACE**.

C Detailed Experiment Results

This section presents the comprehensive end-to-end efficiency and performance for all evaluated models (Qwen 4B, Mistral 7B, and Llama 8B) across both the SciFact and QuanTemp datasets. Tables 4, 5, and 6 detail for each configuration the balanced accuracy, UQ flag rate (% of claims re-verified by RAG), average per-claim token consumption, and the relative % difference compared to the *Always RAG* baseline. To facilitate direct comparison, we include our standard baselines alongside all evaluated UQ-based pipeline setups.

For each model and dataset pair, the UQ method that achieves the highest balanced accuracy is highlighted in blue, while the method that achieves the best accuracy-efficiency trade-off (determined by the ratio of relative token savings to accuracy loss ($\Delta\text{Tok}/\Delta\text{Acc}$)) is highlighted in green.

Fine-grained RAG offers the worst accuracy-efficiency trade-off. As Figure 3a and b illustrate, the decomposition-based *Fine-grained RAG* baseline is the most expensive strategy by a wide margin, inflating token consumption by 130.0% to 233.7% over *Always RAG* on SciFact and by 188.9% to 263.9% on QuanTemp. Yet, it surpasses *Always RAG* in accuracy only once (Llama on SciFact, as discussed above). On the other hand, our UQ-based configurations consistently came close to or outperformed *Fine-grained RAG* at a fraction of the cost. We attribute this drop in performance to the compounding noise introduced during claim decomposition and evidence retrieval, as described earlier in Section 1.

D Calibration Results

Table 7 presents the AUROC scores achieved by our uncertainty methods on the training/calibration sets. These calibration sets were strictly used to train our logistic regression ensemble and establish the optimal uncertainty score thresholds (Except ACE) for selective RAG re-verification.

Method	SciFact			QuanTemp		
	Qwen 4B	Mistral 7B	Llama 8B	Qwen 4B	Mistral 7B	Llama 8B
<i>Information-Theoretic (LM-Polygraph)</i>						
Maximum Sequence Probability	0.468	0.609 ± 0.003	0.609	0.687	0.649 ± 0.001	0.652
Perplexity	0.458	0.611 ± 0.003	0.609	0.686	0.605 ± 0.001	0.651
Mean Token Entropy	0.448	0.615 ± 0.002	0.669	0.683	0.576 ± 0.001	0.646
<i>Logical Consistency</i>						
Adversarial Consistency (ACE)	0.618	0.590 ± 0.002	0.512	0.470	0.401 ± 0.002	0.483
<i>UQ Ensemble</i>						
Ensemble w/o ACE	0.561 ± 0.006	0.617 ± 0.005	0.654 ± 0.003	0.657 ± 0.002	0.686 ± 0.001	0.654 ± 0.001
Ensemble w/ ACE	0.647 ± 0.003	0.629	0.664 ± 0.001	0.665 ± 0.004	0.688 ± 0.002	0.653

Table 7: AUROC scores measured on the validation (calibration) split, used for threshold optimization and ensemble training. We report the result averaged over three runs. The best-performing method for each model within a dataset is **bold**. Non-zero standard deviations are shown as subscripts.

E Prompts

E.1 Fact-checking Prompts

In our experiments, we use two prompt versions per dataset: **parametric**, which extracts parametric verdicts from the model, and **RAG**, which is used for evidence-grounded re-verification. The exact templates are detailed below.

Prompt: Parametric prompt (SciFact)

You are a biomedical domain expert and scientific fact-checker. Verify the veracity of the following claim using your internal scientific knowledge.

CRITICAL INSTRUCTIONS:

1. Evaluate the claim based on the most likely scientific consensus, making your best assessment even if you are not 100% certain.
2. Choose "SUPPORT" if the scientific evidence generally leans towards the claim being true.
3. Choose "CONTRADICT" if the scientific evidence generally leans towards the claim being false.
4. Choose "NOT ENOUGH INFO" ONLY IF the claim is fundamentally subjective, unscientific, or impossible to test. Do not use this label simply because you are unsure of the answer.

Claim: {claim}

Response:

Prompt: RAG prompt (SciFact)

You are a biomedical domain expert and scientific fact-checker. Verify the veracity of the following claim by cross-referencing it strictly against the provided evidence.

CRITICAL INSTRUCTIONS:

1. Choose "SUPPORT" ONLY IF the evidence explicitly contains scientific findings or data that confirm the claim.
2. Choose "CONTRADICT" ONLY IF the evidence explicitly contains findings or data that prove the claim is scientifically incorrect.
3. Choose "NOT ENOUGH INFO" IF the evidence is unrelated, tangential, or insufficient to make a scientific conclusion. Base your decision entirely on the provided text.
4. Format: Output EXACTLY ONE LABEL. Do not output JSON. Do not output an explanation. Do not use punctuation.

Evidence: {context}

Claim: {claim}

Response:

Prompt: Parametric prompt (QuanTemp)

You are an expert political and economic fact-checker. Verify the veracity of the following claim using your internal knowledge only.

CRITICAL INSTRUCTIONS:

1. Evaluate the claim based on the most likely historical or economic data, making your best assessment.
2. Choose "True" if standard consensus leans towards the numbers or dates being correct.
3. Choose "False" if the claim likely contains incorrect numbers, dates, or factual errors.
4. Choose "Conflicting" ONLY IF you specifically know of credible, contradictory datasets or reports that make a single truth impossible. Do not use this label simply because you cannot remember the exact numbers.

Claim: {claim}

Response:

Prompt: RAG prompt (QuanTemp)

You are an expert political and economic fact-checker. Verify the veracity of the following claim by cross-referencing it strictly against the provided evidence.

CRITICAL INSTRUCTIONS:

1. Choose "True" if the evidence contains findings, numbers, or data that confirm the claim.
2. Choose "False" if the evidence contains findings or data that prove the claim is incorrect.
3. Choose "Conflicting" if the provided evidence contains multiple pieces of information or statistics that conflict with each other regarding the claim.

Evidence: {context}

Claim: {claim}

Response:

E.2 ACE Negation Prompt

Prompt: ACE Negation

"Negate the core factual premise of the claim by inverting its logical meaning while preserving the original grammatical structure. Output EXACTLY ONE SENTENCE. No explanations, no quotes, no conversational filler.

Claim: {claim}

Response:

E.3 Decomposition Prompt

The following prompt is used during claim decomposition for the *Fine-grained Verification* baseline.

Prompt: Claim decomposition

You are a strict, highly precise text processing system. Your only task is to break down complex claims into a list of simple, independent atomic facts.

CRITICAL RULES:

1. Break the claim into self-contained, simple statements.
2. Split compound subjects (e.g., "A and B do X") or lists into separate independent claims.
3. Resolve all pronouns (e.g., "it", "they") to their specific entities.
4. If the claim is already a single, simple fact, just output that one fact.
5. NEVER output conversational filler (e.g., "Sure", "Here is the list", "I'm ready").
6. You MUST format every single fact as a bullet point starting exactly with "- ".
7. Do not generate new examples. Stop generating immediately after processing the provided claim.

EXAMPLE 1:

Claim: "Barack Obama, the 44th US President, was born in Hawaii and later moved to Indonesia."

Output:

- Barack Obama was the 44th US President
- Barack Obama was born in Hawaii.
- Barack Obama moved to Indonesia.

EXAMPLE 2:

Claim: "A 40mg dosage of Drug X and a 2mg dosage of Drug Y do not affect progression of disease Z."

Output:

- A 40mg dosage of Drug X does not affect progression of disease Z.
- A 2mg dosage of Drug Y does not affect progression of disease Z.

EXAMPLE 3:

Claim: "Drug A targets the cardiac muscle."

Output:

- Drug A targets the cardiac muscle.

NOW PROCESS THIS CLAIM:

Claim: {claim}

Output:

F Dataset Statistics

For threshold calibration and ensemble training, we use the train split for SciFact and the val split for QuanTemp. As for evaluations, we use the test split for QuanTemp and the val split for SciFact. Additionally, for QuanTemp, we combine the full oracle evidence provided under each claim across all splits into a unified corpus. Both datasets assign each claim one of three veracity labels, which we present alongside detailed dataset statistics below.

Dataset	Split	Claims	Label	Count (Proportion)
SciFact	Train (<i>Training & Calibration</i>)	809	SUPPORT	332 (41.0%)
			NOT ENOUGH INFO	304 (37.6%)
			CONTRADICT	173 (21.4%)
	Val (<i>Evaluation</i>)	300	SUPPORT	124 (41.3%)
			NOT ENOUGH INFO	112 (37.3%)
			CONTRADICT	64 (21.3%)
	Test (<i>Excluded</i>)	300	SUPPORT	N/A
			NOT ENOUGH INFO	N/A
			CONTRADICT	N/A
QuanTemp	Train (<i>Excluded</i>)	9,935	False	5,770 (58.1%)
			Conflicting	2,341 (23.6%)
			True	1,824 (18.4%)
	Val (<i>Training & Calibration</i>)	3,084	False	1,795 (58.2%)
			Conflicting	672 (21.8%)
			True	617 (20.0%)
	Test (<i>Evaluation</i>)	2,495	False	1,423 (57.0%)
			Conflicting	598 (24.0%)
			True	474 (19.0%)

Dataset	Corpus Chunks	Avg. Chunks / Doc	Avg. Relevant Docs / Claim	Language
SciFact	5,218	1.0068	0.6967	English
QuanTemp	28,620	1.9032	1.0000	English, German

Table 8: Dataset and corpus statistics

G Examples

Table 9: Example workflows of our uncertainty-guided pipeline under different scenarios. The only alternative that leaves a flagged parametric verdict unchanged after re-verification is omitted from the figures for brevity.

Scenario 1: Incorrect Parametric Verdict → Flagged → Corrected	
[SciFact Qwen 4B] A high microerythrocyte count raises vulnerability to severe anemia in homozygous alpha (+)-thalassemia trait subjects.	
Gold Label: CONTRADICT	Initial Prediction: SUPPORT (<i>Incorrect</i>)
Ensemble Uncertainty Score: <u>0.6773</u> (> 0.4657 Threshold → Flagged)	
Final Prediction: CONTRADICT (<i>Correct</i>)	

[QuanTemp Qwen 4B] In September 2021, the U.K. government announced its intention to create a new criminal offense of pet abduction.	
Gold Label: TRUE	Initial Prediction: FALSE (<i>Incorrect</i>)
Ensemble Uncertainty Score: <u>0.5482</u> (> 0.5200 Threshold → Flagged)	
Final Prediction: TRUE (<i>Correct</i>)	
Scenario 2: Correct Parametric Verdict → Skipped → Correct Verdict Retained	
[SciFact Qwen 4B] 1,000 genomes project enables mapping of genetic sequence variation, including rare variants with larger penetrance than common variants.	
Gold Label: SUPPORT	Initial Prediction: SUPPORT (<i>Correct</i>)
Ensemble Uncertainty Score: <u>0.3863</u> (≤ 0.4657 Threshold → Skipped)	
Final Prediction: SUPPORT (<i>Correct</i>)	

[QuanTemp Qwen 4B] [In 2014-2015] coverage for the rotavirus vaccine exceeded the 95% target and the pneumococcal vaccine reached 91.5%.	
Gold Label: FALSE	Initial Prediction: FALSE (<i>Correct</i>)
Ensemble Uncertainty Score: <u>0.3842</u> (≤ 0.5200 Threshold → Skipped)	
Final Prediction: FALSE (<i>Correct</i>)	
Scenario 3: Incorrect Parametric Verdict → Skipped → Uncorrected	
[SciFact Qwen 4B] Bone marrow cells contribute to adult macrophage compartments.	
Gold Label: CONTRADICT	Initial Prediction: SUPPORT (<i>Incorrect</i>)
Ensemble Uncertainty Score: <u>0.3602</u> (≤ 0.4657 Threshold → Skipped)	
Final Prediction: SUPPORT (<i>Incorrect</i>)	

[QuanTemp Qwen 4B] When you throw 23 million people off of health insurance – people with cancer, people with heart disease, people with diabetes – thousands of people will die. ... This is study after study making this point.	
Gold Label: TRUE	Initial Prediction: FALSE (<i>Incorrect</i>)
Ensemble Uncertainty Score: <u>0.4214</u> (≤ 0.5200 Threshold → Skipped)	
Final Prediction: FALSE (<i>Incorrect</i>)	
Scenario 4: Correct Parametric Verdict → Flagged → Verdict Becomes Incorrect	
[SciFact Qwen 4B] CX3CR1 on the Th2 cells impairs T cell survival	
Gold Label: CONTRADICT	Initial Prediction: CONTRADICT (<i>Correct</i>)
Ensemble Uncertainty Score: <u>0.6826</u> (> 0.4657 Threshold → Flagged)	
Final Prediction: SUPPORT (<i>Incorrect</i>)	

[QuanTemp Qwen 4B] Japanese citizens boycotting goods made in USA over nuclear bombing in 1945	
Gold Label: FALSE	Initial Prediction: FALSE (<i>Correct</i>)
Ensemble Uncertainty Score: <u>0.5482</u> (> 0.5200 Threshold → Flagged)	
Final Prediction: TRUE (<i>Incorrect</i>)	

H Retrieval Configuration Experiments

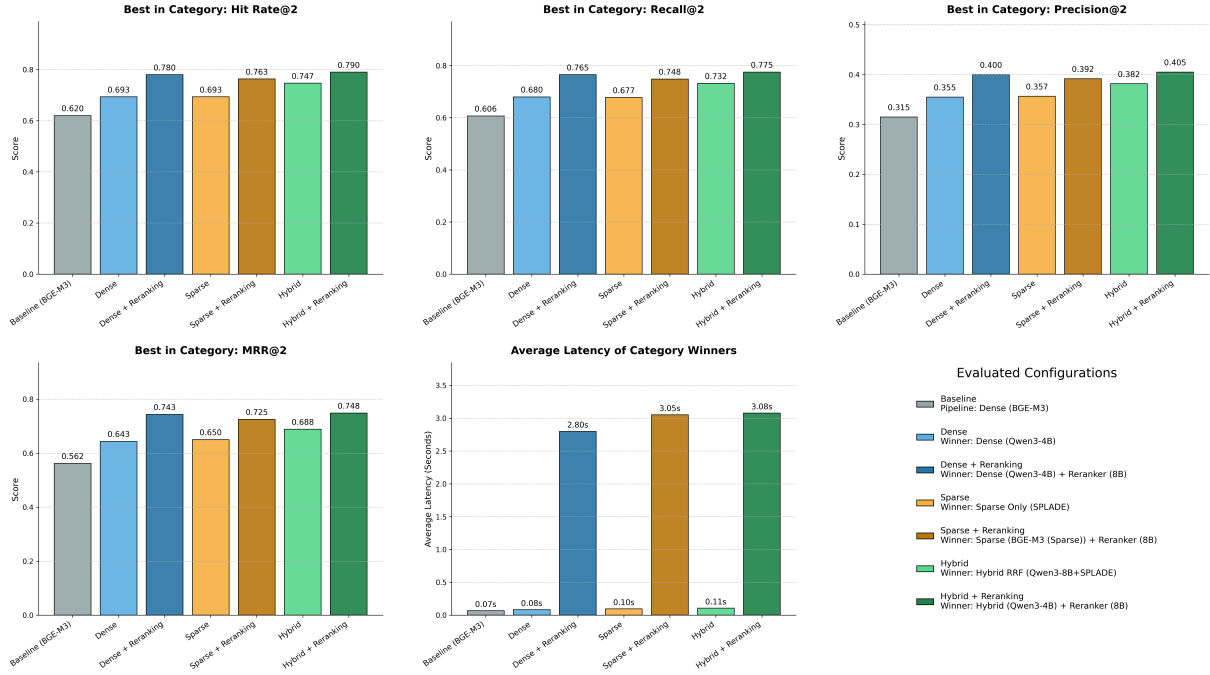


Figure 5: Top-2 retrieval performance of retrieval configurations on SciFact Val split.

We evaluated 41 retrieval configurations on the SciFact Val set (300 claims) to optimize our evidence retrieval setup. Specifically, the claims were used as queries to retrieve the top two chunks from the SciFact corpus, and performance was evaluated using the relevant document IDs included in the datasets as ground truth. The tested configurations included dense, sparse, and hybrid retrieval. We also tested those configurations with reranking added as an additional step. As shown in Figure 5, configurations using a reranker consistently outperformed standalone retrievers across all metrics. Although 8B models provide a marginal edge in retrieval accuracy, 4B variants require about 40% less runtime per claim. Therefore, for this thesis, we selected a hybrid retrieval + reranking setup consisting of Qwen3-Embedding-4B (dense), splade-cocondenser-ensembledistil (sparse), and Qwen3-Reranker-4B.

I Responsible NLP Research checklist

A1. Limitations A detailed analysis of this work’s limitations is outlined on page 13.

A2. Potential Risks The primary risk of our approach is that the selective bypassing of retrieval, combined with the imperfect discriminative power of our UQ methods, can cause incorrect parametric verdicts to be accepted without evidence-grounded re-verification (Appendix G, Scenario 3). Such failures are particularly problematic in high-stakes domains such as healthcare, science, and politics, where unquestioningly trusting a verdict without evidence-grounding could contribute to the propagation of harmful misinformation.

B1. Used Artifacts All datasets, LLMs, and retrieval models are cited in Section 4. Meanwhile, the LM-Polygraph UQ framework is cited under section 2.

B2. Artifact Licenses The licenses for the assets used in this work are as follows:

- Datasets: SciFact (CC BY-NC 2.0), QuanTemp (CC BY-NC 4.0)
- LLMs: Mistral-7B-Instruct-v0.3 & Qwen3-4B-Instruct-2507 (Apache-2.0), Llama-3.1-8B-Instruct (Llama 3.1 Community License)

- Retrieval Models: Qwen3-4B-Embedding & Qwen3-4B-Reranker (Apache-2.0), splade-cocondenser-ensembledistil (CC-BY-NC-SA-4.0)
- Frameworks: LM-Polygraph (MIT)

All assets are used solely for non-commercial research. The SciFact and QuanTemp datasets, including our derived QuanTemp corpus, are redistributed within our code repository under their original CC BY-NC licenses. We release our own code under the MIT license.

B3. Intended Use All datasets and models are used strictly for their intended non-commercial research purposes. Our released code and derived datasets are not intended for usage outside research purposes.

B4. Data Privacy We performed no additional anonymization or offensive-content filtering, instead relying on the curation of the established benchmark datasets (SciFact, QuanTemp). No private personal data is used, and while names of public figures and potentially charged content may appear, these are publicly available and inherent to the fact-checking task, so they are retained as-is.

B5. Artifact Documentation The dataset languages and relevant domains are documented in Section 4.

B6. Data Statistics Detailed dataset statistics are presented in Appendix F.

C1. Compute Model details, computational budget, and hardware setup are reported in Section 4.

C2. Experimental Setup The experimental setup is described in Section 4. The only tuned parameters in this work are the thresholds for each UQ method, calibrated on the train/validation splits and never on the held-out evaluation set.

C3. Results Reporting All experiments are repeated three times, and we report the mean and standard deviation for each result. For performance evaluations of our UQ methods, we report their AUROC scores on both the training/calibration and evaluation sets (Tables 7 & 2)

C4. Package Usage We use LM-Polygraph’s (Fadееva et al., 2023) (v0.5.0) implementations of Maximum Sequence Probability, Perplexity, and Mean Token Entropy. Note that their Perplexity estimator returns the mean negative log-likelihood, which we exponentiate to obtain the conventional perplexity score. Ensemble training, threshold calibration, and evaluation metrics (balanced accuracy, AUROC) use NumPy (Harris et al., 2020) (v1.26.4) and scikit-learn (Pedregosa et al., 2011) (v1.8.0). Models are loaded with Hugging Face Transformers (Wolf et al., 2020) (v4.57.6) and served with vLLM (Kwon et al., 2023) (v0.17.1).

D. Human Annotators & Research with Human Participants This section (D1–D5) is not applicable, as our work uses only existing benchmark datasets and does not involve human annotators or research with human participants.

E1. AI Assistance AI assistants were used in a limited, tool-assistive capacity, with the author retaining full responsibility for all content. Specifically: (1) For writing, to polish drafts written by the author by improving flow and correcting spelling and grammar, with all final proofreading and editing done by the author. (2) For coding, to implement components more efficiently, given author-specified instructions, with all design decisions made by the author and all resulting code independently reviewed and tested by the author. (3) During the literature review, to explain unfamiliar concepts (e.g., the mathematics behind uncertainty quantification methods and Youden’s Index).