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

Anonymous ACL submission

Abstract

Large Language Models (LLMs) enhanced with external knowledge retrieval, an approach known as Retrieval-Augmented Generation (RAG), have achieved strong performance in open-domain question answering. However, RAG systems remain prone to hallucinations: factually incorrect outputs that may arise from inconsistencies between 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 (Faithfulness-aware Retrieval-Augmented UNcertainty Quantification), a new method for hallucination detection in RAG outputs. FRANQ applies distinct uncertainty quantification techniques to estimate factuality, conditioning on whether a statement is faithful to the retrieved context. To evaluate FRANQ and competing uncertainty estimation 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 uncertainty quantification and hallucination detection approaches.

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 hallucinations, arising from factors such as insufficient training data coverage, input ambiguity, and architectural constraints (Huang et al., 2025). To address these limitations, Retrieval-Augmented Generation (RAG; Lewis et al., 2020) integrates dynamically retrieved external knowledge into the generation process. This external retrieval has been shown to partially mitigate factual inaccuracies (Shuster et al., 2021).

However, RAG-based 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 a RAG system, 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). In the context of fact-checking of RAG output, detecting non-factual claims is more critical than identifying unfaithful ones. Distinguishing faithfulness from factuality allows us

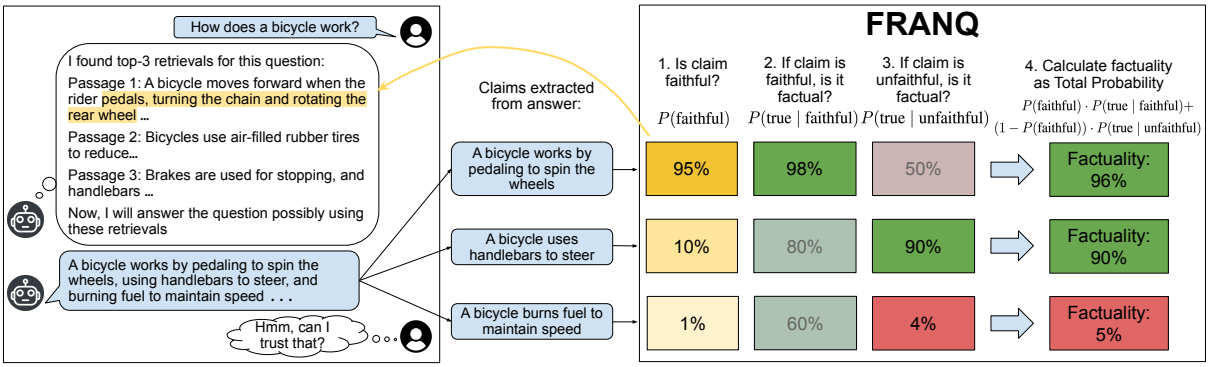


Figure 1: FRANQ illustration. *Left*: A user poses a question, and the RAG retrieves relevant documents and formulates an answer, potentially using information from the retrieved documents. *Middle*: The RAG output is decomposed into atomic claims. *Right*: The FRANQ method assesses factuality by evaluating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

to disentangle two fundamental failure modes in the RAG system: (*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).

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 effectively 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-form and short-form QA tasks. For long-form question answering (QA), where answers include multiple claims, we assess factuality at the claim level. For that, we introduce a dataset with factuality annotations, combining automated labeling with manual validation of challenging examples. For short-form QA, where answers are brief, we treat each response as a single claim and evaluate it as a whole. We test our method on four QA datasets in this setting.

Our key **contributions** are as follows.

- We develop a new UQ method for RAG, FRANQ, that estimates uncertainty by first assessing faithfulness, and then enabling separate 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 build by combining automatic annotation with manual validation for difficult cases; see Section 3.
- We conduct comprehensive experimental validation on both long-form and short-form QA across 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 first 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 model 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 $y_t \sim p(\cdot \mid \mathbf{y}_{<t})$, 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., $y_t = \arg\max_y p(y \mid \mathbf{y}_{<t})$. 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

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

Table 1: Summary of general baseline methods.

main objective is to assess the *factuality* of claims.

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

For example, a standard UQ baseline is the probability of a claim c according to the LLM:

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

We experiment with UQ methods implemented in the LM-Polygraph library (Fadeeva et al., 2023), which we summarize in Table 1.

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. The overall factuality of a claim c is estimated as:

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

Here, we calculate $P(c \text{ is unfaithful to } \mathbf{r})$ as $1 - P(c \text{ is faithful to } \mathbf{r})$. This decomposition integrates three probability components, each instantiated using a suitable UQ method:

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 our approach is visually depicted in Figure 1. Detailed examples of FRANQ applied to specific claims are given in Appendix E.

2.2 FRANQ Components

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

Faithfulness. To assess whether a claim c_i is entailed by the retrieved evidence $\mathbf{r}$, we use *AlignScore*, a semantic similarity metric based on RoBERTa fine-tuned for factual alignment (Zha et al., 2023). AlignScore evaluates the consistency between a claim and its supporting context. A comparison with alternative methods is provided in Appendix C.1.

In long-form QA, we compare the claim c_i with the concatenated retrieved documents $\mathbf{r}$ using AlignScore. For short-form QA, where the full answer $\mathbf{y}$ is treated as a single claim, we provide question context by calculating AlignScore on pairs of the form $(\mathbf{x} \circ \mathbf{y}, r_j)$, with '$\circ$' denoting string concatenation.

Factuality under unfaithful condition. When a claim c is unfaithful (i.e., not entailed by $\mathbf{r}$), it likely originates from the LLM's internal knowledge. We estimate its factuality using the model's own probability estimates. Specifically, we introduce the *Parametric Knowledge* UQ method, which computes the likelihood of c based solely on the LLM's parametric knowledge (Mallen et al., 2023) without providing retrieved evidence $\mathbf{r}$:

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

Note that this method does not require generating a new response; instead, it reuses the original tokens, but performs a new forward pass through the LLM without the retrieved evidence $\mathbf{r}$ in the input. This approach is effective in the long-form QA setting (see Appendix C.2). However, for short-form QA, more alternative general UQ baselines are available, such as the ones employing sample diversity. We find that *Sum of Eigenvalues* (Lin et al., 2024) better approximates factuality in this scenario (see Appendix C.2).

Therefore, we estimate the factuality of unfaithful claims using Parametric Knowledge in long-form QA and using Sum of Eigenvalues in short-form QA.

Factuality under faithful condition. To handle residual errors, for instance, inaccuracies in assessing factuality, we estimate uncertainty within the faithful branch by reverting to a simple probability baseline $p(c \mid \mathbf{x}, \mathbf{r})$. Our ablation studies in Appendix D.1 further demonstrate that the choice of the method in this branch has little impact; even a constant baseline performs well.

Again, for short-form QA, we find alternative baselines more suitable, particularly *Semantic Entropy* (Kuhn et al., 2023), which better captures uncertainty in this scenario (see Appendix C.2).

Therefore, we estimate the factuality for faithful claims with 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:

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

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-form QA and short-form QA scenarios. For long-form QA, we use 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 $\text{UQ}_{\text{unfaith}}$.

Note that we consistently apply the same uncertainty methods across all datasets within each QA setting (short-form and long-form). We select uncertainty techniques based on the nature of the task (using token-level likelihoods for long-form QA and sampling-based metrics for short-form QA) and do not tune any methods per dataset.

2.3 Calibrating FRANQ

Since the UQ methods UQ_{faith} and $\text{UQ}_{\text{unfaith}}$ of equation (4) are not inherently probabilistic and 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 a 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 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:

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

where $\mathcal{F}$ denotes the set of all non-decreasing functions mapping real numbers to probabilities in the interval $[0, 1]$. During inference, we apply the calibration function $\widehat{f}$ to each UQ score to produce probabilistically meaningful output.

Since UQ_{faith} and $\text{UQ}_{\text{unfaith}}$ represent probabilities 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 faithful subset $\mathcal{D}_{\text{faith}}$ and $\text{UQ}_{\text{unfaith}}$ using the unfaithful 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 scores.** Both UQ methods are calibrated on the entire training dataset $\mathcal{D}$, disregarding claim faithfulness.

3. **Condition-calibrated scores.** 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 (see the discussion in Section 5.2). We argue that an effective dataset should reflect generation behavior in the presence of retrieval, capturing both factual errors and contextual misuse. To this end, we introduce a new dataset specifically designed for long-form generations, enabling fine-grained analysis of atomic claims.

Overall, we consider two QA settings: *long-form QA*, where the questions require comprehensive answers composed of multiple claims;

and *short-form QA*, where the questions are answered concisely with single claims. For both QA settings, we use `Llama-3.2-3B-Instruct`, `Llama-3.1-8B-Instruct` (Grattafiori et al., 2024) and `Falcon3-3B-Base` (Team, 2024) as the LLMs. For long-form QA, we also use `Gemma-4B-Instruct` and for short-form QA we use `Gemma-12B-Instruct` (Team et al., 2025).

Long-form QA dataset. Generating high-quality data for RAG hallucination detection remains challenging. Most benchmarks focus on faithfulness (i.e., entailment given the retrieved content) rather than factuality. However, we argue that factuality is more critical in RAG applications, with faithfulness serving as a complementary perspective. Since automatic annotation lacks precision and manual annotation is costly, we adopt a hybrid approach to ensure data quality.

Our long-form QA dataset consists of 76 questions: 44 from RAGTruth (Niu et al., 2024) filtered to include questions where Llama 3B Instruct provides at least two false claims, and 32 technical questions generated via GPT-4 (e.g., *"How does solar power generate electricity?"*).

We extract the claims and their corresponding token spans from the model outputs using the methods introduced in (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 training, reserving the remainder for testing.

We annotate factuality and faithfulness using GPT-4o-search and manually verify the claims that were classified as incorrect. For faithfulness annotation, each claim is automatically assigned one of three categories: *faithful*, *unfaithful-contra*, or *unfaithful-neutral*. These categories are binarized for evaluation: faithful is mapped to 1, and both unfaithful categories to 0, as unfaithful-contra comprises less than 5% of the data. Similarly, factuality annotations include *True*, *False*, and *unverifiable*. We only retain the verifiable claims (True or False), with binarized labels. Further details about the prompts and statistics about the datasets are given in Appendix B.

Short-form QA datasets. For short-form QA, we evaluate FRANQ on four widely used QA datasets: TriviaQA (Joshi et al., 2017), SimpleQA (Wei et al., 2024a), Natural Questions (Kwiatkowski et al., 2019), and PopQA (Mallen et al., 2023). This time, we treat the output as a single claim. From each dataset, we sample 200 questions for training and 1,000 for testing.

We evaluate the factuality by comparing the RAG outputs to the gold-standard answers using GPT-4, following the approach described in (Wei et al., 2024a).

RAG models. For each question, we select the top-k passages using the `Facebook/contriever-msmarco` retriever and Wikipedia embeddings based on the 2018 English Wikipedia, to ensure accurate retrievals. For long-form QA, we set $k = 3$, while for short-form QA we use $k = 5$. The choice reflects a trade-off between recall and noise: long-form QA benefits from more focused contexts, whereas short-form QA gains from broader evidence coverage.

4 Experiments

For all experiments, we fix both the retrieval process and the white-box LLM, verifying the factual accuracy of the LLM-generated claims.

We test FRANQ with three calibration strategies (see Section 2.3). First, we compare it to the unsupervised baselines introduced earlier. While the uncalibrated version of FRANQ is unsupervised, the calibrated versions require labeled training data. Therefore, we further compare our method against a supervised XGBoost baseline (Kuleshov and Liang, 2015) trained using the outputs of unsupervised UQ methods as features.

Furthermore, in ablation studies, we investigate the impact of different FRANQ components $P(\text{faithful})$, UQ_{faith}, and $\text{UQ}_{\text{unfaith}}$, as well as varying training sizes.

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. To implement the baselines, we use the LM-Polygraph library (Fadeeva et al.,

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 ↑
Max Claim Probability	.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
AlignScore	.075	.108	.104	.233	.068	.119	.061	.058
Parametric Knowledge	.064	.018	.067	.029	.059	.047	.112	.183
XGBoost (all UQ features)	.124	.206	.088	.198	.044	-.015	.073	.085
XGBoost (FRANQ features)	.111	.149	.080	.086	.048	.017	.090	.158
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.

2023). For short-form QA, we use 13 sequence-level methods; for long-form QA, we use 5 claim-level methods. A complete list of these methods is provided in Table 1.

RAG-specific baselines. We also evaluate three baseline methods, which we previously introduced as components of FRANQ: AlignScore, Claim Probability, and Parametric Knowledge (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}, $UQ_{unfaith}$), and (2) all available unsupervised UQ methods (for long-form QA: 5 general + 3 RAG-specific baselines; for short-form QA: 13 general + 3 RAG-specific baselines).

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

Evaluation measures. Each UQ method provides factuality estimates, which we compare to the binary gold-standard factuality labels using PR-AUC (the false claims are designated as the positive class to emphasize the method's effectiveness in identifying them). To further check rejection performance, we calculate the Prediction Rejection Ratio (PRR; Mallen et al., 2023) with a maximum rejection threshold of 0.5. It quantifies how well the model can reject uncertain predictions while retaining accurate ones, thus providing a direct measure of the model's ability to identify and prioritize reliable outputs.

4.2 Long-Form QA Results

For long-form QA, we evaluate each UQ method using two measures (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 variant 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 compute two evaluation measures (PR-AUC and PRR) for each UQ method across four models (Llama 3B Instruct, Llama 8B Instruct, Falcon 3B Base, and Gemma 12B Instruct) and four QA datasets. For each measure and dataset, we determine the relative rank of each method (1st, 2nd, etc.). Because these detailed results vary across datasets, we report only the mean values and mean ranks averaged over the four datasets for clarity, following Vashurin et al. (2025). Summary results are shown in Table 3, with full per-dataset results in Appendix A.

Across all models, the condition-calibrated FRANQ outperforms baselines, achieving the best mean rank and mean value on both measures, except for mean PRR on Gemma 12B Instruct, where it ranks second. The calibrated FRANQ variant

Method	Llama 3B Instruct				Falcon 3B Base				Llama 8B Instruct				Gemma 12B Instruct			
	MeanRank ↓		MeanValue ↑		MeanRank ↓		MeanValue ↑		MeanRank ↓		MeanValue ↑		MeanRank ↓		MeanValue ↑	
	PR-AUC	PRR	PR-AUC	PRR	PR-AUC	PRR	PR-AUC	PRR	PR-AUC	PRR	PR-AUC	PRR	PR-AUC	PRR	PR-AUC	PRR
Max Sequence Probability	12.50	11.50	.558	.454	12.50	14.00	.628	.256	12.50	13.00	.569	.407	7.75	17.50	.400	.162
CCP	13.00	13.00	.551	.443	10.50	10.50	.641	.304	14.50	14.00	.553	.417	14.75	14.75	.412	.198
Mean Token Entropy	6.75	8.00	.594	.481	14.25	15.00	.613	.242	4.50	8.00	.640	.491	14.25	12.50	.423	.230
P(True)	20.00	20.00	.292	-.011	20.00	20.00	.529	.042	20.00	20.00	.365	.083	13.75	14.75	.434	.191
Lexical Similarity	10.50	8.25	.564	.479	14.00	13.50	.618	.277	4.00	2.50	.639	.532	14.00	12.00	.430	.240
Degree Matrix	3.75	5.75	.629	.520	**3.00**	3.25	.702	.464	7.00	7.00	.627	.492	8.50	6.50	.464	.260
Eccentricity	5.00	6.50	.615	.515	8.75	7.25	.662	.406	9.00	8.00	.613	.485	10.50	8.00	.451	.256
Sum of Eigenvalues	4.25	5.50	.628	.518	3.50	3.75	.700	.460	6.00	8.50	.628	.489	7.75	7.25	.467	.260
Number of Semantic Sets	16.50	16.75	.472	.312	10.25	8.25	.646	.382	15.50	14.50	.477	.273	18.75	17.50	.383	.172
Semantic Entropy	5.25	4.75	.613	.525	13.25	13.25	.623	.278	4.50	3.50	.637	.519	7.25	7.00	.466	.261
Perplexity	10.75	11.75	.558	.432	15.75	16.25	.603	.225	8.50	9.50	.609	.456	20.75	19.25	.376	.140
TokenSAR	13.50	13.75	.540	.413	13.25	14.50	.615	.243	12.50	12.50	.579	.430	17.00	14.25	.413	.228
SentenceSAR	10.25	9.00	.571	.483	15.00	14.25	.602	.263	12.00	11.50	.556	.414	13.25	12.50	.416	.174
AlignScore	18.25	17.50	.415	.207	7.50	7.25	.666	.372	18.50	18.00	.432	.224	21.00	18.50	.376	.158
Parametric Knowledge	18.00	17.75	.425	.247	19.25	18.75	.556	.104	16.50	17.50	.499	.330	21.75	22.00	.364	.105
XGBoost (all UQ features)	8.50	7.25	.594	.494	3.25	3.75	.705	.462	5.50	6.50	.634	.503	5.75	6.00	.474	**.301**
XGBoost (FRANQ features)	14.75	14.00	.526	.409	7.25	8.00	.670	.368	16.00	15.50	.524	.385	14.75	15.75	.414	.196
FRANQ no calibration	12.25	14.75	.553	.403	10.00	10.75	.641	.345	16.00	15.50	.523	.340	9.75	12.75	.447	.225
FRANQ calibrated	4.00	2.75	.628	.537	7.75	6.00	.672	.411	4.50	3.00	.644	.534	6.00	9.25	.481	.258
FRANQ condition-calibrated	**3.00**	**2.25**	**.631**	**.541**	**3.00**	**2.75**	**.711**	**.477**	**3.00**	**2.00**	**.647**	**.540**	**2.75**	**5.25**	**.496**	.283

Table 3: Results averaged across four QA datasets for Llama 3B Instruct, Falcon 3B Base, Llama 8B Instruct and Gemma 12B Instruct. For each model, the first two columns show the mean rank across all 4 datasets, and the last two columns report the mean metric values. Lower mean ranks and higher mean values indicate better performance. The condition-calibrated FRANQ is top-performing across all settings, except mean PRR on Gemma 12B Instruct, where it ranks second.

ranks second in both mean rank and mean value for PRR on Llama 3B Instruct and in mean value for both PR-AUC and PRR on Llama 8B Instruct. Among the unsupervised methods, Degree Matrix and Lexical Similarity showed strong performance, ranking second-best in PR-AUC or PRR under different scenarios.

4.4 Ablation Studies

We conducted ablation studies to evaluate the contributions of the individual FRANQ components, the effect of the training set size on the supervised variants, and the calibration quality of the tested UQ methods. In this section, we summarize the key findings, and we give a more detailed analysis in Appendix D.

Analysis of FRANQ's components. We begin by analyzing different combinations of UQ_{faith} and $UQ_{unfaith}$ within the condition-calibrated FRANQ (see Figure 10 in in Appendix D). For long-form QA with Llama 3B Instruct, the performance is highly sensitive to the choice of $UQ_{unfaith}$, with Parametric Knowledge emerging as the most effective option. In contrast, $UQ_{unfaith}$ has a relatively minor impact. For short-form QA with Llama 3B Instruct, FRANQ is robust across multiple UQ configurations, suggesting greater flexibility in component selection.

Next, we evaluate alternative faithfulness estimation strategies (see Tables 14 and 15 in Appendix D). Substituting AlignScore probabilities with binary thresholding consistently degrades the performance, underscoring the benefit of using continuous faithfulness signals. Surprisingly, calibrating AlignScore with gold-standard labels also reduces the performance, indicating that the raw AlignScore values are more effective when used directly within FRANQ.

Impact of the training set size. We also examine the impact of the training set size on both FRANQ and XGBoost variants (see Figure 11 in Appendix D). The unsupervised FRANQ baseline remains stable, while supervised versions with calibration generally improve with more data. In long-form QA, the condition-calibrated FRANQ peaks at 300 training instances before slightly declining; for short-form QA, the performance saturates around 120 instances. Notably, the supervised FRANQ methods remain highly effective even with limited data (e.g., 100–200 samples) and consistently outperform XGBoost across all training sizes.

XGBoost approximates the decision-making behavior of FRANQ. In order to better understand XGBoost's behavior, we analyzed its first decision tree trained on FRANQ features (see Figure 12 in Appendix D). We found that the tree mimics FRANQ's decision logic: initially splitting on AlignScore, followed by Claim Probability or Parametric Knowledge depending on the score. This suggests that supervised methods such as XGBoost learn decision patterns that are closely aligned with those encoded in FRANQ.

Calibration evaluation. Finally, we assess the calibration of all UQ methods using Expected Calibration Error (ECE; Guo et al., 2017); see Table 16 in Appendix D. As expected, the two calibrated FRANQ variants achieve the lowest ECE scores for both long- and short-form QA, indicating that they produce well-calibrated confidence estimates that are well-aligned with the empirical accuracy.

5 Related Work

5.1 Uncertainty Quantification for RAG

Several uncertainty quantification (UQ) methods have been proposed for RAG that assess how retrieved knowledge influences the LLM output. Chuang et al. (2024) detect hallucinations using a linear classifier over the lookback ratio, measuring attention weights on the context versus generated tokens. Sun et al. (2025) regress features reflecting the model's reliance on retrieved and parametric knowledge. Li et al. (2024) estimate uncertainty via the signal-to-noise ratio of output probabilities across samples, while Hu et al. (2024) compute a relevance matrix between the prompt and response to classify hallucinations.

All these approaches identify hallucinations only with respect to the retrieved context, without verifying against world knowledge, and add computational overhead through auxiliary models and training. Search-based verification methods such as SAFE (Wei et al., 2024b) instead employ an LLM agent for claim decomposition and web-based verification, introducing dependencies on external services and multi-step reasoning. In contrast, FRANQ offers a lightweight, self-contained UQ framework that separates uncertainty from retrieved and internal knowledge. It models factuality as a probabilistic combination of (1) faithfulness to retrieved context and (2) truthfulness under both faithful and unfaithful conditions, without extra training or external verification.

When retrieval is omitted, uncertainty is often estimated solely 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 bridges these perspectives by jointly modeling retrieval-related and intrinsic uncertainty within a single probabilistic framework, enabling consistent evaluation across both faithful (retrieval-grounded) and unfaithful (parametric) generations.

5.2 Hallucination and Factuality Datasets for RAG

Hallucination detection in RAG requires labeled data capturing model hallucinations through class annotations. RAGTruth (Niu et al., 2024) is a multi-domain dataset featuring manually annotated, span-level hallucinations. However, it does not account for cases where the model generates knowledge independently of the retrieved context, responses that may still be factually correct and useful to the user.

Knowledge-grounded dialogue datasets such as Wizard of Wikipedia (Dinan et al., 2019) and Faith-Dial (Dziri et al., 2022) similarly pair responses with external sources. However, their primary goal is conversational coherence, and any information not grounded in the provided context is labeled as hallucination regardless of factual correctness.

Hallucination benchmarks for RAG are often derived from QA datasets that include relevant context, such as RAGBench (Friel et al., 2024) and AdaptiveRAG (Moskvoretskii et al., 2025), where hallucinations are defined strictly relative to the given context.

By contrast, FRANQ extends this formulation by estimating factuality even when the generated content diverges from the retrieved evidence, enabling a more comprehensive view of model reliability.

6 Conclusion and Future Work

We introduced FRANQ (Faithfulness-aware Retrieval Augmented uNcertainty Quantification), 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.

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.

Volodymyr Kuleshov and Percy S Liang. 2015. Calibrated structured prediction. In Advances in Neural Information Processing Systems, volume 28. Curran Associates, Inc.

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 Lysyuk, 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. arXiv preprint arXiv:2501.12835.

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 MeanRank and MeanValue 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 4) and Falcon 3B Base (see Table 5).

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.

Method	NQ			PopQA			TriviaQA			SimpleQA		
	ROC-AUC ↑	PR-AUC↑	PRR↑	ROC-AUC ↑	PR-AUC↑	PRR↑	ROC-AUC ↑	PR-AUC↑	PRR↑	ROC-AUC ↑	PR-AUC↑	PRR↑
Max Sequence Probability	.680	.440	.292	.745	.550	.421	.774	.529	.478	.833	.712	.625
CCP	.705	.471	.357	.709	.526	.393	.767	.528	.471	.800	.680	.552
Mean Token Entropy	.723	.503	.389	.768	**.607**	.455	.796	.569	.523	.809	.697	.555
P(True)	.463	.256	-.042	.550	.374	.100	.474	.244	-.022	.419	.294	-.082
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
Eccentricity	.750	.528	**.424**	.741	.569	.420	.805	.589	.545	.852	.773	.671
Sum of Eigenvalues	.749	.553	.411	.740	.564	.416	.816	.621	.561	.861	.774	.686
Number of Semantic Sets	.680	.459	.336	.534	.374	.069	.695	.461	.382	.737	.595	.461
Semantic Entropy	.727	.518	.373	**.776**	.602	**.496**	.801	.565	.546	.863	.766	.684
Perplexity	.705	.485	.366	.745	.573	.422	.763	.522	.460	.768	.653	.480
TokenSAR	.696	.476	.359	.728	.560	.397	.745	.488	.426	.769	.635	.468
SentenceSAR	.678	.395	.269	.762	.562	.459	.794	.560	.521	.858	.767	.682
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 (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 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 4: Results on 4 QA datasets for Llama 3B Instruct.

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	NQ			PopQA			TriviaQA			SimpleQA		
	ROC-AUC ↑	PR-AUC↑	PRR↑	ROC-AUC ↑	PR-AUC↑	PRR↑	ROC-AUC ↑	PR-AUC↑	PRR↑	ROC-AUC ↑	PR-AUC↑	PRR↑
Max Sequence Probability	.599	.555	.186	.653	.649	.259	.590	.487	.163	.625	.820	.416
CCP	.632	.576	.258	.659	.648	.297	.620	.518	.212	.635	.822	.448
Mean Token Entropy	.599	.542	.184	.657	.662	.279	.557	.432	.108	.656	.814	.396
P(True)	.516	.458	.025	.515	.510	.012	.480	.383	.002	.551	.764	.129
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**
Eccentricity	.632	.536	.204	.776	.773	.569	.633	.474	.216	.767	.867	.621
Sum of Eigenvalues	.651	.568	.260	**.789**	**.780**	.570	.661	.559	.299	.791	.894	.713
Number of Semantic Sets	.622	.534	.205	.691	.670	.413	.651	.524	.282	.738	.858	.625
Semantic Entropy	.561	.494	.086	.718	.698	.415	.584	.468	.155	.685	.831	.456
Perplexity	.593	.545	.199	.626	.629	.215	.547	.428	.097	.645	.809	.388
TokenSAR	.602	.555	.210	.656	.654	.249	.553	.433	.106	.654	.818	.405
SentenceSAR	.509	.455	.012	.755	.707	.463	.523	.395	.026	.739	.850	.552
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 (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 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 5: Results on 4 QA datasets for Falcon 3B Base.

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 6, full results in Table 7). 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	MeanRank ↓		MeanValue ↑	
	PR-AUC	PRR	PR-AUC ↑	PRR ↑
Max Sequence Probability	11	10	.638	.489
CCP	12	11	.629	.472
Mean Token Entropy	9	12	.651	.46
P(True)	21	21	.339	-.117
Lexical Similarity	7	9	.668	.512
Degree Matrix	3	3	.687	.548
Eccentricity	6	5	.669	.54
Sum of Eigenvalues	4	7	.686	.536
Number of Semantic Sets	17	16	.569	.379
Semantic Entropy	8	6	.666	.537
Perplexity	13	13	.616	.416
TokenSAR	14	17	.591	.379
SentenceSAR	10	8	.648	.523
AlignScore	18	18	.507	.234
Parametric Knowledge	19	19	.502	.22
XGBoost (all UQ features)	5	4	.684	.544
XGBoost (FRANQ features)	16	14	.579	.402
FRANQ no calibration	15	15	.586	.393
FRANQ calibrated	2	2	.692	.549
FRANQ condition-calibrated	**1**	**1**	**.695**	**.553**

Table 6: Aggregated results with shuffled retrievals on 4 QA datasets for Llama 3B Instruct.

Method	MeanRank ↓		NQ		PopQA		TriviaQA		SimpleQA	
	PR-AUC	PRR	PR-AUC ↑	PRR ↑	PR-AUC ↑	PRR ↑	PR-AUC ↑	PRR ↑	PR-AUC ↑	PRR ↑
Max Sequence Probability	10.5	10	.477	.216	.645	.509	.556	.477	**.874**	.755
CCP	11.5	11.25	.494	.269	.629	.481	.571	.482	.822	.657
Mean Token Entropy	7.25	10.75	.571	.346	.678	.51	.58	.46	.777	.525
P(True)	20.75	20.75	.348	-.032	.348	-.065	.254	-.151	.404	-.22
Lexical Similarity	6	7	.58	.37	.696	.547	.579	.489	.819	.639
Degree Matrix	5	5	.57	.359	.661	.537	.645	.553	.871	.745
Eccentricity	8.25	5.25	.536	.313	.665	.548	.617	.545	.856	.753
Sum of Eigenvalues	4.25	5.75	.573	.36	.664	.516	.637	.542	.872	.728
Number of Semantic Sets	14.75	13.5	.514	.329	.468	.19	.563	.474	.731	.524
Semantic Entropy	6.75	5.75	.537	.317	**.697**	**.587**	.565	.493	.867	.752
Perplexity	12.25	13.5	.55	.309	.63	.463	.542	.423	.742	.47
TokenSAR	14.5	15	.536	.283	.596	.429	.522	.382	.711	.422
SentenceSAR	9.5	7.5	.465	.217	.695	.546	.566	.517	.866	**.811**
AlignScore	17.5	17.5	.493	.259	.506	.225	.417	.213	.613	.24
Parametric Knowledge	18.5	18.5	.406	.104	.564	.342	.494	.377	.545	.059
XGBoost (all UQ features)	6.5	5.5	.561	.354	.657	.474	**.656**	**.594**	.86	.753
XGBoost (FRANQ features)	15.5	14.75	.488	.269	.579	.376	.471	.37	.779	.592
FRANQ no calibration	15.5	15.25	.419	.127	.626	.452	.527	.417	.774	.575
FRANQ calibrated	**4**	**4.5**	.601	.401	.675	.549	.624	.529	.868	.717
FRANQ condition-calibrated	**4**	**4.5**	**.603**	**.413**	.675	.549	.636	.536	.867	.715

Table 7: Results with shuffled retrievals on 4 QA datasets for Llama 3B Instruct.

B Prompts and Setup

```
Contents (not necessarily includes answer to the following question):

Title: {title1}
Content: {retrieval1}

Title: {title2}
Content: {retrieval2}
...

Title: {title5}
Content: {retrieval5}

Question: {question}

Answer (single line):
```

Figure 2: 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 3: Prompt used in long-form QA datasets. Retrievals corresponds to the Wikipedia passage retrieved for input question.

B.1 Short-form QA

For short-form QA experiments, we used each input question along with the top-5 retrieved passages from Wikipedia. The prompt format used across all evaluated models is shown in Figure 2. For annotation, we provided GPT-4 with the input question, the model-generated answer, and the corresponding gold answer, and asked it to categorize the model's response as correct, incorrect, or not attempted (the latter category was excluded from evaluation). This annotation method follows the approach proposed by Wei et al. (2024a). Table 8 presents the statistics of all short-form QA datasets.

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.

B.2 Long-form QA

For long-form QA experiments, we used each input question along with the top-3 retrieved passages from Wikipedia. The prompt format employed across all evaluated models is illustrated in Figure 3. We then decomposed the extracted answers into atomic claims using the prompt shown in Figure 4. For each claim, we identified its corresponding span in the original sentence using the prompt in Figure 5. Claims for which we could not extract spans (due to inconsistencies in annotation, e.g., words appearing in a different order than in the text) were excluded from evaluation. Finally, we annotated the resulting claims for both factuality and faithfulness using a combination of automatic annotation and subsequent manual validation (see Appendix B.3 and B.4). Table 9 presents the statistics of all long-form QA dataset.

Compared to other decomposition operators, such as that used in FActScore (Min et al., 2023), we performed long-form QA claim decomposition more carefully. Specifically, (1) we decomposed the entire text rather than splitting each sentence, which reduced redundancy and ambiguity in the resulting claims; (2) the extracted claims were decontextualized, unlike in the original FActScore, simplifying their verification. During manual validation, we also examined the quality of decomposed claims in complex cases.

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

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.

```
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 5: 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.

B.3 Automatic Annotation of Claims

As for faithfulness annotation, each claim is assigned one of three categories during the automatic annotation. The categories of faithfulness are “faithful” if the context contains information that directly supports the statement, “unfaithful-contra” if the context contains information that directly contradicts the statement, “unfaithful-neutral” if the context does not contain any information supporting or contradicting the statement. Next, the obtained labels are binarized for experimental evaluation: 1 if labeled faithful and 0 if labeled unfaithful-contra, or unfaithful-neutral. This is done since the unfaithful-contra category is quite small (less than 5% of the data).

Regarding the factuality annotation, the automatic annotation process also assigns one of three categories to each claim. The categories of factuality are “True” if the statement is factually correct according to your knowledge, “False” if the statement is factually incorrect according to your knowledge, “unverifiable” if the factual accuracy of the statement cannot be determined based on available knowledge (should be assessed without referring to the provided context). After that, the annotation for factuality is binarized, namely only verifiable claims are taken (we exclude labels of the unverifiable category), 1 if the claim class is false and 0 if the class is true. Figure 6 presents the annotation prompt used with GPT-4o-search.

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

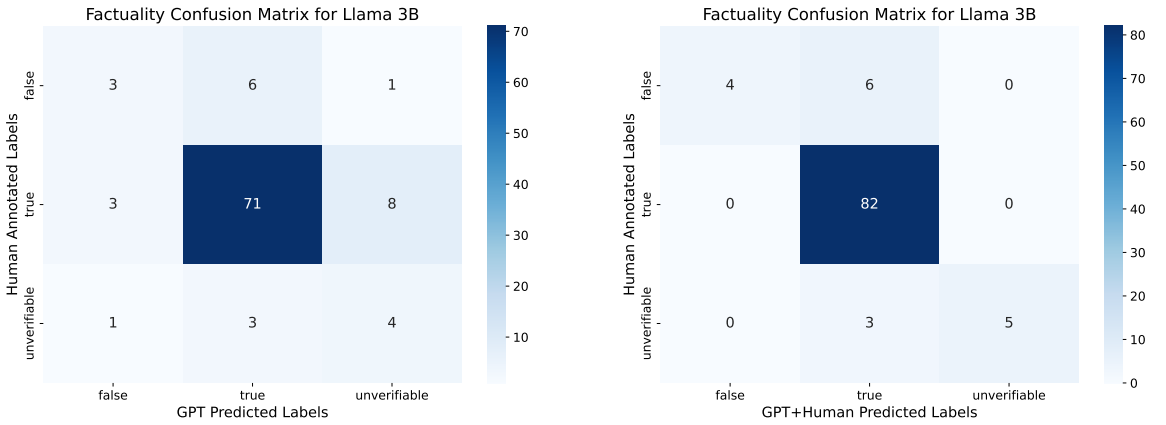
B.4 Manual Enhancement of Automatic Annotation

For manual verification of the automatic annotation, we used reliable sources obtained from Google search results. An initial validation of the automatic labeling was conducted by comparing automatic and manual annotations on randomly selected claims to assess class balance. Specifically, 100 claims were reviewed for Llama 3B Instruct and 76 for Falcon 3B Base. The resulting class distributions are shown in Figure 7(a) and Figure 8(a), respectively. We also present the results of comparison of manual and automatic annotation for faithfulness for both models; see Figure 9(a, b).

Given the difficulty of assessing the false and unverifiable categories in automatic annotations, these cases were prioritized for manual review. Enhanced annotations for the same 100 (Llama 3B Instruct) and 76 (Falcon 3B Base) claims are shown in Figure 7(b) and Figure 8(b). In addition, a full manual re-check of all False and Unverifiable claims was performed, resulting in 359 manually reviewed claims for Llama 3B Instruct and 240 for Falcon 3B Base.

We used six student annotators for this study. Each annotator spent approximately three hours on the annotation task. Instructions were provided orally in the form of informal discussions; no written guidelines were used. All annotators participated voluntarily and were not financially compensated.

To assess annotation consistency, we also conducted an agreement analysis on the 100 Llama 3B Instruct claims, each independently reviewed by two annotators (see Table 10 for details). These results suggest that the annotators generally aligned well in both tasks, particularly for factuality, though there is some room for improvement, especially in clarifying ambiguous or borderline cases.



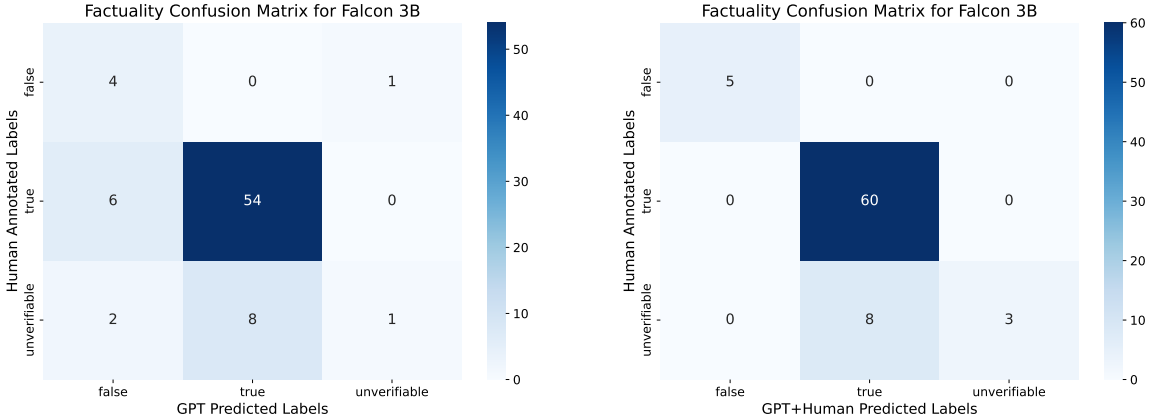
(a) Before manual enhancement of automatic annotation

(b) After manual enhancement of automatic annotation

Figure 7: 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.

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

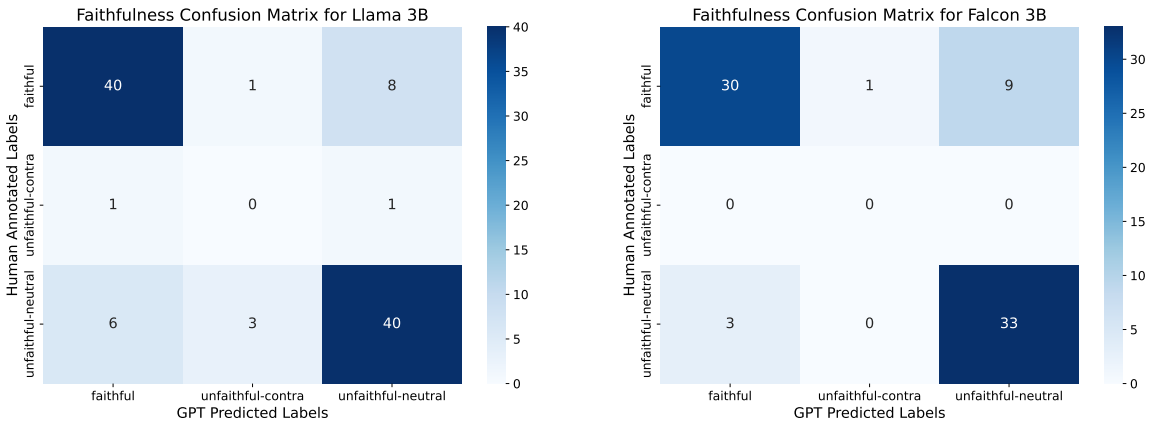
Table 10: Inter-annotator agreement for factuality and faithfulness annotations based on 100 claims of Llama 3B Instruct. Accuracy measures raw agreement, Cohen's Kappa adjusts for chance agreement.



(a) Before manual enhancement of automatic annotation

(b) After manual enhancement of automatic annotation

Figure 8: 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 9: 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 Results

C.1 Faithfulness Evaluation on Long-Form QA

Table 11 shows the long-form QA dataset results when using faithfulness as the target metric. In this evaluation, all methods remained consistent with those used in the factuality dataset. Among the compared methods, AlignScore achieved the best performance, indicating its effectiveness in approximating the faithfulness within the FRANQ formula.

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

Table 11: Llama 3B Instruct results on long-form QA benchmark with faithfulness target.

C.2 Factuality under Faithful/Unfaithful Conditions

Tables 12 present the results for high and low AlignScore, respectively. The results indicate that Semantic Entropy is the top-performing method for faithful claims, and sum of Eigenvalues of Graph Laplacian is the top-performing method for unfaithful claims.

Table 13 presents the results of all unsupervised methods with only unfaithful claims left in the test. Parametric Knowledge outperforms other methods with respect to ROC-AUC and PRR metrics.

Method	Llama 3B Instruct					
	MeanRank ↓			MeanValue ↑		
	ROC-AUC	PR-AUC	PRR	ROC-AUC	PR-AUC	PRR
Max Sequence Probability	7.00	7.00	7.50	.754	.518	.454
CCP	9.00	7.75	8.75	.742	.512	.434
Mean Token Entropy	5.25	5.25	5.00	.767	.540	.472
P(True)	15.25	15.25	15.25	.486	.252	-.001
Lexical Similarity	7.00	7.00	7.00	.758	.500	.457
Degree Matrix	4.50	4.00	4.25	.770	.549	.488
Eccentricity	5.50	5.25	4.75	.768	.547	.482
Sum of Eigenvalues	5.50	6.50	6.25	.767	.538	.476
Number of Semantic Sets	13.00	12.50	12.50	.625	.383	.232
Semantic Entropy	**2.50**	**3.00**	**2.50**	**.781**	**.562**	**.510**
Perplexity	8.25	7.50	8.00	.741	.509	.426
TokenSAR	9.25	9.25	9.75	.733	.493	.408
SentenceSAR	5.00	6.75	5.50	.766	.518	.473
AlignScore	13.25	13.50	13.50	.606	.321	.170
Parametric Knowledge	10.25	10.25	10.00	.657	.413	.295

(a) Only claims with AlignScore > 0.5

Method	Llama 3B Instruct					
	MeanRank ↓			MeanValue ↑		
	ROC-AUC	PR-AUC	PRR	ROC-AUC	PR-AUC	PRR
Max Sequence Probability	8.50	8.00	8.50	.752	.648	.446
CCP	7.75	9.25	8.00	.741	.631	.445
Mean Token Entropy	6.75	5.00	6.75	.755	.673	.462
P(True)	15.50	15.50	15.25	.474	.405	-.016
Lexical Similarity	6.25	8.00	7.00	.767	.662	.469
Degree Matrix	3.00	2.75	2.50	.796	.728	.551
Eccentricity	3.50	3.50	4.25	.793	.709	.527
Sum of Eigenvalues	**1.50**	**2.50**	**2.00**	**.807**	**.735**	**.560**
Number of Semantic Sets	11.50	9.50	10.00	.693	.605	.377
Semantic Entropy	3.75	4.25	5.25	.782	.689	.502
Perplexity	9.25	8.75	9.25	.725	.634	.411
TokenSAR	11.00	10.50	10.50	.710	.616	.390
SentenceSAR	6.25	6.75	5.50	.770	.667	.473
AlignScore	14.00	14.00	14.25	.555	.488	.142
Parametric Knowledge	12.50	13.00	12.25	.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.

Method	Llama 3B Instruct		
	ROC-AUC ↑	PR-AUC ↑	PRR ↑
MCP	.538	.115	.028
PTrue	.463	.112	.002
Perplexity	.480	.092	-.068
MTE	.580	.167	.122
CCP	.585	.134	.152
AlignScore	.477	.094	-.007
Parametric Knowledge	**.667**	**.190**	**.303**

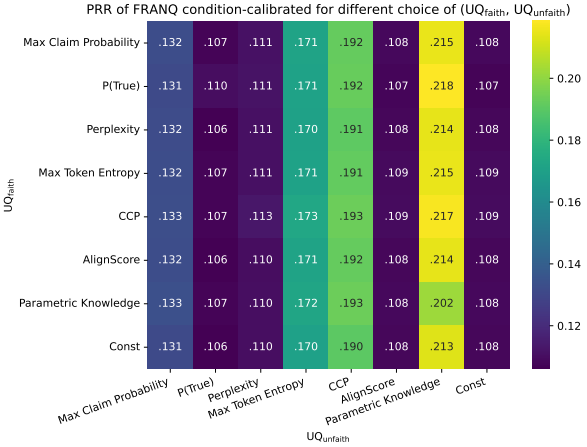
Table 13: Results on long-form QA benchmark with factuality target (only unfaithful claims).

D Ablation Studies

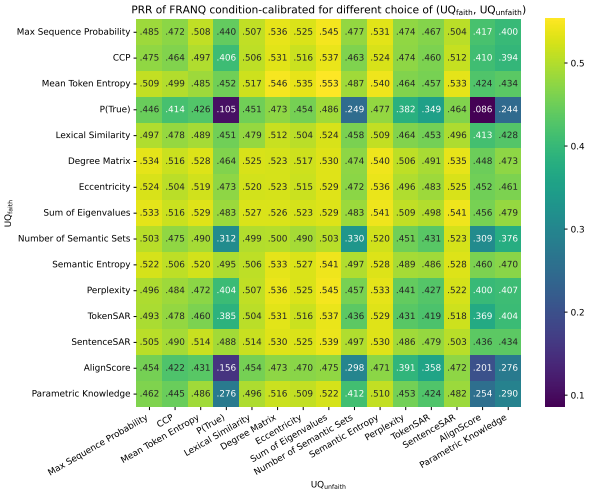
D.1 FRANQ with Different Choices of UQ_{faith} and $UQ_{unfaith}$

Figure 10 presents the PRR of FRANQ condition-calibrated, for different choices of UQ_{faith} and $UQ_{unfaith}$ methods. On the long-form dataset (see Figure 10(a)), the selection of UQ_{faith} has a minimal impact, while opting for Parametric Knowledge as the $UQ_{unfaith}$ method is essential.

The short-form QA results (see Figure 10(b)) show that many combinations of UQ_{faith} and $UQ_{unfaith}$ achieve PR-AUC scores close to the top, indicating flexibility in selecting these components. The pair chosen for short-form QA experiments (Semantic Entropy and Sum of Eigenvalues) achieves a PR-AUC of 0.541, which is close to the highest score of 0.553 obtained by Max Token Entropy and Sum of Eigenvalues.



(a) PRR on long-form QA dataset.



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

Figure 10: Comparison of FRANQ condition-calibrated with different choice of UQ_{faith} and $UQ_{unfaith}$.

D.2 FRANQ with Alternative Faithfulness Estimators

Table 14 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 15 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.

Method	Llama 3B Instruct, long-form QA			Llama 3B Instruct, short-form QA (mean values across 4 datasets)		
	ROC-AUC ↑	PR-AUC ↑	PRR ↑	ROC-AUC ↑	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

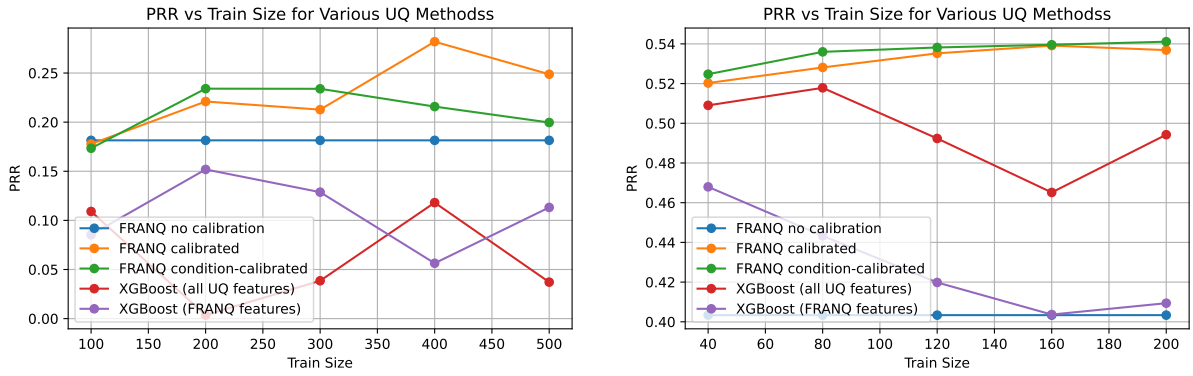
Table 14: Comparison of FRANQ performance on Llama 3B Instruct benchmarks, when using AlignScore with and without threshold.

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

Table 15: Comparison of FRANQ performance on Llama 3B Instruct long-form QA benchmark, when applying calibration for faithfulness estimator, AlignScore.

D.3 Impact of Train Size on FRANQ

Figure 11 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.



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

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

Figure 11: PRR comparison of FRANQ and XGBoost methods with different train size.

D.4 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 12 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.

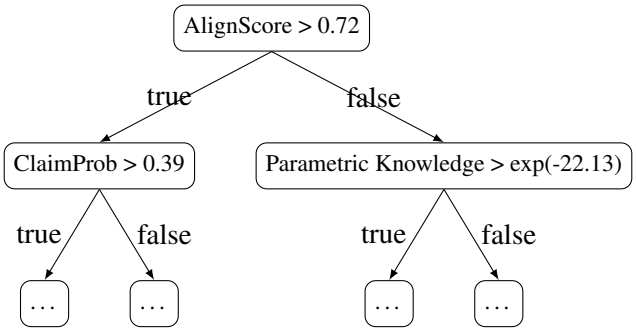


Figure 12: Top vertices of first XGBoost tree trained on FRANQ components (ClaimProb) for long-form QA Llama 3B Instruct behchmark.

D.5 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 16 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	ECE ↓
Max Claim Probability	.72
P(True)	.94
Perplexity	.18
CCP	.21
AlignScore	.40
Parametric Knowledge	.80
XGBoost (all UQ features)	.05
XGBoost (FRANQ features)	.06
FRANQ no calibration	.44
FRANQ calibrated	**.02**
FRANQ condition-calibrated	.03

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

Method	Mean ECE ↓
Max Sequence Probability	.46
CCP	.23
P(True)	.71
Lexical Similarity	**.07**
Degree Matrix	.14
Eccentricity	.46
Sum of Eigenvalues	.54
Number of Semantic Sets	.48
Perplexity	.17
AlignScore	.13
Parametric Knowledge	.23
XGBoost (all UQ features)	.15
XGBoost (FRANQ features)	.17
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).

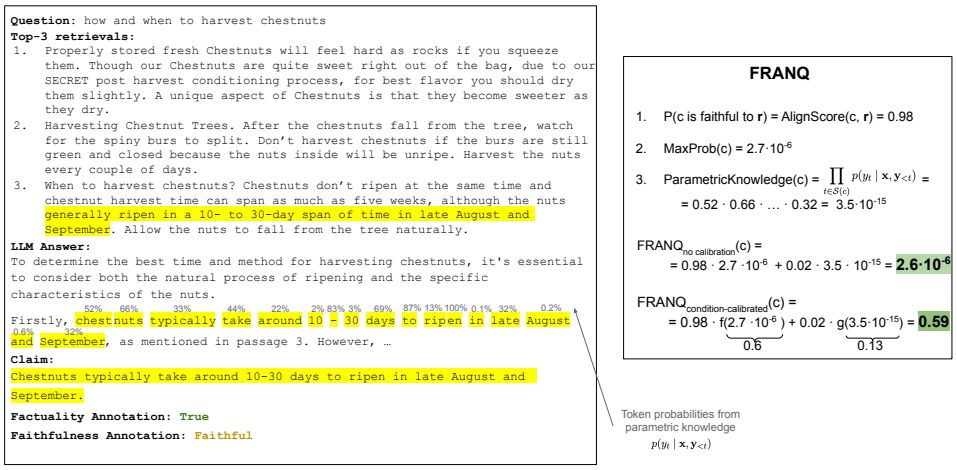
Table 16: Expected Calibration Error (ECE) for all tested UQ methods with Llama 3B Instruct.

E FRANQ Examples

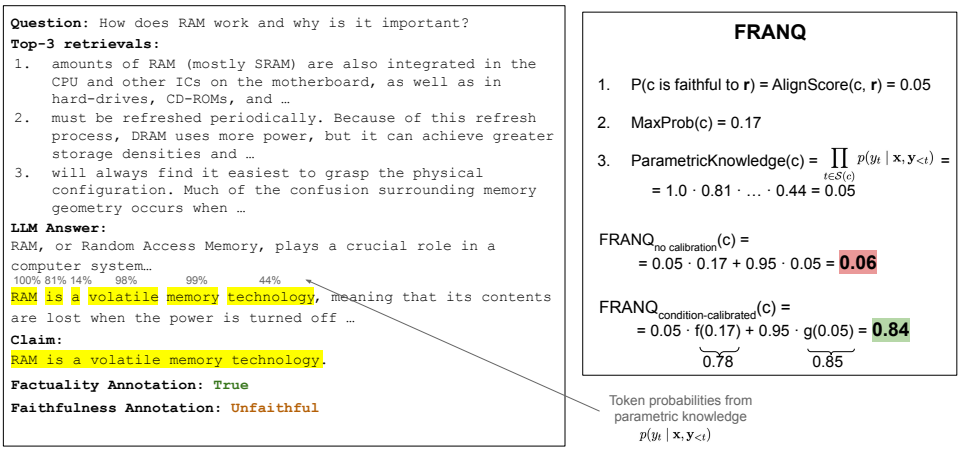
In Figure 13, we demonstrate the behavior of FRANQ using three examples from a long-form QA dataset evaluated with Llama 3B Instruct. We selected three representative claims and present their corresponding FRANQ scores for both the uncalibrated version and condition-calibrated version. The latest uses monotonic functions f and g, fitted via isotonic regression for Claim Probability and Parametric Knowledge methods, respectively.

F Resource and Expenses

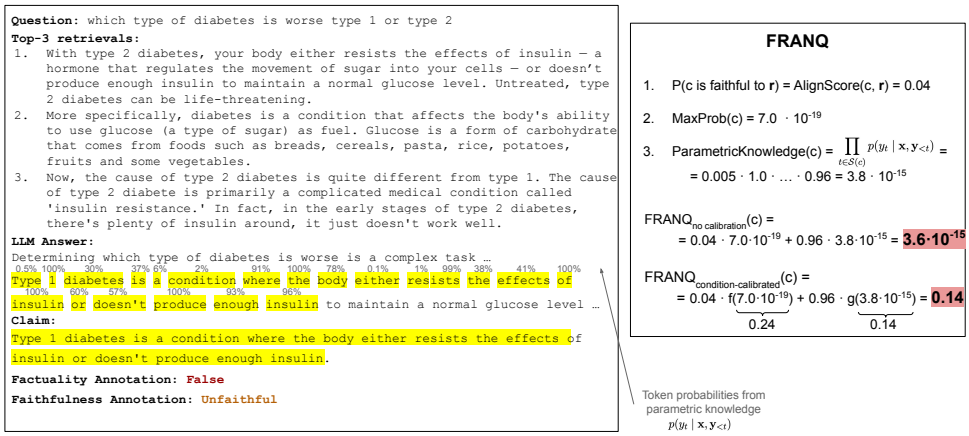
For a single run of data generation and evaluation of UQ baseline methods, we used approximately 8 days of compute time on an NVIDIA V100 32GB GPU across all datasets and models for the long-form QA task. In comparison, the short-form QA task required less than one day of GPU time. The OpenAI API was employed for atomic claim splitting, matching, and annotation. The total API cost for processing a single model run (Llama 3B Instruct) was approximately $100. In addition to automated processing, human annotations were conducted by six student annotators, each contributing roughly three hours of work.



(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 13: 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.