Faithfulness-Aware Uncertainty Quantification for Fact-Checking the Output of Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) enhanced with retrieval, an approach known as Retrieval-Augmented Generation (RAG), have achieved strong performance in open-domain question answering. However, RAG remains prone to hallucinations: factually incorrect outputs may arise from inaccuracies in the model's internal knowledge and the retrieved context. Existing approaches to mitigating hallucinations often conflate factuality with faithfulness to the retrieved evidence, incorrectly labeling factually correct statements as hallucinations if they are not explicitly supported by the retrieval. In this paper, we introduce FRANQ, a new method for hallucination detection in RAG outputs. FRANQ applies distinct uncertainty quantification (UQ) techniques to estimate factuality, conditioning on whether a statement is faithful to the retrieved context. To evaluate FRANQ and competing UQ methods, we construct a new long-form question answering dataset annotated for both factuality and faithfulness, combining automated labeling with manual validation of challenging cases. Extensive experiments across multiple datasets, tasks, and LLMs show that FRANQ achieves more accurate detection of factual errors in RAG-generated responses compared to existing approaches. Our implementation is available at http://anonymous.for.review.

1 Introduction

Large Language Models (LLMs) are increasingly employed across a wide range of tasks. However, LLMs are prone to generating plausible but factually incorrect generations, a phenomenon known as hallucination, arising from factors such as insufficient training data coverage, input ambiguity, and architectural constraints (Huang et al., 2025). Retrieval-Augmented Generation (RAG; Lewis et al., 2020) mitigates this issue by incorporating dynamically retrieved external knowledge into the generation process, which can partially mitigate factual inaccuracies (Shuster et al., 2021).

However, RAG systems still produce hallucinations (Shi et al., 2023). Moreover, the use of retrieved information makes it more challenging to detect hallucinations and to determine their original source. Models become more confident in generating statements that appear in the retrieval, regardless of their factual correctness (Kim et al., 2025). At the same time, the retrieved passages themselves may be erroneous, incomplete, or completely irrelevant with respect to the query (Shi et al., 2023; Ding et al., 2024). Conversely, even when retrieval is accurate, inconsistencies can emerge between the model's internal knowledge and the retrieved data (Wang et al., 2024a, 2025).

Thus, an important question is how to define *hallucination* in RAG, given the interplay between the model's internal knowledge and the retrieved context. One approach is to consider any content that is not directly supported by the retrieved context as a hallucination (Niu et al., 2024). However, we argue that hallucination should be defined based on factual inaccuracies rather than strict contextual alignment. Specifically, a generated statement that originates from the LLM's internal knowledge but lies outside the retrieved context should not be considered a hallucination if it is factually correct.

To address this distinction, we differentiate between *factuality* and *faithfulness*. Faithfulness refers to whether the generated output is semantically entailed by the retrieved context, while factuality indicates whether the content is objectively correct (Maynez et al., 2020; Dziri et al., 2022; Yang et al., 2024). For RAG fact-checking, detecting non-factual claims is more critical than identifying unfaithful ones. This distinction disentangles two core RAG failure modes: (*i*) hallucinations caused by erroneous grounding in the retrieved context, and (*ii*) factual errors stemming from the model's internal knowledge (Zhou et al., 2024).

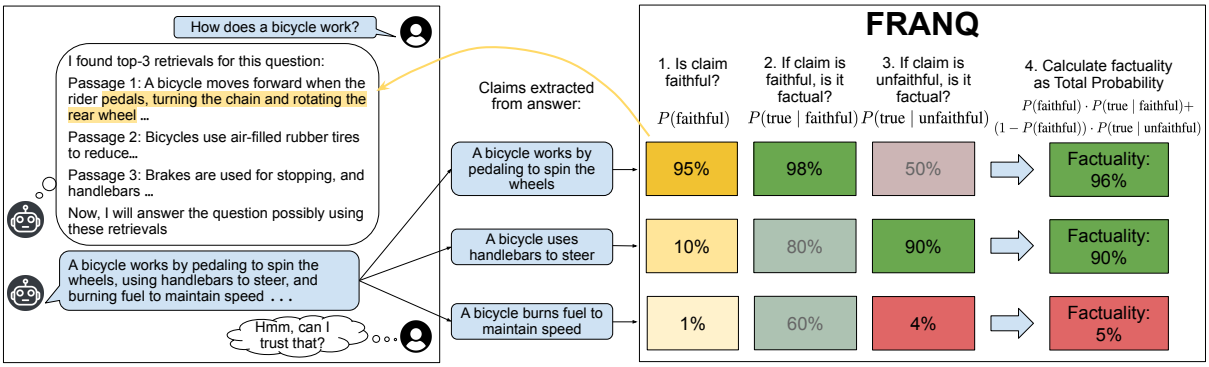


Figure 1: FRANQ illustration. *Left*: A user poses a question, and the RAG retrieves relevant documents and formulates an answer, potentially using information from the retrieved documents. *Middle*: The RAG output is decomposed into atomic claims. *Right*: The FRANQ method assesses factuality by evaluating three components: (1) faithfulness, (2) factuality under faithful condition, and (3) factuality under unfaithful condition.

In this paper, we investigate the detection of non-factual statements produced by RAG using Uncertainty Quantification (UQ) techniques. We introduce FRANQ (**F**aithfulness-aware **R**etrieval **A**ugmented UN**c**ertainty **Q**uantification), a novel method that first evaluates the faithfulness of the generated response and subsequently applies different UQ methods based on the outcome. With this separation, FRANQ tailors its strategy to the specific RAG failure mode: whether it originates from retrieval grounding or from the model's own knowledge.

We evaluate FRANQ on both long- and short-form question answering (QA) tasks. For long-form QA, where answers include multiple claims, we assess factuality at the claim level and introduce a new dataset with factuality annotations, combining automated labeling with manual validation. For short-form QA, we test our method on four QA datasets and treat each response as a single claim.

Our key **contributions** are as follows.

- We develop a new UQ method for RAG, FRANQ, that estimates uncertainty by first assessing faithfulness, and then using uncertainty quantification methods for faithful and unfaithful outputs; see Section 2.
- We develop a long-form QA factuality dataset for RAG. The dataset incorporates both factuality and faithfulness labels, and was built by combining automatic annotation with manual validation for difficult cases; see Section 3.
- We conduct comprehensive experiments on both long- and short-form QA with several LLMs, demonstrating that FRANQ improves the detection of factual errors in RAG outputs compared to other approaches; see Section 4.

2 Uncertainty Quantification for RAG

Let $\mathbf{x}$ be the user query submitted to the RAG system. The system retrieves k passages denoted by $\mathbf{r} = \{r_1, \ldots, r_k\}$, from an external knowledge source using $\mathbf{x}$ as the query. The RAG system then uses an LLM to generate an output $\mathbf{y}$, conditioned on both $\mathbf{x}$ and $\mathbf{r}$.

Autoregressive LLMs produce text sequentially, generating one token at a time. At each step t, the model samples a token $y_t \sim p(\cdot \mid \mathbf{y}_{<t}, \mathbf{x}, \mathbf{r})$, where $\mathbf{y}_{<t}$ denotes the sequence of previously generated tokens. In the case of greedy decoding, this token is selected as the most likely outcome, i.e., $y_t = \arg\max_y p(y \mid \mathbf{y}_{<t}, \mathbf{x}, \mathbf{r})$. From $\mathbf{y}$, we extract a set of l atomic claims denoted as $c_1, \ldots, c_l$. Each claim c_i is associated with a specific span of tokens, $\mathcal{S}(c_i)$, which represents the indices of the tokens in $\mathbf{y}$ that correspond to this particular claim.

A claim c is considered *factually true* if it is generally true, and *false* otherwise. A claim is deemed *faithful* with respect to the retrieved documents $\mathbf{r}$, if it is entailed by them, and *unfaithful* otherwise. While most current benchmarks for evaluating RAG outputs focus on evaluating faithfulness (Dziri et al., 2022; Niu et al., 2024), our main objective is to assess the *factuality* of claims.

General baselines. A straightforward approach to hallucination detection is to apply standard UQ methods to the LLM output conditioned on the joint prompt containing both the user query $\mathbf{x}$ and the retrieved context $\mathbf{r}$. However, this strategy ignores the structural asymmetry between $\mathbf{x}$ and $\mathbf{r}$.

As an illustrative example, a common UQ baseline is to estimate the negative log-probability of a

Category	Uncertainty Quantification Method	Suitable for long-form	Suitable for short-form
Information-based	Max Claim/Sequence Probability	✓	✓
	Perplexity (Fomicheva et al., 2020)	✓	✓
	Mean/Max Token Entropy (Fomicheva et al., 2020)	✓	✓
	CCP (Fadeeva et al., 2024)	✓	✓
Reflexive	P(True) (Kadavath et al., 2022)	✓	✓
Sample diversity	Lexical Similarity (Fomicheva et al., 2020)		✓
	Degree Matrix (Lin et al., 2024)		✓
	Sum of Eigenvalues (Lin et al., 2024)		✓
	Semantic Entropy (Kuhn et al., 2023)		✓
	SentenceSAR (Duan et al., 2024)		✓

Table 1: Summary of UQ methods used as baselines.

claim c under the model distribution:

$$U(c \mid \mathbf{x}, \mathbf{r}) = -\sum_{t \in \mathcal{S}(c)} \log p(y_t \mid \mathbf{x}, \mathbf{r}, \mathbf{y}_{<t}). \quad (1)$$

Table 1 summarizes several other UQ methods that can be applied in this general baseline setting.

2.1 Faithfulness-aware Retrieval Augmented uNcertainty Quantification (FRANQ)

We introduce FRANQ, a new approach for assessing the factuality of claims in RAG outputs by leveraging UQ and explicitly treating $\mathbf{x}$ and $\mathbf{r}$ as separate inputs. The key idea is to first assess whether a generated claim is faithful to $\mathbf{r}$ and then apply different UQ methods depending on the outcome. This yields the following decomposition of the probability that a claim c is true:

$$\begin{aligned} &P(c \text{ is true}) = \quad (2)\\ &P(c \text{ is faithful to } \mathbf{r}) \cdot P(c \text{ is true} \mid \text{faithful}) + \\ &P(c \text{ is unfaithful to } \mathbf{r}) \cdot P(c \text{ is true} \mid \text{unfaithful}), \end{aligned}$$

where $P(c \text{ unfaithf. to } \mathbf{r}) = 1 - P(c \text{ faithful to } \mathbf{r})$. This decomposition isolates three probability components, each of which we approximate using specialized techniques described in Section 2.2:

1. $P(c \text{ is faithful to } \mathbf{r})$;
2. $P(c \text{ is true} \mid \text{faithful})$;
3. $P(c \text{ is true} \mid \text{unfaithful})$.

An overview of FRANQ is visually depicted in Figure 1, and illustrative examples applied to individual claims are provided in Appendix G.

2.2 FRANQ Components

We now describe the three components in equation (2).

Faithfulness. To determine the degree to which a claim c_i is entailed by the retrieved evidence $\mathbf{r}$, we use *AlignScore*, a RoBERTa-based similarity metric fine-tuned for factual alignment (Zha et al., 2023). *AlignScore* is specifically designed to measure factual consistency between a claim and context evidence, making it well suited for claim-level faithfulness estimation in RAG. Importantly, *AlignScore* yields well-calibrated continuous faithfulness estimates rather than near-binary decisions; in practice, many claims exhibit intermediate values due to partial or implicit grounding. We analyze the distribution, calibration, and alternative faithfulness estimators in Appendix C.

In long-form QA, we apply *AlignScore* to each claim–retrieval pair $(c_i, \mathbf{r})$ to get the faithfulness estimate for claim c_i. In short-form QA, where the entire answer $\mathbf{y}$ is treated as a single claim, we prepend the question context and instead evaluate *AlignScore* on $(\mathbf{x} \circ \mathbf{y}, \mathbf{r})$, with '$\circ$' denoting string concatenation.

Factuality under unfaithful condition. When a claim c is unfaithful (not entailed by $\mathbf{r}$), it originates from the LLM's internal knowledge. In this case, we estimate factuality using the model's probability estimates, avoiding distributional shifts arising from conditioning on retrieved context $\mathbf{r}$. Specifically, we introduce a *Parametric Knowledge* method, which computes the likelihood of c based solely on the LLM's parametric knowledge (Mallen et al., 2023) without the retrieved evidence $\mathbf{r}$:

$$p(c \mid \mathbf{x}) = \prod_{t \in \mathcal{S}(c)} p(y_t \mid \mathbf{x}, \mathbf{y}_{<t}). \quad (3)$$

This method does not require generating new responses; instead, it reuses the original tokens and performs a forward pass through the LLM with the retrieved evidence removed from the input.

In long-form QA, we find that *Parametric Knowledge* provides an effective estimate of factuality for unfaithful claims (see Section 4.4). In contrast, short-form QA admits a wider range of general UQ baselines, including methods based on sample diversity (see Table 1). In this setting, we observe that the *Sum of Eigenvalues* method (Lin et al., 2024) offers a better approximation of factuality (see Section 4.4). Therefore, we estimate the factuality of unfaithful claims using *Parametric Knowledge* in long-form QA and *Sum of Eigenvalues* in short-form QA.

Factuality under faithful condition. When a claim c is assessed as faithful to $\mathbf{r}$, the LLM may still fail to apply that evidence correctly to the user query. For example, the LLM may simply choose one of the entities mentioned in $\mathbf{r}$, producing a faithful but incorrect answer to the query $\mathbf{x}$.

To account for such errors, in long-form QA, we estimate uncertainty within the faithful branch using a simple *Max Claim Probability* baseline, $p(c \mid \mathbf{x}, \mathbf{r})$. In short-form QA, alternative baselines are more suitable, particularly *Semantic Entropy* (Kuhn et al., 2023), which better captures uncertainty in this scenario (see Section 4.4).

Therefore, we estimate the factuality for faithful claims with *Max Claim Probability* for long-form QA, and *Semantic Entropy* for short-form QA.

Resulting formula. In summary, we estimate the factuality of the claim c with FRANQ using the following formula:

$$\begin{aligned}\text{FRANQ}(c) = P_{\text{faithful}}(c, \mathbf{r}) \cdot \text{UQ}_{\text{faith}}(c) \\ +\big(1 - P_{\text{faithful}}(c, \mathbf{r})\big) \cdot \text{UQ}_{\text{unfaith}}(c),\end{aligned} \quad (4)$$

where we use *AlignScore* to estimate faithfulness probability P_{faithful} and two UQ methods, UQ_{faith} and $\text{UQ}_{\text{unfaith}}$, selected based on empirical performance for long- and short-form QA scenarios. For long-form QA, we use *Max Claim Probability* (1) for UQ_{faith} and *Parametric Knowledge* (3) for $\text{UQ}_{\text{unfaith}}$. For short-form QA, we use *Semantic Entropy* (Kuhn et al., 2023) for UQ_{faith} and *Sum of Eigenvalues* (Lin et al., 2024) for $\text{UQ}_{\text{unfaith}}$.

We consistently apply the same uncertainty methods across all datasets within each QA setting (short- and long-form), and select uncertainty techniques only based on the nature of the task (using token-level likelihoods for long-form QA and sampling-based metrics for short-form QA).

2.3 Calibrating FRANQ

Since the UQ methods UQ_{faith} and $\text{UQ}_{\text{unfaith}}$ of equation (4) may have different distributions, to avoid inconsistencies and miscalibration among various UQ measures, we calibrate their outputs using isotonic regression on the training data (Vashurin et al., 2025).

Formally, given training dataset $\mathcal{D} = \{(u_i, \text{fact}_i)\}_{i=1}^{N}$ comprising pairs of UQ scores u_i and corresponding binary factuality labels fact_i for the N claims, we calibrate the UQ scores by fitting a non-decreasing function $f\colon \mathbb{R} \to [0, 1]$ through isotonic regression, minimizing the squared error:

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{N} \big(f(u_i) - \text{fact}_i\big)^2, \quad (5)$$

where $\mathcal{F}$ denotes the set of all non-decreasing functions mapping real numbers to probabilities in the interval $[0, 1]$. Isotonic regression directly optimizes over $\mathcal{F}$ without assuming any parametric or functional form for f. This yields a piecewise-constant function $\widehat{f}$ defined on the observed UQ scores that satisfies the monotonicity constraint. During inference, we apply the calibration function $\widehat{f}$ to each UQ score to produce probabilistically meaningful output.

Condition-Calibrated FRANQ. Since UQ_{faith} and $\text{UQ}_{\text{unfaith}}$ represent factuality scores under faithful and unfaithful conditions, respectively, we propose condition-specific calibration. This involves partitioning the training dataset $\mathcal{D}$ into two subsets: faithful claims $\mathcal{D}_{\text{faith}}$ and unfaithful claims $\mathcal{D}_{\text{unfaith}}$. Then, we calibrate UQ_{faith} using the subset $\mathcal{D}_{\text{faith}}$ and $\text{UQ}_{\text{unfaith}}$ using the subset $\mathcal{D}_{\text{unfaith}}$.

We consider FRANQ with condition-specific calibration as our primary method. To evaluate the impact of calibration, we additionally assess two variants: one without any calibration, and another one in which both UQ methods are calibrated using the full training dataset $\mathcal{D}$. The calibration strategies are summarized as follows:

1. **No calibration.** Raw outputs from UQ_{faith} and $\text{UQ}_{\text{unfaith}}$ are directly used in equation (4) without any calibration.
2. **Calibrated.** Both UQ methods are calibrated on the entire training dataset $\mathcal{D}$, disregarding claim faithfulness.
3. **Condition-calibrated.** Each UQ method is calibrated using a subset of the training data corresponding to the respective condition: UQ_{faith} is calibrated using $\mathcal{D}_{\text{faith}}$, and $\text{UQ}_{\text{unfaith}}$ is calibrated using $\mathcal{D}_{\text{unfaith}}$.

3 Datasets for RAG Uncertainty Quantification

Existing datasets for studying RAG hallucinations have serious limitations, as they typically evaluate only context-relative correctness rather than factuality (see Section 5). We argue that factuality is more critical in RAG applications, with faithfulness serving as a complementary perspective. Consequently, an effective dataset should capture both factual errors and contextual misuse.

To address this need, we introduce a new dataset specifically designed for long-form generations, enabling fine-grained analysis of atomic claims.

3.1 Long-form QA Dataset.

Questions. Our long-form QA dataset consists of 76 questions: 44 most challenging questions from RAGTruth (Niu et al., 2024) (identified by those

with highest number of hallucinated claims), and 32 additional technical "how-to" questions generated using GPT-4 via simple prompts (e.g., requesting challenging, domain-diverse technical questions such as *"How does solar power generate electricity?"*). The generated questions were manually inspected to ensure clarity and relevance.

Retrieval Model. For each question, we retrieve the top-k=3 passages using the Facebook Contriever model (Izacard et al., 2021) with embeddings computed over the 2018 English Wikipedia, ensuring high-quality and reliable evidence passages.

LLMs. We construct four model-specific dataset subsets by generating long-form answers to all 76 questions with their corresponding retrieved passages, using greedy decoding independently for each model: Llama 3B Instruct, Llama 8B Instruct (Grattafiori et al., 2024), Falcon 3B Base (Team, 2024), and Gemma 4B Instruct (Team et al., 2025). These subsets enable UQ methods to be evaluated on top of in-policy generations for each model.

Claim Extraction. For each generated answer, we extract atomic claims and their corresponding token spans using the approach of (Wang et al., 2024b; Vashurin et al., 2025). First, GPT-4o extracts decontextualized atomic claims from the entire text paragraph through a dedicated prompt. Then, for each claim, a second prompt instructs GPT-4o to list the relevant words from the original text, which we map to token spans. Applying this procedure, we obtain 1,782 claims for Llama 3B Instruct and 1,548 claims for Falcon 3B Base. From these claims, we select 500 claims for train set, reserving the remainder for test set. The prompts used for claims extraction and mapping are listed in Appendix B.

Annotation. We annotate factuality and faithfulness using GPT-4o-search with dedicated prompts. Faithfulness labels categorize each claim as either *faithful* or *unfaithful*. Factuality annotations include three categories: *True*, *False*, and *Unverifiable*. We retain only verifiable claims (True or False), binarizing the labels accordingly.

During verification of the automatic annotations, we found that the *False* and *Unverifiable* categories are particularly difficult to assess automatically, so we manually reviewed all claims assigned to either category and corrected their labels when necessary.

Further details on prompts, dataset statistics and annotation scheme are provided in Appendix B.

3.2 Short-form QA Datasets

In contrast to long-form QA, where evaluating factuality requires extracting model-specific claims and annotating them, short-form QA provides gold-standard answers for each question. This allows us to directly compare each model's generated answer with the ground-truth answer, yielding an automatic factuality judgment without additional manual annotation or claim-level verification.

Questions. We adapt four short-form QA datasets for RAG evaluation: TriviaQA (Joshi et al., 2017), SimpleQA (Wei et al., 2024a), Natural Questions (Kwiatkowski et al., 2019), and PopQA (Mallen et al., 2023). For each dataset, we sample 200 questions for training and 1000 for testing, and we treat each model response as a single claim.

RAG Models. We use the same retrieval model as in the long-form setting, selecting the top-k=5 passages per question. For LLMs, we use the same two Llama models and the Falcon model, along with an additional model: Gemma 12B Instruct (Team et al., 2025).

Annotation. We evaluate factuality of each generated answer by comparing it against the gold-standard answer using GPT-4o, following the procedure of (Wei et al., 2024a), which has been shown to yield reliable factuality judgments.

4 Experiments

In this section, we evaluate FRANQ and corresponding baselines on both the short-form and long-form benchmarks described in Section 3. For all experiments, we fix the retrieval process and the underlying white-box LLM, and we assess the factual accuracy of the model-generated claims.

Later, through ablation studies, we examine the contribution of the individual FRANQ components, $P(\text{faithful})$, UQ_{faith}, and $\text{UQ}_{\text{unfaith}}$, as well as the effect of varying the amount of training data.

4.1 Experimental Setup

UQ baselines. We group all UQ methods into four categories: (1) general baselines, (2) RAG-specific baselines, (3) XGBoost-based methods, and (4) three variants of our proposed FRANQ method, each using a different calibration strategy.

General baselines. We compare FRANQ with general baselines, which consist of standard UQ methods applied directly to the LLM's output distribution without using any RAG-specific structure.

Method	Llama 3B Instruct		Falcon 3B Base		Llama 8B Instruct		Gemma 4B Instruct	
	PR-AUC ↑	PRR ↑	PR-AUC ↑	PRR ↑	PR-AUC ↑	PRR ↑	PR-AUC ↑	PRR ↑
General Baselines								
Max Claim Prob.	.058	-.029	.126	.258	.055	.118	.061	.0
P(True)	.117	.207	.077	.170	.071	.112	.096	.148
Perplexity	.056	-.081	.090	.165	.075	.090	.048	-.071
Max Token Entropy	.109	.115	.130	.219	**.102**	.138	.051	-.003
CCP	.085	.169	.162	.181	.061	.108	.087	.216
RAG-Specific Baselines								
AlignScore	.075	.108	.104	.233	.068	.119	.061	.058
Parametric Knowledge	.064	.018	.067	.029	.059	.047	.112	.183
XGBoost								
XGBoost (all UQ features)	.124	.206	.088	.198	.044	-.015	.073	.085
XGBoost (FRANQ features)	.111	.149	.080	.086	.048	.017	.090	.158
FRANQ								
FRANQ no calibration	.100	.181	.135	**.362**	.063	.162	.080	.200
FRANQ calibrated	.103	**.256**	.074	.090	.043	-.047	**.150**	**.401**
FRANQ condition-calibrated	**.140**	.223	**.173**	.354	.081	**.184**	.090	.208

Table 2: Results on long-form QA benchmark with factuality target. Higher values indicate better performance. In every setting, the top-performing method is one of the FRANQ variants.

For implementation, we use the LM-Polygraph library (Fadeeva et al., 2023). A complete list of methods we used is provided in Table 1.

RAG-specific baselines. We also evaluate the two FRANQ components in isolation, *AlignScore* and *Parametric Knowledge*, to assess how much their combination in FRANQ improves over using each component individually (see Section 2.2).

XGBoost methods. We include XGBoost models trained on factuality labels using two feature sets: (1) the three components used in FRANQ (*AlignScore*, UQ_{faith}, $\text{UQ}_{\text{unfaith}}$), and (2) all available unsupervised UQ method.

FRANQ. Finally, we evaluate three FRANQ variants with different calibration strategies for UQ_{faith} and $\text{UQ}_{\text{unfaith}}$ (see Section 2.3): *no calibration*, *calibrated*, and *condition-calibrated*.

Evaluation measures. Each UQ method produces factuality estimates, which we compare against binary gold-standard labels using PR-AUC, treating false claims as the positive class to emphasize their detection. We also assess rejection performance using the Prediction Rejection Ratio (PRR; Mallen et al., 2023) with a maximum rejection threshold of 0.5. PRR measures how effectively the model rejects uncertain predictions while retaining accurate ones, capturing its ability to prioritize reliable outputs.

4.2 Long-Form QA Results

For long-form QA, we evaluate each UQ method using PR-AUC and PRR across four models (Llama 3B Instruct, Falcon 3B Base, Llama 8B Instruct and Gemma 4B Instruct), see Table 2. The condition-calibrated FRANQ achieves the best PR-AUC and second-best PRR for Llama 3B Instruct and Falcon 3B Base, while for Llama 8B Instruct it attains the best PRR and second-best PR-AUC. The calibrated FRANQ achieves the highest PRR for Llama 3B Instruct and the highest PR-AUC and PRR for Gemma 4B Instruct. The non-calibrated FRANQ also performs strongly, ranking first and second in PRR for Falcon 3B Base and Llama 8B Instruct, respectively. Overall, FRANQ demonstrates strong and consistent performance across all models.

4.3 Short-Form QA Results

For short-form QA, we evaluate UQ methods using PR-AUC and PRR across four models (Llama 3B Instruct, Llama 8B Instruct, Falcon 3B Base, Gemma 12B Instruct) and four datasets. To account for dataset variability, we report mean scores averaged over datasets, following Vashurin et al. (2025) (Table 3); per-dataset results appear in Appendix A.

Condition-calibrated FRANQ achieves the best mean performance across all models and both measures, except for mean PRR on Gemma 12B Instruct, where it ranks second. Calibrated FRANQ ranks second for PRR on Llama 3B Instruct and for both PR-AUC and PRR on Llama 8B Instruct. Among unsupervised methods, Degree Matrix and Lexical Similarity perform strongly, ranking second-best in several settings.

Method	Llama 3B Instruct		Falcon 3B Base		Llama 8B Instruct		Gemma 12B Instruct	
	PR-AUC ↑	PRR ↑	PR-AUC ↑	PRR ↑	PR-AUC ↑	PRR ↑	PR-AUC ↑	PRR ↑
General Baselines								
Max Sequence Prob.	.558	.454	.628	.256	.569	.407	.400	.162
Mean Token Entropy	.594	.481	.613	.242	.640	.491	.423	.230
CCP	.551	.443	.641	.304	.553	.417	.412	.198
Lexical Similarity	.564	.479	.618	.277	.639	.532	.430	.240
Degree Matrix	.629	.520	.702	.464	.627	.492	.464	.260
Sum of Eigenvalues	.628	.518	.700	.460	.628	.489	.467	.260
Semantic Entropy	.613	.525	.623	.278	.637	.519	.466	.261
SentenceSAR	.571	.483	.602	.263	.556	.414	.416	.174
RAG-specific Baselines								
AlignScore	.415	.207	.666	.372	.432	.224	.376	.158
Parametric Knowledge	.425	.247	.556	.104	.499	.330	.364	.105
XGBoost								
XGBoost (all UQ features)	.594	.494	.705	.462	.634	.503	.474	**.301**
XGBoost (FRANQ features)	.526	.409	.670	.368	.524	.385	.414	.196
FRANQ								
FRANQ no calibration	.553	.403	.641	.345	.523	.340	.447	.225
FRANQ calibrated	.628	.537	.672	.411	.644	.534	.481	.258
FRANQ condition-calibrated	**.631**	**.541**	**.711**	**.477**	**.647**	**.540**	**.496**	.283

Table 3: Results in PR-AUC↑ and PRR↑, averaged across four QA datasets for Llama 3B Instruct, Falcon 3B Base, Llama 8B Instruct and Gemma 12B Instruct. The condition-calibrated FRANQ is top-performing across all settings, except mean PRR on Gemma 12B Instruct, where it ranks second.

4.4 Ablation Studies

In this section, we summarize the main observations from ablation studies examining (1) the contribution of FRANQ's components, (2) robustness to retrieval noise, (3) the effect of supervision and (4) computational efficiency. Complete experimental descriptions, tables, and additional ablations are provided in Appendix D.

Analysis of FRANQ's components. Figure 2 reports the PRR of FRANQ with condition calibration for different choices of UQ_{faith} and $\text{UQ}_{\text{unfaith}}$, evaluated on a subset of 200 questions from each dataset using the Llama 3B Instruct model.

On the long-form QA dataset (see Figure 2(a)), performance is largely insensitive to the choice of UQ_{faith}, whereas the choice of $\text{UQ}_{\text{unfaith}}$ is decisive: using Parametric Knowledge as $\text{UQ}_{\text{unfaith}}$ yields the best PRR across essentially all UQ_{faith} options.

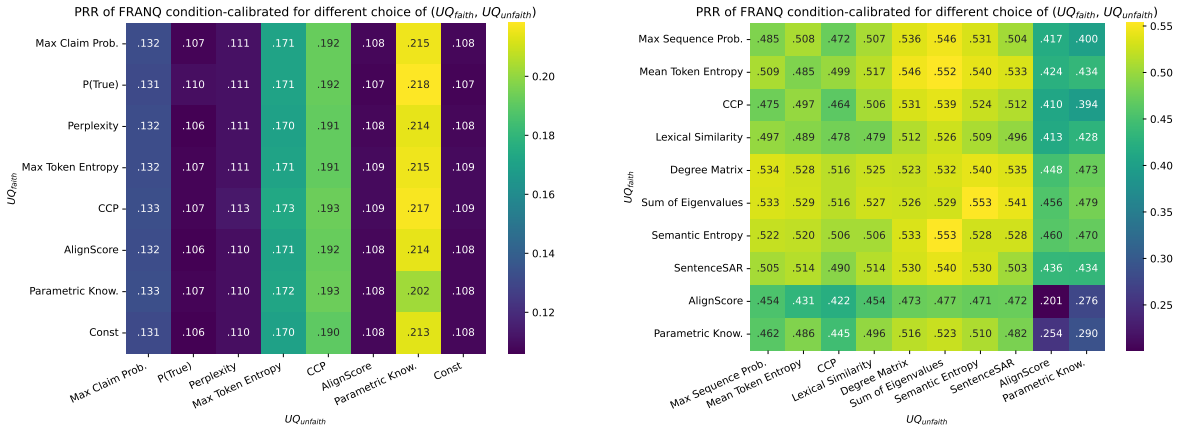
On short-form QA (see Figure 2(b)), many combinations of UQ_{faith} and $\text{UQ}_{\text{unfaith}}$ perform similarly, suggesting that FRANQ is relatively robust to these design choices. The configuration used in our short-form experiments (Semantic Entropy for UQ_{faith} and Sum of Eigenvalues for $\text{UQ}_{\text{unfaith}}$) achieves best observed value of PRR $= 0.553$.

We further examine alternative faithfulness modeling strategies in Appendix D.1. Replacing continuous AlignScore probabilities with binary thresholding degrades performance, underscoring the value of probabilistic faithfulness weighting.

Robustness to retrieval noise. We evaluate FRANQ under corrupted retrievals by randomly replacing a fraction of retrieved documents with unrelated passages, simulating noisy RAG settings (Appendix A.1). Results show that even with 50% retrieval noise, calibrated and condition-calibrated FRANQ achieve top-ranked performance, indicating that the method can effectively handle substantial amounts of irrelevant evidence.

Effect of supervision. We analyze the impact of training set size on supervised FRANQ variants by varying the number of labeled calibration examples used for training (see Appendix D.2). Performance improves with additional data and then saturates, with the optimal training size occurring at approximately 300 instances for long-form QA, while for short-form QA performance stabilizes at around only 120 training samples.

Computational efficiency. We analyze the runtime overhead, training cost, and model size of FRANQ's uncertainty components in Appendix E. Overall, FRANQ introduces minimal training cost (fitting isotonic regression takes less than a second), while producing extremely compact models (on the order of hundreds of bytes), making calibrated FRANQ variants practical for real-world deployment.



(a) PRR on long-form QA dataset.

(b) PRR on short-form QA benchmark (mean across 4 datasets).

Figure 2: Comparison of FRANQ condition-calibrated with different choice of UQ_{faith} and $UQ_{unfaith}$.

5 Related Work

Uncertainty Quantification for RAG. Several UQ methods for RAG analyze how retrieved knowledge influences LLM outputs, including lookback-ratio classifiers (Chuang et al., 2024), feature-based regression of retrieved versus parametric reliance (Sun et al., 2025), uncertainty estimation via the signal-to-noise ratio of output probabilities across samples (Li et al., 2024), and prompt–response relevance modeling (Hu et al., 2024). These methods evaluate hallucinations only relative to retrieved context and often incur computational overhead, while search-based approaches such as SAFE (Wei et al., 2024b) rely on LLM agents and web verification. In contrast, FRANQ is a lightweight, self-contained UQ framework that probabilistically combines faithfulness to retrieved context with truthfulness under both faithful and unfaithful conditions, without additional training or external verification.

When retrieval is absent, uncertainty is typically estimated from internal model knowledge using white-box (Fomicheva et al., 2020; Kadavath et al., 2022; Kuhn et al., 2023; Fadeeva et al., 2024; Duan et al., 2024) or black-box (Fomicheva et al., 2020; Lin et al., 2024) methods. FRANQ unifies these settings by jointly modeling retrieval-related and intrinsic uncertainty within a single probabilistic framework.

Factuality/Hallucination Datasets for RAG. Hallucination detection in RAG relies on labeled datasets of factual errors. RAGTruth (Niu et al., 2024) offers multi-domain, span-level annotations, but excludes cases where models produce correct information independent of the retrieved context.

Knowledge-grounded dialogue datasets such as Wizard of Wikipedia (Dinan et al., 2019) and FaithDial (Dziri et al., 2022) pair responses with external sources but prioritize conversational coherence, treating any content not grounded in the provided context as hallucination regardless of factual correctness.

QA-based benchmarks such as RAGBench (Friel et al., 2024) and AdaptiveRAG (Moskvoretskii et al., 2025) define hallucinations strictly relative to the given context, whereas FRANQ estimates factuality even when generation diverges from retrieved evidence, enabling a more comprehensive assessment of model reliability.

6 Conclusion and Future Work

We introduced FRANQ, a new method for quantifying the factuality of claims in RAG output based on their faithfulness. Across both long-form and short-form QA tasks and multiple LLMs, FRANQ consistently outperforms existing unsupervised UQ baselines, RAG-specific methods, and supervised classifiers. We also presented a new long-form QA dataset annotated for both factuality and faithfulness using a hybrid of automatic and manual labeling.

Our approach opens several promising directions for future research. One direction is to extend uncertainty modeling to the retrieval stage, thus allowing systems to account for noisy, incomplete, or conflicting evidence. Another is to leverage FRANQ's uncertainty signals for generation-time control and post-editing, thus enabling more reliable and interpretable RAG systems.

Limitations

While FRANQ provides strong hallucination detection performance on average, it does not guarantee ideal hallucination detection in every situation, as this is a challenging task.

FRANQ assumes that the retrieved evidence is always factual and takes precedence over the LLM's parametric knowledge. In theory, this can be achieved through careful selection and curation of document sources within the search index. However, ensuring complete factual accuracy in real-world applications might be challenging as the size of the index grows.

Since FRANQ leverages the calibration of its components, it might be considered as supervised. To address this concern, we showed that it also outperforms supervised methods.

Ethical Considerations

FRANQ is designed to reduce the spread of factual errors by enhancing the interpretability and reliability of language model outputs. By distinguishing between factuality and faithfulness, it helps prevent misclassification of factually correct but unsupported claims. However, FRANQ does not actively prevent the generation of hallucinations and instead relies on downstream filtering. Its effectiveness, therefore, depends on integration into larger pipelines with proper safeguards.

FRANQ assumes that the retrieved context is factual and trustworthy. In real-world applications, the retrieved documents may be biased, outdated, or incorrect, which could compromise the method's output. Careful curation of retrieval sources and monitoring of retrieval quality are crucial to avoid reinforcing harmful biases or misinformation.

The dataset used for evaluation relies on the output from GPT-4o. While we manually validated a subset of the annotations, some inherent biases from the underlying model may persist. We encourage future work to explore more diverse annotation strategies, including community-sourced validation.

Improving the factuality estimation can support safer AI deployment, especially in knowledge-intensive domains such as education, healthcare, or law. However, the system should not be considered a replacement for human fact-checkers. It is best used as a decision-support tool rather than a source of truth.

References

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. Lookback Lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 1419–1436. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In International Conference on Learning Representations.

Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2024. Retrieve only when it needs: Adaptive retrieval augmentation for hallucination mitigation in large language models. arXiv preprint arXiv:2402.10612.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5050–5063. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M Ponti, and Siva Reddy. 2022. FaithDial: A faithful benchmark for information-seeking dialogue. Transactions of the Association for Computational Linguistics, 10:1473–1490.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In Findings of the Association for Computational Linguistics: ACL 2024, pages 9367–9385. 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. Association for Computational Linguistics.

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.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. RAGBench: Explainable benchmark for retrieval-augmented generation systems. arXiv preprint arXiv:2407.11005.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. Preprint, arXiv:2407.21783.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 1321–1330. PMLR.

Haichuan Hu, Yuhan Sun, and Quanjun Zhang. 2024. LRP4RAG: Detecting hallucinations in retrieval-augmented generation via layer-wise relevance propagation. arXiv preprint arXiv:2408.15533.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2):1–55.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1601–1611. 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, and 1 others. 2022. Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221.

Hyuhng Joon Kim, Youna Kim, Sang-goo Lee, and Taeuk Kim. 2025. When to speak, when to abstain: Contrastive decoding with abstention. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9710–9730, Vienna, Austria. Association for Computational Linguistics.

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.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7:452–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems, volume 33, pages 9459–9474.

Zixuan Li, Jing Xiong, Fanghua Ye, Chuanyang Zheng, Xun Wu, Jianqiao Lu, Zhongwei Wan, Xiaodan Liang, Chengming Li, Zhenan Sun, and 1 others. 2024. UncertaintyRAG: Span-level uncertainty enhanced long-context modeling for retrieval-augmented generation. arXiv preprint arXiv:2410.02719.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. Transactions on Machine Learning Research.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9802–9822. Association for Computational Linguistics.

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.

Sewon Min, Kalpesh Krishna, Xinxi 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. Association for Computational Linguistics.

Viktor Moskvoretskii, Maria Marina, Mikhail Salnikov, Nikolay Ivanov, Sergey Pletenev, Daria Galimzianova, Nikita Krayko, Vasily Konovalov, Irina Nikishina, and Alexander Panchenko. 2025. Adaptive retrieval without self-knowledge? bringing uncertainty back home. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6355–6384.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10862–10878.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In International Conference on Machine Learning, pages 31210–31227. PMLR.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3784–3803. Association for Computational Linguistics.

ZhongXiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025. ReDeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In The Thirteenth International Conference on Learning Representations.

Falcon-LLM Team. 2024. The Falcon 3 family of open models.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. arXiv preprint arXiv:2503.19786.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Akim Tsvigun, Daniil Vasilev, Rui Xing, Abdelrahman Boda Sadallah, Kirill Grishchenkov, Sergey Petrakov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. Benchmarking uncertainty quantification methods for large language models with LM-Polygraph. Transactions of the Association for Computational Linguistics, 13:220–248.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö Arık. 2024a. Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. arXiv preprint arXiv:2410.07176.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025. Retrieval-augmented generation with conflicting evidence. arXiv preprint arXiv:2504.13079.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024b. Factcheck-Bench: Fine-grained evaluation benchmark for automatic fact-checkers. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 14199–14230. Association for Computational Linguistics.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024a. Measuring short-form factuality in large language models. arXiv 2411.04368.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, and 1 others. 2024b. Long-form factuality in large language models. In Advances in Neural Information Processing Systems, volume 37, pages 80756–80827.

Chenxu Yang, Zheng Lin, Chong Tian, Liang Pang, Lanrui Wang, Zhengyang Tong, Qirong Ho, Yanan Cao, and Weiping Wang. 2024. A factuality and diversity reconciled decoding method for knowledge-grounded dialogue generation. arXiv preprint arXiv:2407.05718.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11328–11348. Association for Computational Linguistics.

Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S Yu. 2024. Trustworthiness in retrieval-augmented generation systems: A survey. arXiv preprint arXiv:2409.10102.

A Additional Short-form QA Results

In Table 3 of the main text, we report aggregated results for short-form QA using mean values for ease of presentation. Here, we provide the full results for each of the four QA datasets (Natural Questions, PopQA, TriviaQA, SimpleQA) for both Llama 3B Instruct (see Table 6) and Falcon 3B Base (see Table 7).

For Llama 3B Instruct, FRANQ calibrated and FRANQ condition-calibrated methods consistently rank among the top performers. They are the top two methods on TriviaQA and SimpleQA. On PopQA, FRANQ condition-calibrated ranks among the top three methods, alongside Semantic Entropy and Max Token Entropy. On Natural Questions, it ranks in the top four, along with DegreeMatrix, Eccentricity, and Sum of Eigenvalues. Overall, both FRANQ variants achieve the best average performance across all datasets.

For Falcon 3B Base, FRANQ condition-calibrated achieves the top performance on TriviaQA and second-best performance on Natural Questions. It also ranks among the top three methods on PopQA and among the top four on SimpleQA, alongside Degree Matrix, Sum of Eigenvalues, and XGBoost (all features). On average, FRANQ condition-calibrated is the leading method across the four datasets.

Method	**MeanValue ↑**	
	PRAUC ↑	**PRR ↑**
General Baselines		
Max Sequence Prob.	.638	.489
Mean Token Entropy	.651	.460
CCP	.629	.472
Lexical Similarity	.668	.512
Degree Matrix	.687	.548
Sum of Eigenvalues	.686	.536
Semantic Entropy	.666	.537
SentenceSAR	.648	.523
RAG-specific Baselines		
AlignScore	.507	.234
Parametric Knowledge	.502	.220
XGBoost		
XGBoost (all UQ features)	.684	.544
XGBoost (FRANQ features)	.579	.402
FRANQ		
FRANQ no calibration	.586	.393
FRANQ calibrated	<u>.692</u>	<u>.549</u>
FRANQ condition-calibrated	**.695**	**.553**

Table 4: Aggregated results with shuffled retrievals on 4 QA datasets for Llama 3B Instruct.

A.1 Short-form QA with Corrupted Retrievals

Real-world retrieval systems are often imperfect, occasionally introducing irrelevant passages that degrade both LLM and FRANQ performance. To simulate this, we randomly shuffled 50% of the retrieved passages across four QA datasets, ensuring no shuffled sample retained its original retrievals. Corrupted retrievals added substantial noise to the prompt, causing an average 7% accuracy drop. Nevertheless, FRANQ methods remained robust, maintaining competitive results (aggregated in Table 4, full results in Table 5). Under this setting, FRANQ condition-calibrated and FRANQ calibrated variants achieve top-1 and top-2 overall performance in average quality across the four datasets, measured by PR-AUC and PRR.

Method	NQ		PopQA		TriviaQA		SimpleQA		Mean Value	
	PRAUC ↑	PRR ↑	PRAUC ↑	PRR ↑	PRAUC ↑	PRR ↑	PRAUC ↑	PRR ↑	PRAUC ↑	PRR ↑
General Baselines										
Max Sequence Prob.	.477	.216	.645	.509	.556	.477	**.874**	.755	.638	.489
Mean Token Entropy	.571	.346	.678	.510	.580	.460	.777	.525	.652	.460
CCP	.494	.269	.629	.481	.571	.482	.822	.657	.629	.472
Lexical Similarity	.580	.370	.696	.547	.579	.489	.819	.639	.669	.511
Degree Matrix	.570	.359	.661	.537	.645	.553	.871	.745	.687	.549
Sum of Eigenvalues	.573	.360	.664	.516	.637	.542	.872	.728	.687	.537
Semantic Entropy	.537	.317	**.697**	**.587**	.565	.493	.867	.752	.667	.537
SentenceSAR	.465	.217	.695	.546	.566	.517	.866	**.811**	.648	.523
RAG-specific Baselines										
AlignScore	.493	.259	.506	.225	.417	.213	.613	.240	.507	.234
Parametric Knowledge	.406	.104	.564	.342	.494	.377	.545	.059	.502	.221
XGBoost										
XGBoost (all UQ features)	.561	.354	.657	.474	**.656**	**.594**	.860	.753	.684	.544
XGBoost (FRANQ features)	.488	.269	.579	.376	.471	.370	.779	.592	.579	.402
FRANQ										
FRANQ no calibration	.419	.127	.626	.452	.527	.417	.774	.575	.587	.393
FRANQ calibrated	.601	.401	.675	.549	.624	.529	.868	.717	.692	.549
FRANQ condition-calibrated	**.603**	**.413**	.675	.549	.636	.536	.867	.715	.695	.553

Table 5: Results in PRAUC↑ and PRR↑ with shuffled retrievals on 4 QA datasets for Llama 3B Instruct, including mean performance across datasets.

Method	NQ			PopQA			TriviaQA			SimpleQA		
	AUROC ↑	PR-AUC↑	PRR↑	AUROC ↑	PR-AUC↑	PRR↑	AUROC ↑	PR-AUC↑	PRR↑	AUROC ↑	PR-AUC↑	PRR↑
General Baselines												
Max Sequence Prob.	.680	.440	.292	.745	.550	.421	.774	.529	.478	.833	.712	.625
Mean Token Entropy	.723	.503	.389	.768	**.607**	.455	.796	.569	.523	.809	.697	.555
CCP	.705	.471	.357	.709	.526	.393	.767	.528	.471	.800	.680	.552
Lexical Similarity	.720	.494	.386	.763	.571	.462	.775	.508	.485	.818	.685	.585
Degree Matrix	**.751**	**.557**	.421	.738	.570	.421	.816	**.626**	.570	.852	.764	.668
Sum of Eigenvalues	.749	.553	.411	.740	.564	.416	.816	.621	.561	.861	.774	.686
Semantic Entropy	.727	.518	.373	**.776**	.602	**.496**	.801	.565	.546	.863	.766	.684
SentenceSAR	.678	.395	.269	.762	.562	.459	.794	.560	.521	.858	.767	.682
RAG-specific Baselines												
AlignScore	.682	.427	.312	.566	.371	.079	.631	.387	.215	.645	.473	.221
Parametric Knowledge	.626	.371	.203	.664	.470	.290	.727	.467	.397	.490	.393	.096
XGBoost												
XGBoost (all UQ features)	.712	.504	.375	.744	.565	.433	.773	.546	.486	.835	.760	.683
XGBoost (FRANQ features)	.651	.412	.283	.690	.503	.350	.692	.441	.328	.860	.747	.676
FRANQ												
FRANQ no calibration	.637	.456	.268	.676	.481	.278	.773	.557	.467	.826	.717	.601
FRANQ calibrated	.735	.529	.405	.765	.597	.468	**.821**	.623	**.580**	.869	.761	.695
FRANQ condition-calibrated	.748	.526	.409	.763	.605	.477	.821	.618	.576	**.877**	**.776**	**.703**

Table 6: Results on 4 QA datasets for Llama 3B Instruct.

Method	NQ			PopQA			TriviaQA			SimpleQA		
	AUROC ↑	PR-AUC↑	PRR↑	AUROC ↑	PR-AUC↑	PRR↑	AUROC ↑	PR-AUC↑	PRR↑	AUROC ↑	PR-AUC↑	PRR↑
General Baselines												
Max Sequence Prob.	.599	.555	.186	.653	.649	.259	.590	.487	.163	.625	.820	.416
Mean Token Entropy	.599	.542	.184	.657	.662	.279	.557	.432	.108	.656	.814	.396
CCP	.632	.576	.258	.659	.648	.297	.620	.518	.212	.635	.822	.448
Lexical Similarity	.581	.486	.115	.721	.691	.412	.587	.476	.157	.650	.818	.422
Degree Matrix	.653	.571	.258	.787	.777	**.571**	.660	.565	.311	**.795**	**.896**	**.718**
Sum of Eigenvalues	.651	.568	.260	**.789**	**.780**	.570	.661	.559	.299	.791	.894	.713
Semantic Entropy	.561	.494	.086	.718	.698	.415	.584	.468	.155	.685	.831	.456
SentenceSAR	.509	.455	.012	.755	.707	.463	.523	.395	.026	.739	.850	.552
RAG-specific Baselines												
AlignScore	.655	.613	.320	.639	.652	.262	.685	.540	.341	.748	.860	.566
Parametric Knowledge	.556	.486	.089	.611	.590	.210	.567	.420	.086	.512	.729	.030
XGBoost												
XGBoost (all UQ features)	**.679**	**.617**	**.340**	.772	.748	.507	.693	.572	.340	.787	.885	.661
XGBoost (FRANQ features)	.640	.596	.292	.694	.712	.414	.624	.517	.236	.731	.853	.532
FRANQ												
FRANQ no calibration	.576	.496	.113	.732	.716	.448	.609	.492	.205	.738	.862	.616
FRANQ calibrated	.617	.541	.215	.773	.749	.520	.626	.513	.228	.769	.885	.682
FRANQ condition-calibrated	.668	.591	.331	.781	.764	.533	**.695**	**.606**	**.377**	.776	.886	.668

Table 7: Results on 4 QA datasets for Falcon 3B Base.

B Prompts and Setup

B.1 Short-form QA

For short-form QA experiments, we paired each question with the top-5 retrieved Wikipedia passages and used the prompt format in Figure 3. For annotation, GPT-4o was given the question, model-generated answer, and gold answer, and asked to label responses as correct, incorrect, or not attempted (excluded from evaluation), following Wei et al. (2024a). Table 8 reports dataset statistics.

B.2 Long-form QA

For long-form QA experiments, we used each question with the top-3 retrieved Wikipedia passages. All models followed the prompt format shown in Figure 4. Extracted answers were decomposed into atomic claims using the prompt in Figure 5, and each claim was matched to its corresponding span in the original sentence using Figure 6. Claims without identifiable spans (e.g., due to annotation inconsistencies) were excluded. The remaining claims were annotated for factuality and faithfulness using automatic annotation followed by manual validation (Appendix B.3, B.4). Table 9 reports dataset statistics.

Compared to prior decomposition methods such as FActScore (Min et al., 2023), our approach is more careful: we decompose entire texts rather than individual sentences to reduce redundancy and ambiguity, and we produce decontextualized claims to simplify verification. Claim quality was further examined during manual validation in complex cases.

```
Contents (not necessarily includes answer to the following question):
Title: {title1}
Content: {retrieval1}
...
Title: {title5}
Content: {retrieval5}
Question: {question}
Answer (single line):
```

Figure 3: Prompt used in short-form QA datasets. Titles and retrievals correspond to the Wikipedia page title and the passage retrieved from it.

```
Using the context provided below, answer the question with a balanced approach. Ensure your response contains an equal number of claims or details drawn directly from the context and from your own knowledge:
Context: passage 1:{retrieval1}
passage 2:{retrieval2}
passage 3:{retrieval3}
Question: {question}
Answer:
```

Figure 4: Prompt used in long-form QA datasets. Retrievals corresponds to the Wikipedia passage retrieved for input question.

```
Your task is to decompose the text into atomic claims.
Let's define a function named decompose(input:str).
The returned value should be a list of strings, where each string should be a context-independent, fully atomic claim, representing one fact. Atomic claims are simple, indivisible facts that do not bundle multiple pieces of information together.

### Guidelines for Decomposition:
1. **Atomicity**: Break down each statement into the smallest possible unit of factual information. Avoid grouping multiple facts in one claim. For example:
    - Instead of: "Photosynthesis in plants converts sunlight, carbon dioxide, and water into glucose and oxygen."
    - Output: ["Photosynthesis in plants converts sunlight into glucose.", "Photosynthesis in plants converts carbon dioxide into glucose.", "Photosynthesis in plants converts water into glucose.", "Photosynthesis in plants produces oxygen."]
    - Instead of: "The heart pumps blood through the body and regulates oxygen supply to tissues."
    - Output: ["The heart pumps blood through the body.", "The heart regulates oxygen supply to tissues."]
    - Instead of: "Gravity causes objects to fall to the ground and keeps planets in orbit around the sun."
    - Output: ["Gravity causes objects to fall to the ground.", "Gravity keeps planets in orbit around the sun."]
2. **Context-Independent**: Each claim must be understandable and verifiable on its own without requiring additional context or references to other claims. Avoid vague claims like "This process is important for life."
3. **Precise and Unambiguous**: Ensure the claims are specific and avoid combining related ideas that can stand independently.
4. **No Formatting**: The response must be a Python list of strings without any extra formatting, code blocks, or labels like "python".

### Example:
If the input text is: "Mary is a five-year-old girl. She likes playing piano and doesn't like cookies."
The output should be: ["Mary is a five-year-old girl.", "Mary likes playing piano.", "Mary doesn't like cookies."]
Note that your response will be passed to the python interpreter, SO NO OTHER WORDS!

decompose("{text}")
```

Figure 5: Prompt template used with GPT-4o for decomposing an answer into a set of atomic claims.

```
Task: Analyze the given text and the claim (which was extracted from the text). For each sentence in the text:
1. Copy the sentence exactly as it appears in the text.
2. Identify the words from the sentence that are related to the claim, in the same order they appear in the sentence. If no words are related, output "No related words."

Example:
Text: "Sure! Here are brief explanations of each type of network topology mentioned in the passages: [...]"
Claim: "Distributed Bus topology connects all network nodes to a shared transmission medium via multiple endpoints."
Answer:
Sentence: "Sure! Here are brief explanations [...]"
Related words from this sentence (same order they appear in the sentence): No related words
Sentence: "2. Distributed Bus: In a Distributed Bus topology, [...]"
Related words from this sentence (same order they appear in the sentence): "Distributed", "Bus", "topology", "all", "network", [...]
Sentence: [... More sentences follow ...]

Now analyze the following text using this format:
Text: {text}
Claim: {claim}
Answer:
```

Figure 6: Prompt template used with GPT-4o to identify the span in the original text corresponding to each atomic claim. The model is instructed to process each sentence and extract words relevant to the claim, preserving their order. Parts of the 1-shot example have been omitted for brevity.

Model	Dataset	Train Size	Test Size	True	False	Unverifiable	Mean Generation Length (characters)
Llama 3B Instruct	NQ	200	1000	62.4 %	27.6 %	10.0 %	180.1
	PopQA	200	1000	50.2 %	22.4 %	27.3 %	149.2
	TriviaQA	200	1000	68.0 %	22.3 %	9.7 %	114.4
	SimpleQA	200	1000	29.5 %	14.4 %	56.1 %	159.9
Falcon 3B Base	NQ	200	1000	44.1 %	37.6 %	18.3 %	352.2
	PopQA	200	1000	42.6 %	41.6 %	15.8 %	260.3
	TriviaQA	200	1000	57.2 %	32.8 %	10.0 %	324.7
	SimpleQA	200	1000	25.5 %	65.6 %	8.8 %	286.9

Table 8: Statistics of datasets used in short-form QA benchmark.

Model	Train Size	Test Size	True	False	Unverifiable	Faithful	Unfaithful	Undefined	Mean Generation Length (characters)
Llama 3B Instruct	600	1182	91.0 %	5.8 %	3.1 %	37.3 %	62.6 %	0.1 %	1725.4
Falcon 3B Base	600	948	91.4 %	6.0 %	2.6 %	38.2 %	61.5 %	0.3 %	1720.2
Llama 8B Instruct	300	500	89.4 %	5.0 %	5.6 %	34.6 %	64.6 %	0.8 %	1856.4
Gemma 4B Instruct	300	500	88.8 %	5.7 %	5.5 %	44.7 %	54.5 %	0.8 %	1708.3

Table 9: Statistics of datasets used in long-form QA benchmark.

B.3 Automatic Annotation of Claims

For faithfulness annotation, each claim is automatically assigned one of three categories: "faithful" (the context supports the statement), "unfaithful-contra" (the context contradicts it), or "unfaithful-neutral" (the context provides neither support nor contradiction). For experimental evaluation, these labels are binarized: faithful $\rightarrow$ 1, and unfaithful-contra or unfaithful-neutral $\rightarrow$ 0, since the unfaithful-contra class constitutes less than 5% of the data.

For factuality annotation, each claim is likewise assigned one of three categories: "True" (factually correct), "False" (factually incorrect), and "unverifiable" (accuracy cannot be determined without relying on the provided context). Factuality is then binarized by retaining only verifiable claims, mapping False $\rightarrow$ 1 and True $\rightarrow$ 0. The prompt used with GPT-4o-search is shown in Figure 7.

B.4 Manual Enhancement of Automatic Annotation

Manual verification of the automatic annotation was performed using reliable sources identified through Google search. To validate the automatic labels and assess class balance, we compared automatic and manual annotations on randomly selected claims: 100 for Llama 3B Instruct and 76 for Falcon 3B Base. The resulting class distributions are shown in Figure 8(a) and Figure 9(a), with corresponding faithfulness comparisons in Figure 10(a, b).

Because false and unverifiable categories are particularly difficult to assess automatically, these cases were prioritized for manual review. Enhanced annotations for the same 100 (Llama 3B Instruct) and 76 (Falcon 3B Base) claims appear in Figure 8(b) and Figure 9(b). We additionally conducted a full manual re-check of all False and Unverifiable claims, yielding 359 manually reviewed claims for Llama 3B Instruct and 240 for Falcon 3B Base.

Six student annotators contributed to the study, each spending about three hours on the task. Instructions were delivered informally through oral discussion, with no written guidelines. All annotators volunteered and received no financial compensation.

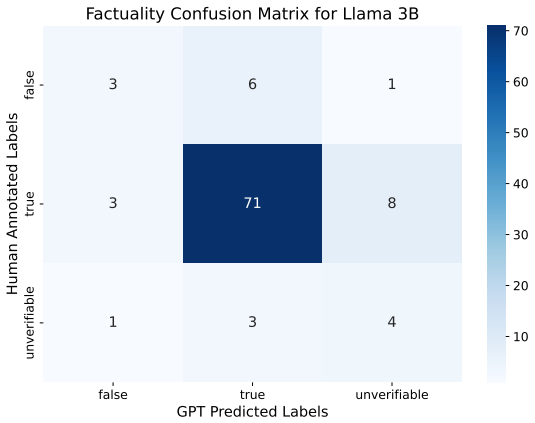
To evaluate annotation consistency, we ran an agreement analysis on the 100 Llama 3B Instruct claims, each independently reviewed by two annotators (see Table 10). The results indicate generally strong alignment across annotators, particularly for factuality, while highlighting some ambiguity in borderline cases.

Annotation Type	Num of Claims	Accuracy	Cohen's Kappa
Factuality	100	.87	.552
Faithfulness	100	.78	.586

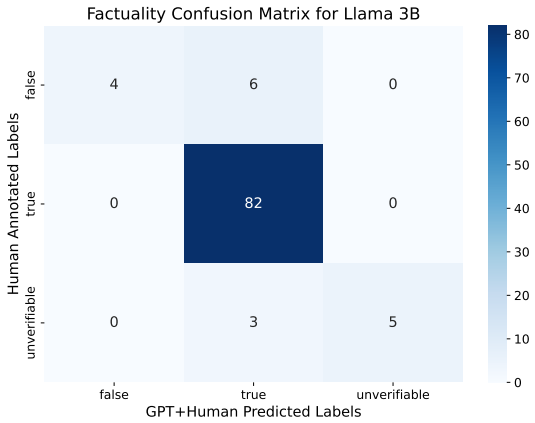
Table 10: Inter-annotator agreement for factuality and faithfulness annotations based on 100 claims of Llama 3B Instruct. Accuracy measures raw agreement, Cohen's Kappa adjusts for chance agreement.

```
Evaluate the given claim using two criteria: **faithfulness** and **factuality**.
- **Faithfulness** assesses how accurately the claim reflects the *context document*. Assign one of the following labels:
- "faithful" — The claim is directly supported by the context.
- "unfaithful-contra" — The claim directly contradicts the context.
- "unfaithful-neutral" — The claim is not present in or supported by the context.
- **Factuality** assesses the truth of the claim *independently of the context*, based on the most up-to-date and reliable sources of knowledge available to humanity. Assign one of the following labels:
- "True" — The claim is factually correct.
- "False" — The claim is factually incorrect.
- "unverifiable" — The truth of the claim cannot be determined with current knowledge.
Return your answer in the exact format: ("faithfulness label", "factuality label")
Context Document: {retrievals}
Claim: {claim}
Label:
```

Figure 7: Prompt used with GPT-4o-search to automatically annotate claims for faithfulness and factuality in long-form QA benchmark.

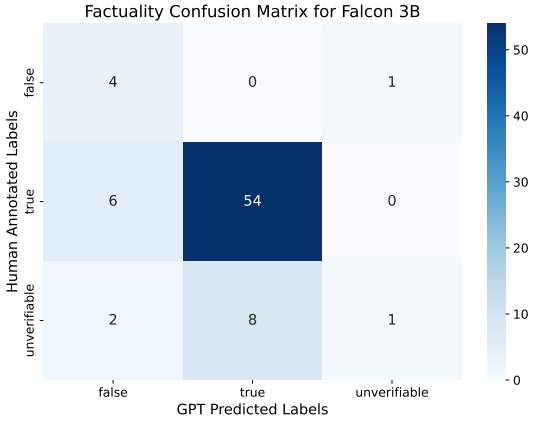


(a) Before manual enhancement of automatic annotation

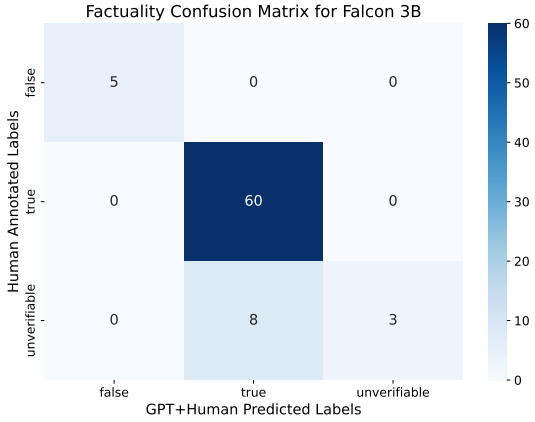


(b) After manual enhancement of automatic annotation

Figure 8: Balance of classes of factuality annotations for the Llama 3B Instruct model. Each matrix is based on 100 randomly selected claims, comparing annotations produced by the model with those from human annotators.

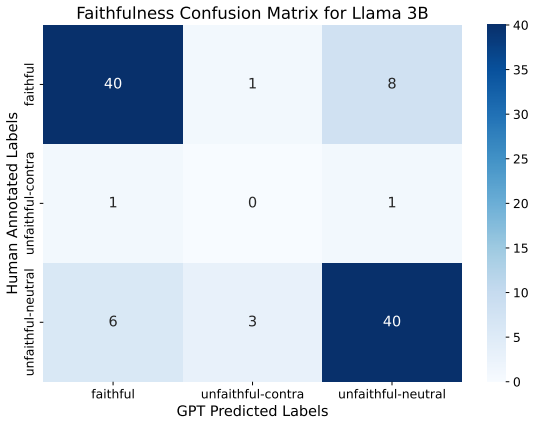


(a) Before manual enhancement of automatic annotation

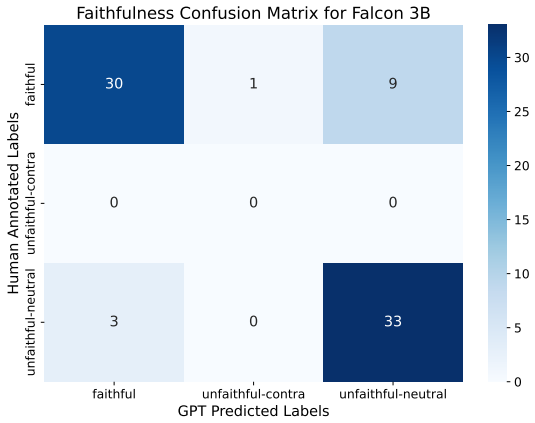


(b) After manual enhancement of automatic annotation

Figure 9: Balance of classes of factuality annotations for the Falcon 3B Base model. Each matrix is based on 76 randomly selected claims, comparing annotations produced by the model with those from human annotators.



(a) Llama 3B Instruct faithfulness classes



(b) Falcon 3B Base faithfulness classes

Figure 10: Balance of classes of faithfulness annotations for Llama 3B Instruct and Falcon 3B Base models. The matrices are based on 100 and 76 randomly selected claims, correspondingly, comparing annotations produced by the model with those from human annotators.

C Additional Faithfulness-Related Analysis

In this section, we provide additional analysis of the behavior and effectiveness of AlignScore as a faithfulness estimator within the FRANQ framework.

C.1 Faithfulness Distribution and Calibration

We examine the empirical behavior of AlignScore when used to estimate claim-level faithfulness. Figure 12 shows the distribution of AlignScore values computed between model-generated claims and their corresponding retrieved documents on the long-form QA benchmark, for two representative models.

We observe that a substantial fraction of claims receive intermediate faithfulness scores, reflecting cases where claims are only partially supported or rely on implicit inferences from the retrieved evidence. Across both models, more than 40% of claims fall within the range $[0.1, 0.9]$.

AlignScore also demonstrates strong calibration with respect to gold faithfulness labels, achieving low expected calibration error (ECE $= 0.05$). This indicates that AlignScore provides a reliable continuous estimate of faithfulness suitable for probabilistic combination in FRANQ equation 2.

Figure 11 provides a representative qualitative example illustrating how intermediate faithfulness values arise in practice.

C.2 Faithfulness Evaluation on Long-Form QA

We next evaluate the effectiveness of AlignScore as a faithfulness estimator on the long-form QA benchmark. Table 11a reports performance when faithfulness is treated as the target metric. All methods follow the same experimental setup used for the factuality evaluation.

Among the compared approaches, AlignScore achieves the strongest performance across metrics, indicating its effectiveness in approximating claim-level faithfulness within the FRANQ decomposition.

Question: *How and when to harvest chestnuts?*

Retrieved passage (excerpt): "When to harvest chestnuts? Chestnuts don't all ripen at once. Harvest typically spans up to five weeks, but most nuts ripen within a 10–30 day period in late August and September."

Model-generated claim: "The best time to harvest chestnuts is during the 10–30 day ripening window."

AlignScore: 0.61

Analysis: While the retrieved passage mentions a 10–30 day ripening period, it does not specify that this window constitutes the best time for harvesting. Accordingly, AlignScore assigns this claim an intermediate faithfulness score of 0.61, reflecting partial grounding.

Figure 11: Example illustrating intermediate AlignScore values arising from partial grounding between a claim and retrieved evidence.

C.3 Factuality Under Faithful and Unfaithful Conditions

We further analyze factuality estimation under faithful and unfaithful conditions. Table 11b reports results for unsupervised methods when restricting evaluation to unfaithful claims only. In this setting, methods leveraging parametric knowledge perform best, achieving the highest AUROC and PRR scores.

Tables 12 report results averaged across four QA datasets for Llama 3B Instruct, considering only claims with high and low AlignScore, respectively. For faithful claims, Semantic Entropy achieves the best performance, whereas for unfaithful claims, the sum of eigenvalues of the Graph Laplacian performs best. These results further motivate the use of different uncertainty estimators conditioned on faithfulness within FRANQ.

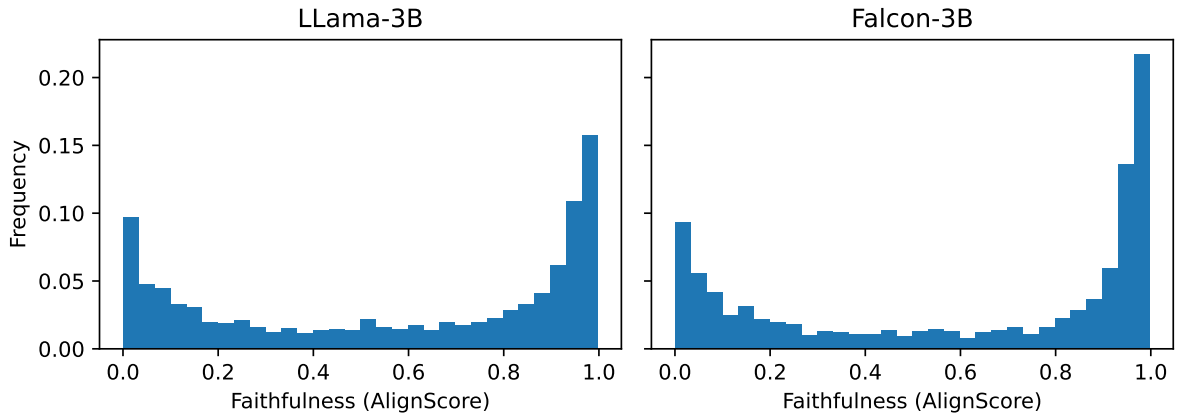


Figure 12: Distribution of AlignScore-based faithfulness estimates on the long-form QA benchmark for Llama 3B Instruct and Falcon 3B Base. A substantial mass lies in the intermediate region, and low ECE values indicate good calibration.

Method	Llama 3B Instruct AUROC ↑	PR-AUC ↑	PRR ↑
General Baselines			
Max Claim Prob.	.614	.751	.298
P(True)	.447	.624	-.242
Perplexity	.642	.782	.315
Mean Token Entropy	.596	.743	.208
CCP	.569	.727	.135
RAG-Specific Baselines			
AlignScore	**.856**	**.907**	**.789**
Parametric Knowledge	.273	.559	-.704

(a) Results on long-form QA benchmark with faithfulness target.

Method	Llama 3B Instruct AUROC ↑	PR-AUC ↑	PRR ↑
General Baselines			
Max Claim Prob.	.538	.115	.028
P(True)	.463	.112	.002
Perplexity	.480	.092	-.068
Mean Token Entropy	.580	.167	.122
CCP	.585	.134	.152
RAG-specific Baselines			
AlignScore	.477	.094	-.007
Parametric Knowledge	**.667**	**.190**	**.303**

(b) Results on long-form QA benchmark with factuality target (only unfaithful claims).

Table 11: Additional faithfulness-related results

Method	Llama 3B Instruct AUROC ↑	PR-AUC ↑	PRR ↑
General Baselines			
Max Sequence Prob.	.754	.518	.454
Mean Token Entropy	.767	.540	.472
CCP	.742	.512	.434
Lexical Similarity	.758	.500	.457
Degree Matrix	.770	.549	.488
Sum of Eigenvalues	.767	.538	.476
Semantic Entropy	**.781**	**.562**	**.510**
SentenceSAR	.766	.518	.473
RAG-Specific Baselines			
AlignScore	.606	.321	.170
Parametric Knowledge	.657	.413	.295

(a) Only claims with AlignScore > 0.5

Method	Llama 3B Instruct AUROC ↑	PR-AUC ↑	PRR ↑
General Baselines			
Max Sequence Prob.	.752	.648	.446
Mean Token Entropy	.755	.673	.462
CCP	.741	.631	.445
Lexical Similarity	.767	.662	.469
Degree Matrix	.796	.728	.551
Sum of Eigenvalues	**.807**	**.735**	**.560**
Semantic Entropy	.782	.689	.502
SentenceSAR	.770	.667	.473
RAG-Specific Baselines			
AlignScore	.555	.488	.142
Parametric Knowledge	.602	.512	.230

(b) Only claims with AlignScore < 0.5

Table 12: Results averaged across 4 QA datasets for Llama 3B Instruct considering only claims with high and low AlignScore.

D Ablation Studies

D.1 FRANQ with Alternative Faithfulness Estimators

Table 13 compares the performance of three original FRANQ versions (each employing a different calibration strategy) with three modified versions that use a thresholded AlignScore instead of raw AlignScore probabilities. In the thresholded versions, the faithfulness probability is defined as $P(c \text{ is faithful to } \mathbf{r}) = \mathbb{1}\left(\text{AlignScore}(c) > T\right)$ with $T = 0.5$. These methods are denoted by the 'T=0.5' label. The results indicate that, overall, the continuous versions of FRANQ outperform their thresholded counterparts.

Table 14 further compares the performance of three original FRANQ versions with a condition-calibrated version of FRANQ that also calibrates AlignScore for faithfulness estimation (this method is denoted 'FRANQ condition-calibrated, faithfulness-calibrated'). In this version, the AlignScore is calibrated using a training set with binary gold faithfulness targets and then incorporated into the FRANQ formula. The results suggest that calibrating AlignScore may reduce the PRR of FRANQ, indicating that it might be more effective to use AlignScore without faithfulness calibration.

D.2 Impact of Train Size on FRANQ

Figure 13 shows the PRR for 3 FRANQ variants and 2 XGBoost variants, evaluated across varying training set sizes on both long-form and short-form QA datasets. The uncalibrated FRANQ, being unsupervised, exhibits constant performance regardless of training size. In contrast, the supervised FRANQ variants generally improve with larger training sets, except for the condition-calibrated FRANQ on the long-form QA dataset, which peaks at 300 training samples and slightly declines thereafter. Across all training sizes, calibrated versions of FRANQ consistently outperform XGBoost. The results indicate that the optimal training size for condition-calibrated FRANQ is approximately 300 for long-form QA, while for short-form QA, its performance stabilizes at around 120 training samples.

D.3 Analysis of XGBoost

We examine the first tree from an XGBoost model trained on FRANQ features (AlignScore, Claim Probability, and Parametric Knowledge) for long-form QA with Llama 3B Instruct. While XGBoost uses multiple trees, the first tree often captures key decision patterns.

Figure 14 presents the first several nodes in first XGBoost tree. The root splits on AlignScore. If it's high, the model next considers Claim Probability; if low, it turns to Parametric Knowledge. This mirrors FRANQ's logic: leading with faithfulness assessment with AlignScore, followed by either Claim Probability or Parametric Knowledge. The tree thus exhibits structure similar of FRANQ's decision process.

D.4 Calibration Properties of UQ Methods

We evaluate the calibration properties of all our UQ methods using the Expected Calibration Error (ECE; Guo et al., 2017). ECE quantifies the alignment between predicted confidence scores and observed accuracy. Specifically, predictions are partitioned into 10 equally spaced confidence bins. Within each bin, we compute the average predicted confidence and compare it to the empirical accuracy. Lower ECE values indicate better-calibrated models.

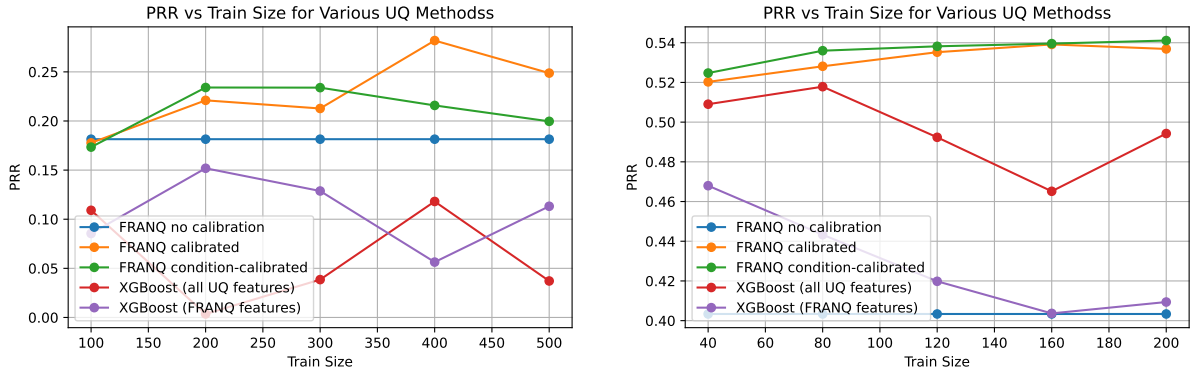
Table 15 reports ECE scores for both long-form QA dataset and short-form QA benchmark using the Llama 3B Instruct model. Only UQ methods that produce confidence values within the [0, 1] interval are included, as this is a prerequisite for ECE computation. Notably, the two calibrated variants of FRANQ achieve the best calibration performance across datasets.

Method	Llama 3B Instruct, long-form QA			Llama 3B Instruct mean across 4 short-form QA		
	AUROC ↑	PR-AUC ↑	PRR ↑	AUROC ↑	PR-AUC ↑	PRR ↑
FRANQ no calibration	.646	.100	.181	.646	.100	.181
FRANQ no calibration T=0.5	.629	.105	.170	.629	.105	.170
FRANQ calibrated	**.653**	.103	**.256**	**.653**	.103	**.256**
FRANQ calibrated T=0.5	.607	.085	.084	.607	.085	.084
FRANQ condition-calibrated	.641	**.140**	.223	.641	**.140**	.223
FRANQ condition-calibrated T=0.5	.587	.111	.180	.587	.111	.180

Table 13: Comparison of FRANQ performance on Llama 3B Instruct benchmarks, when using AlignScore with and without threshold.

Method	Llama 3B Instruct, long-form QA		
	AUROC ↑	PR-AUC ↑	PRR ↑
FRANQ no calibration	.646	.100	.181
FRANQ calibrated	**.653**	.103	**.256**
FRANQ condition-calibrated	.641	**.140**	.223
FRANQ condition-, faithfulness-calibrated	.587	.124	.112

Table 14: Comparison of FRANQ performance on Llama 3B Instruct long-form QA benchmark, when applying calibration for faithfulness estimator, AlignScore.



(a) Long-form QA, Llama 3B Instruct

(b) Short-form QA, Llama 3B Instruct

Figure 13: PRR comparison of FRANQ and XGBoost methods with different train size.

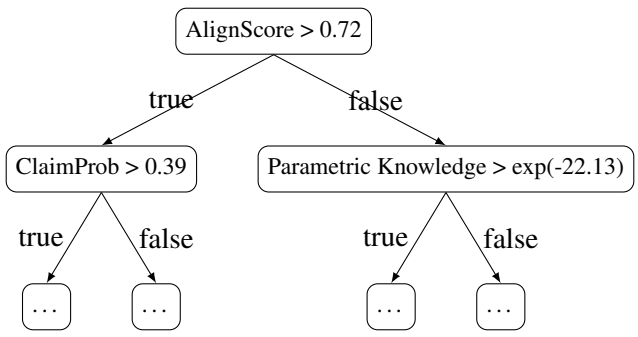


Figure 14: Top vertices of first XGBoost tree trained on FRANQ components (ClaimProb) for long-form QA Llama 3B Instruct behchmark.

Method	ECE ↓
General Baselines	
Max Claim Prob.	.72
P(True)	.94
Perplexity	.18
CCP	.21
RAG-Specific Baselines	
AlignScore	.40
Parametric Knowledge	.80
XGBoost	
XGBoost (all UQ features)	.05
XGBoost (FRANQ features)	.06
FRANQ	
FRANQ no calibration	.44
FRANQ calibrated	**.02**
FRANQ condition-calibrated	.03

(a) Long-form QA Llama 3B Instruct dataset.

Method	Mean ECE ↓
General Baselines	
Max Sequence Prob.	.46
Lexical Similarity	**.07**
Degree Matrix	.14
Sum of Eigenvalues	.54
CCP	.23
RAG-Specific Baselines	
AlignScore	.13
Parametric Knowledge	.23
XGBoost	
XGBoost (all UQ features)	.15
XGBoost (FRANQ features)	.17
FRANQ	
FRANQ no calibration	.64
FRANQ calibrated	**.07**
FRANQ condition-calibrated	**.07**

(b) Short-form QA Llama 3B Instruct benchmark (ECE is averaged across 4 QA datasets).

Table 15: Expected Calibration Error (ECE) for all tested UQ methods with Llama 3B Instruct.

Method	Inference Mean Runtime	Training Time	Model Size
Max Claim Probability	< 0.1 s	—	—
P(True)	1.3 s	—	—
Perplexity	< 0.1 s	—	—
Max Token Entropy	< 0.1 s	—	—
CCP	1.7 s	—	—
AlignScore	0.5 s	—	—
Parametric Knowledge	1.6 s	—	—
XGBoost (all UQ features)	1.9 s	0.60 s	10 kB
XGBoost (FRANQ features)	1.7 s	0.12 s	14 kB
FRANQ (no calibration)	1.7 s	—	—
FRANQ (calibrated)	1.7 s	0.57 s	312 B
FRANQ (condition-calibrated)	1.7 s	0.58 s	244 B

Table 16: Inference runtime, training time, and model size for uncertainty estimators and FRANQ variants, measured on Llama 3B Instruct for short-form QA. All runtimes represent overhead beyond base LLM generation.

E Runtime Analysis

In Table 16 we present the runtime overhead, training cost, and model size of FRANQ's uncertainty components. All runtimes are measured as incremental cost beyond a completed LLM forward pass using Llama 3B Instruct on short-form NQ dataset (average generation time: 2.2 s per instance).

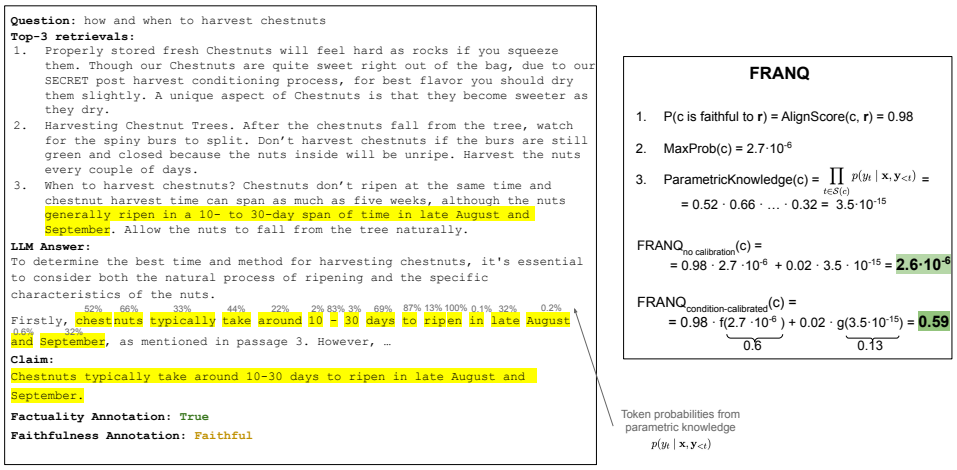
Overall, FRANQ incurs modest overhead, while isotonic calibration adds negligible cost and yields compact models (hundreds of bytes), supporting production deployment.

F Resource and Expenses

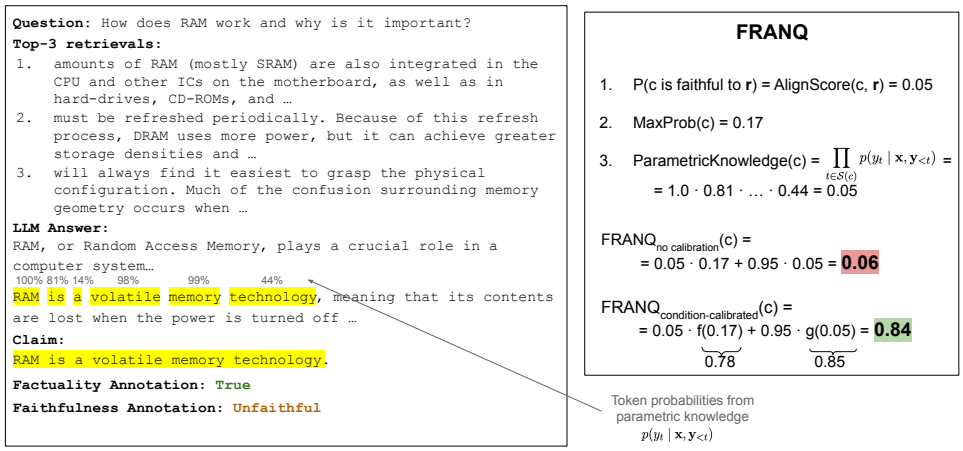
A full data-generation and UQ-baseline evaluation run required about 8 days of compute on an NVIDIA V100 32GB GPU for long-form QA, while short-form QA needed under one day. The OpenAI API was used for claim splitting, matching, and annotation, costing roughly $100 per model run (Llama 3B Instruct). Human annotation involved six student annotators, each contributing about three hours of work.

G FRANQ Examples

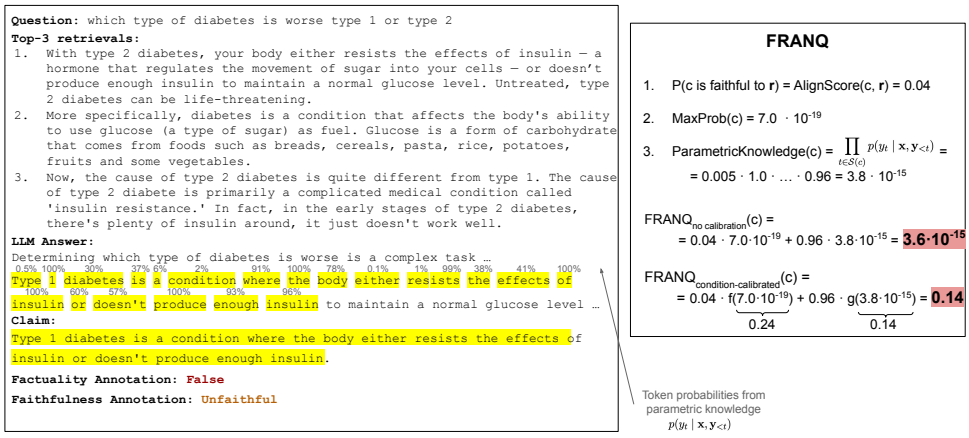
In Figure 15, we demonstrate the behavior of FRANQ using three examples from a long-form QA dataset evaluated with Llama 3B Instruct. We selected three representative claims and present their corresponding FRANQ scores for both the uncalibrated version and condition-calibrated version. The latest uses monotonic functions f and g, fitted via isotonic regression for Claim Probability and Parametric Knowledge methods, respectively.



(a) **Faithful–True.** FRANQ correctly identifies the claim as faithful and uses Claim Probability, which detects high entailment with the third retrieved passage. This results in an appropriately high FRANQ score.



(b) **Unfaithful–True.** FRANQ accurately detects the claim's low faithfulness and assigns its factuality score based on Parametric Knowledge, which is relatively high. In the uncalibrated version, the final score is underestimated due to the uncalibrated Parametric Knowledge score. The condition-calibrated version corrects this by assigning a calibrated score of 0.85, resulting in a correctly high factuality estimate.



(c) **Unfaithful–False.** FRANQ correctly identifies the claim as unfaithful and assigns a low factuality score using Parametric Knowledge, consistent across both the uncalibrated and calibrated versions.

Figure 15: Example outputs from FRANQ. *Left*: Each example includes the input question, retrieved passages, the LLM-generated answer, a selected claim from the answer, and corresponding factuality and faithfulness annotations. Claims and their spans in the answer are highlighted in yellow. If a claim is faithful, its corresponding span in the retrieved passages is also highlighted. *Right*: The FRANQ component scores and final factuality estimations, shown for both the uncalibrated and condition-calibrated versions.