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

Ekaterina Fadeeva[1] ♢ **Aleksandr Rubashevskii**[2] ♢ **Dzianis Piatrashyn**[2]
Roman Vashurin[2] **Shehzaad Dhuliawala**[1] **Artem Shelmanov**[2] **Timothy Baldwin**[2]
Preslav Nakov[2] **Mrinmaya Sachan**[1] **Maxim Panov**[2]
[1]ETH Zürich [2]MBZUAI
{efadeeva,sdhuliawala,msachan}@ethz.ch
{aleksandr.rubashevskii,preslav.nakov,maxim.panov}@mbzuai.ac.ae

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 https://github.com/stat-ml/rag_uncertainty.

1 Introduction

Large Language Models (LLMs) are increasingly employed across a wide range of tasks, including natural language understanding, generation, and reasoning. 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, as well as 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). The use of retrieved information complicates both their detection and source attribution, as models become more confident in generating statements that appear in the retrieval, regardless of 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., 2025a,b).

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 considers any content not supported by the retrieved context as hallucination (Niu et al., 2024). However, we argue that hallucination should instead be defined based on factual inaccuracies: statements outside the retrieved context should not be considered hallucinations if they are 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).

♢ Equal contribution

Findings of the Association for Computational Linguistics: ACL 2026, pages 6814–6836
July 2-7, 2026

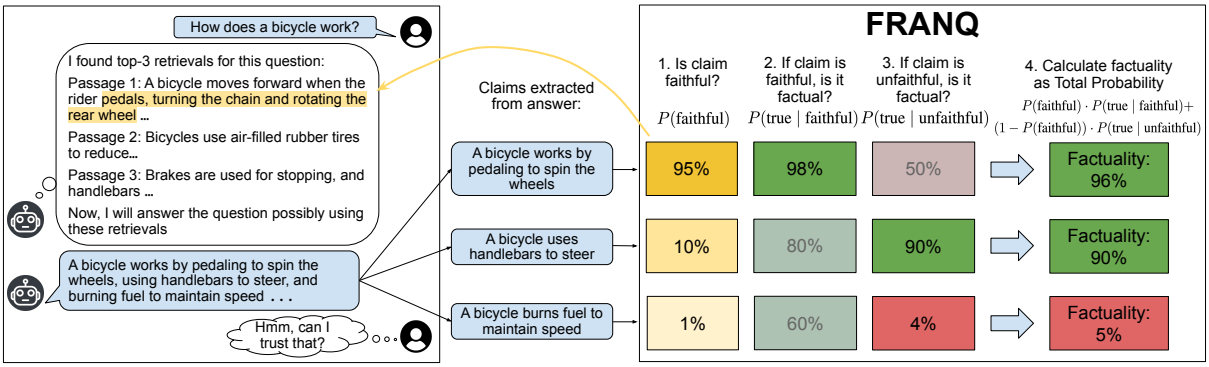


Figure 1: FRANQ illustration. *Left*: A user poses a question, and the RAG retrieves relevant documents and formulates an answer, potentially using information from the retrieved documents. *Middle*: The RAG output is decomposed into atomic claims. *Right*: The FRANQ method assesses factuality by evaluating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

Here, 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 U**N**certainty **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 propose FRANQ, a UQ method for RAG that first assesses faithfulness and then applies different uncertainty estimation methods to faithful and unfaithful outputs (Section 2).
- We introduce a long-form QA factuality dataset for RAG with both factuality and faithfulness labels, built through automatic annotation and manual validation of difficult cases (Section 3).
- We conduct extensive experiments on long- and short-form QA across several LLMs, showing that FRANQ improves factual error detection in RAG outputs over existing approaches (Section 4).

2 Uncertainty Quantification for RAG

Let $\mathbf{x}$ be the user query to the RAG system. The RAG system then retrieves k passages denoted by $\mathbf{r} = \{r_1, \ldots, r_k\}$, from an external knowledge source using $\mathbf{x}$ as the query, and uses an LLM to generate 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 supported by established, verifiable knowledge, and *factually false* otherwise. A claim is deemed *faithful* with respect to the retrieved documents $\mathbf{r}$, if it is entailed by them, and *unfaithful* otherwise. Importantly, factuality and faithfulness capture different aspects of correctness: a claim may be factually true yet not grounded in the retrieved context, or faithful to the retrieval while still being factually incorrect. 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 LLM outputs conditioned on the joint prompt $(\mathbf{x}, \mathbf{r})$. However, this 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 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 $P(c \text{ is unfaithful to } \mathbf{r})$ is calculated as $1 - P(c \text{ faithful to } \mathbf{r})$. 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 F.

2.2 FRANQ Components

We now describe the components of 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.

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.

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, the answer $\mathbf{y}$ is treated as a single claim, and we prepend the question context and 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, *Parametric Knowledge* offers an effective estimate of factuality for unfaithful claims (see Section 4.4). In short-form QA, a broader range of general UQ baselines is applicable, including methods based on sample diversity (see Table 1). In this setting, the *Sum of Eigenvalues* (Lin et al., 2024) offers a better approximation of factuality (see Section 4.4). Thus, we use *Parametric Knowledge* for long-form QA and *Sum of Eigenvalues* for 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:

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

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- 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:

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

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 (Izacard et al., 2022) with embeddings from 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., 2024; Vashurin et al., 2025). Following prior work, GPT-4o first extracts decontextualized atomic claims from the full paragraph through a dedicated prompt. Then, for each claim, a second prompt identifies the corresponding words in 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 claim extraction and mapping are listed in Appendix B.

Annotation. We annotate each atomic claim for both faithfulness and factuality using a two-stage procedure. In the first stage, we use GPT-4o-search, a GPT-4o-based search model augmented with web search over up-to-date web sources, to assign faithfulness labels (*faithful* or *unfaithful*) and factuality labels (*True*, *False*, or *Unverifiable*) using dedicated prompts.

In the second stage, we manually review all claims initially labeled as *False* or *Unverifiable*, which are the most difficult cases for automatic annotation, and correct the labels when necessary. This targeted verification step helps improve the overall quality and reliability of the annotations. We then retain only verifiable claims (*True* or *False*) and binarize the labels accordingly.

This design is supported by our validation analysis: we found that the automatic annotation is substantially more reliable for *True* claims than for *False*/*Unverifiable* ones (90% vs. ∼40% precision), which motivates targeted manual verification of the latter. On the Llama 3B Instruct subset, this second stage improves overall annotation accuracy from 78% to 91%. Human inter-annotator accuracy for factuality is 0.87, indicating strong agreement among the annotators. Further details on the prompts, annotation protocol, dataset statistics, and agreement/error analysis 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.

Method	Llama 3B Instruct PR-AUC ↑	Llama 3B Instruct PRR ↑	Falcon 3B Base PR-AUC ↑	Falcon 3B Base PRR ↑	Llama 8B Instruct PR-AUC ↑	Llama 8B Instruct PRR ↑	Gemma 4B Instruct PR-AUC ↑	Gemma 4B Instruct PRR ↑
General Baselines								
Max Claim Prob.	.058 ± .008	-.029 ± .017	.126 ± .007	.258 ± .015	.055 ± .004	.118 ± .023	.061 ± .005	.000 ± .022
P(True)	.117 ± .012	.207 ± .023	.077 ± .003	.170 ± .013	.071 ± .009	.112 ± .026	.096 ± .012	.148 ± .018
Perplexity	.056 ± .004	-.081 ± .018	.090 ± .004	.165 ± .016	.075 ± .009	.090 ± .024	.048 ± .003	-.071 ± .019
Max Token Entropy	.109 ± .004	.115 ± .020	.130 ± .008	.219 ± .016	**.102** ± .010	.138 ± .023	.051 ± .003	-.003 ± .022
CCP	.085 ± .006	.169 ± .024	.162 ± .010	.181 ± .017	.061 ± .005	.108 ± .022	.087 ± .008	.216 ± .026
RAG-Specific Baselines								
AlignScore	.075 ± .004	.108 ± .020	.104 ± .005	.233 ± .016	.068 ± .007	.119 ± .025	.061 ± .004	.058 ± .021
Parametric Knowledge	.064 ± .006	.018 ± .021	.067 ± .003	.029 ± .015	.059 ± .005	.047 ± .023	.112 ± .011	.183 ± .025
XGBoost								
XGBoost (all UQ features)	.124 ± .006	.206 ± .022	.088 ± .004	.198 ± .014	.044 ± .003	-.015 ± .024	.073 ± .008	.085 ± .023
XGBoost (FRANQ features)	.111 ± .010	.149 ± .020	.080 ± .005	.086 ± .016	.048 ± .003	.017 ± .023	.090 ± .004	.158 ± .022
FRANQ								
FRANQ no calibration	.100 ± .007	.181 ± .024	.135 ± .007	**.362** ± .017	.063 ± .005	.162 ± .026	.080 ± .005	.200 ± .021
FRANQ calibrated	.103 ± .008	**.256** ± .020	.074 ± .003	.090 ± .014	.043 ± .003	-.047 ± .022	**.150** ± .005	**.401** ± .022
FRANQ condition-calibrated	**.140** ± .012	.223 ± .025	**.173** ± .010	.354 ± .017	.081 ± .008	**.184** ± .027	.090 ± .008	.208 ± .025

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.

Method	Llama 3B Instruct PR-AUC ↑	Llama 3B Instruct PRR ↑	Falcon 3B Base PR-AUC ↑	Falcon 3B Base PRR ↑	Llama 8B Instruct PR-AUC ↑	Llama 8B Instruct PRR ↑	Gemma 12B Instruct PR-AUC ↑	Gemma 12B Instruct PRR ↑
General Baselines								
Max Sequence Prob.	.558 ± .007	.454 ± .008	.628 ± .004	.256 ± .007	.569 ± .007	.407 ± .007	.400 ± .006	.162 ± .009
Mean Token Entropy	.594 ± .007	.481 ± .008	.613 ± .004	.242 ± .007	.640 ± .007	.491 ± .008	.423 ± .006	.230 ± .008
CCP	.551 ± .007	.443 ± .008	.641 ± .004	.304 ± .007	.553 ± .007	.417 ± .008	.412 ± .006	.198 ± .009
Lexical Similarity	.564 ± .008	.479 ± .008	.618 ± .004	.277 ± .007	.639 ± .007	.532 ± .007	.430 ± .007	.240 ± .009
Degree Matrix	.629 ± .008	.520 ± .008	.702 ± .003	.464 ± .006	.627 ± .007	.492 ± .007	.464 ± .007	.260 ± .009
Sum of Eigenvalues	.628 ± .008	.518 ± .008	.700 ± .003	.460 ± .006	.628 ± .007	.489 ± .007	.467 ± .007	.260 ± .009
Semantic Entropy	.613 ± .007	.525 ± .008	.623 ± .004	.278 ± .006	.637 ± .007	.519 ± .007	.466 ± .007	.261 ± .008
SentenceSAR	.571 ± .007	.483 ± .008	.602 ± .004	.263 ± .006	.556 ± .007	.414 ± .007	.416 ± .006	.174 ± .008
RAG-specific Baselines								
AlignScore	.415 ± .007	.207 ± .009	.666 ± .005	.372 ± .008	.432 ± .007	.224 ± .008	.376 ± .006	.158 ± .008
Parametric Knowledge	.425 ± .007	.247 ± .009	.556 ± .005	.104 ± .009	.499 ± .007	.330 ± .008	.364 ± .006	.105 ± .009
XGBoost								
XGBoost (all UQ features)	.594 ± .008	.494 ± .008	.705 ± .004	.462 ± .007	.634 ± .007	.503 ± .008	.474 ± .007	**.301** ± .008
XGBoost (FRANQ features)	.526 ± .008	.409 ± .008	.670 ± .004	.368 ± .007	.524 ± .007	.385 ± .008	.414 ± .006	.196 ± .009
FRANQ								
FRANQ no calibration	.553 ± .007	.403 ± .008	.641 ± .003	.345 ± .006	.523 ± .007	.340 ± .008	.447 ± .007	.225 ± .008
FRANQ calibrated	.628 ± .007	.537 ± .008	.672 ± .003	.411 ± .006	.644 ± .007	.534 ± .007	.481 ± .007	.258 ± .009
FRANQ condition-calibrated	**.631** ± .007	**.541** ± .008	**.711** ± .003	**.477** ± .006	**.647** ± .007	**.540** ± .007	**.496** ± .007	.283 ± .009

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 Experiments

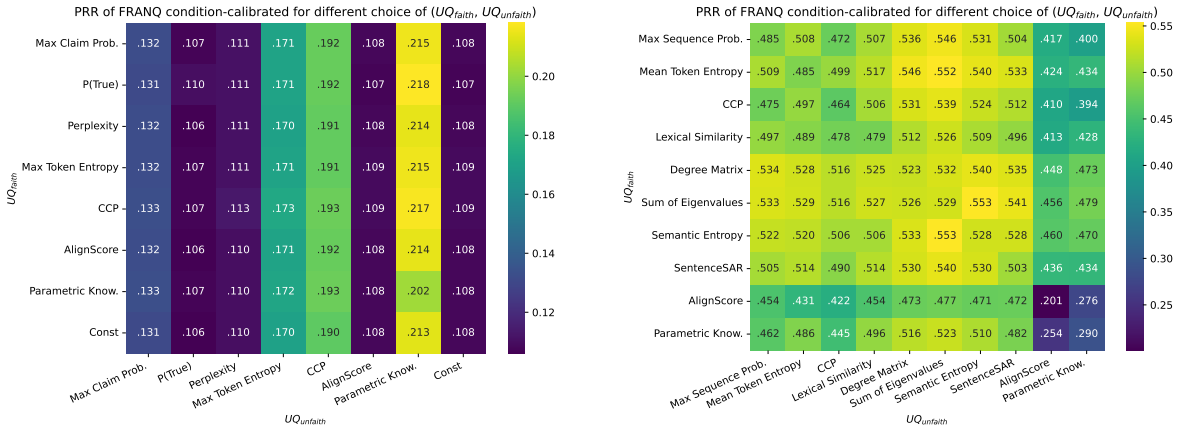
In this section, we evaluate FRANQ and corresponding baselines on both the short-form and long-form benchmarks described in Section 3. For all experiments, we fix the retrieval process and the underlying white-box LLM, and we assess the factual accuracy of the model-generated claims. This allows us to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

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



(a) PRR on long-form QA dataset.

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

Figure 2: PRR of condition-calibrated FRANQ for different choices of UQ_{faith} and $UQ_{unfaith}$.

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}, $UQ_{unfaith}$), and (2) all available unsupervised UQ method.

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 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.

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.

Method	Shuffled retrievals PR-AUC ↑	Shuffled retrievals PRR ↑	Corrupted retrievals PR-AUC ↑	Corrupted retrievals PRR ↑
General Baselines				
Max Sequence Prob.	.638	.489	.776	.374
Mean Token Entropy	.651	.460	.767	.358
CCP	.629	.472	.779	.392
Lexical Similarity	.668	.512	.767	.358
Degree Matrix	.687	.548	.781	.406
Sum of Eigenvalues	.686	.536	.782	.404
Semantic Entropy	.666	.537	.780	.447
SentenceSAR	.648	.523	.772	.381
RAG-specific Baselines				
AlignScore	.507	.234	.659	-.028
Parametric Knowledge	.502	.220	.801	.412
XGBoost				
XGBoost (all UQ features)	.684	.544	.774	.354
XGBoost (FRANQ features)	.579	.402	**.813**	.471
FRANQ				
FRANQ no calibration	.586	.393	.768	.331
FRANQ calibrated	.692	.549	.807	.475
FRANQ condition-calibrated	**.695**	**.553**	**.813**	**.478**

Table 4: Robustness of Llama 3B Instruct under two retrieval corruption settings on four short-form QA datasets. In *shuffled retrievals*, 50% of passages are replaced with unrelated ones; in *factually corrupted retrievals*, 50% are modified to contain incorrect facts.

Analysis of FRANQ's components. Figure 2 shows the PRR of condition-calibrated FRANQ for different choices of UQ_{faith} and $UQ_{unfaith}$, evaluated with Llama 3B Instruct on long-form QA and on a subset of 200 questions from each short-form dataset.

On long-form QA (Figure 2(a)), performance is largely insensitive to the choice of UQ_{faith}, whereas the choice of $UQ_{unfaith}$ is critical: using Parametric Knowledge as $UQ_{unfaith}$ yields the strongest PRR for nearly all UQ_{faith} options. This suggests that, in long-form QA, modeling factuality for unfaithful claims is the key design choice.

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

Additional faithfulness ablations are reported in Appendix D.1. In particular, replacing continuous AlignScore probabilities with binary thresholding consistently degrades performance, highlighting the value of probabilistic faithfulness weighting.

Robustness to retrieval corruption. We evaluate FRANQ under two retrieval failure modes. First, we simulate irrelevant retrievals by randomly replacing 50% of retrieved passages across the four short-form QA datasets, ensuring that no corrupted example retains its original retrieval set. This reduces answer accuracy by 7% on average.

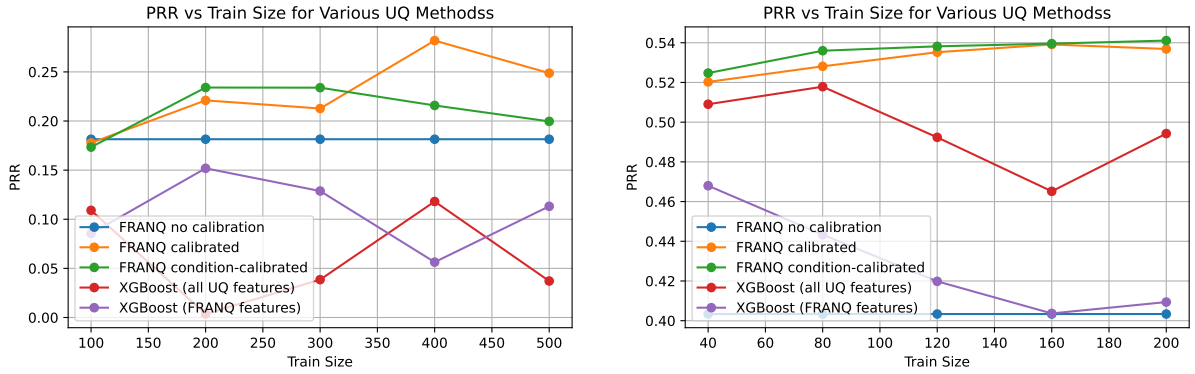
Method	Inference Runtime	Training Time	Model Size
General Baselines			
Max Claim Probability	< 0.1 s	—	—
P(True)	1.3 s	—	—
Perplexity	< 0.1 s	—	—
Max Token Entropy	< 0.1 s	—	—
CCP	1.7 s	—	—
RAG-specific Baselines			
AlignScore	0.5 s	—	—
Parametric Knowledge	1.6 s	—	—
XGBoost			
XGBoost (all UQ features)	1.9 s	0.60 s	10 kB
XGBoost (FRANQ features)	1.7 s	0.12 s	14 kB
FRANQ			
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 5: Inference runtime (averaged across dataset samples), training cost, and model size of uncertainty estimators and FRANQ variants on Llama 3B Instruct for short-form QA, measured beyond base LLM generation.

Second, we consider a more challenging setting in which retrieved passages remain relevant to the input question but contain subtle factual errors. We construct this setting by using GPT-4o to rewrite 50% of retrieved passages with plausible inaccuracies, such as altered dates, people, places, or events. This encourages the LLM to produce answers that are faithful to the retrieval yet factually wrong. In this regime, UQ_{faith} must assess whether the retrieval itself is trustworthy. We find that Parametric Knowledge is the most effective choice for UQ_{faith}, as it captures the model's confidence in the retrieved content.

Table 4 shows that under both shuffled and factually corrupted retrievals, condition-calibrated and calibrated FRANQ rank first and second, respectively, across evaluation metrics. Overall, these results suggest that FRANQ is robust to both irrelevant and misleading retrieved evidence.

Effect of supervision. Figure 3 shows PRR versus training set size for three FRANQ variants and two XGBoost baselines on long-form and short-form QA. As expected, uncalibrated FRANQ is unaffected by training size, while the supervised variants improve with more labeled calibration data and then saturate. On long-form QA, condition-calibrated FRANQ peaks at around 300 training instances; on short-form QA, performance stabilizes at around 120. Across all training sizes, the calibrated FRANQ variants consistently outperform the XGBoost baselines.



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

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

Figure 3: PRR comparison of FRANQ and XGBoost methods across different training set sizes.

Computational efficiency. Table 5 reports the runtime overhead, training cost, and model size of FRANQ's uncertainty components beyond a completed Llama 3B Instruct forward pass on short-form Natural Questions (2.2 s per instance on average). Overall, FRANQ adds only modest inference cost. The supervised variants are especially lightweight: isotonic calibration fits in under one second and yields models of only a few hundred bytes, making calibrated FRANQ practical.

5 Related Work

Uncertainty Quantification for RAG. Several UQ methods for RAG study how retrieved knowledge shapes LLM outputs, including lookback-ratio classifiers (Chuang et al., 2024), feature-based regression models comparing retrieved and parametric knowledge contributions (Sun et al., 2025), uncertainty estimation based on the signal-to-noise ratio of output probabilities across samples (Li et al., 2024), and prompt–response relevance modeling (Hu et al., 2024). These approaches mainly assess hallucinations relative to retrieved context, rather than assessing whether they are factually correct in an absolute sense.

However, they often incur additional computational cost and do not directly address factual correctness beyond the retrieved evidence. Search-based approaches such as SAFE (Wei et al., 2024b) extend verification using LLM agents and web search, but at substantially greater complexity. In contrast, FRANQ is a lightweight, self-contained UQ framework that 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 estimated from internal 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 methods (Fomicheva et al., 2020; Lin et al., 2024). However, these are studied in isolation and ignore interactions with retrieved evidence. Our work unifies both by jointly modeling retrieval-related and intrinsic uncertainty.

Factuality/Hallucination Datasets for RAG. Progress in RAG hallucination detection relies on datasets with reliable factuality annotations. RAGTruth (Niu et al., 2024) provides span-level labels but excludes factually correct content unsupported by retrieval. Dialogue datasets such as Wizard of Wikipedia (Dinan et al., 2019) and FaithDial (Dziri et al., 2022) emphasize grounding, treating unsupported content as hallucinated even if correct. Similarly, QA benchmarks such as RAGBench (Friel et al., 2024) and AdaptiveRAG (Moskvoretskii et al., 2025) define hallucinations relative to context. In contrast, FRANQ focuses on factuality beyond retrieved evidence.

6 Conclusion and Future Work

We introduced FRANQ, a method for quantifying factuality of claims in RAG outputs via faithfulness. Across long- and short-form QA tasks and multiple LLMs, FRANQ outperforms unsupervised UQ baselines, RAG-specific methods, and supervised classifiers. We also presented a QA dataset annotated for factuality and faithfulness using a hybrid labeling process. Our approach opens directions for future work, including using FRANQ's uncertainty signals for generation-time control to enable more reliable RAG systems.

Limitations

While FRANQ achieves strong hallucination detection performance on average, it does not guarantee perfect factuality estimation in every case, especially in difficult settings involving ambiguous claims, limited evidence, or highly uncertain generations.

Although FRANQ is robust to both irrelevant and factually corrupted retrievals, its performance still depends on the quality of the underlying signals used to estimate faithfulness and factuality. In particular, severe retrieval errors or model miscalibration may reduce performance in challenging real-world settings.

Since some FRANQ variants rely on calibration of their components, they require a small amount of labeled data. However, our experiments show that these variants remain effective with limited supervision and outperform stronger supervised baselines.

Ethical Considerations

FRANQ is designed to reduce the spread of factual errors by improving the reliability and interpretability of language model outputs. By distinguishing between factuality and faithfulness, it can avoid penalizing factually correct but unsupported claims. However, FRANQ does not prevent hallucinations directly and instead relies on downstream filtering, so its impact depends on how it is integrated into larger systems.

Although FRANQ is robust to noisy and factually corrupted retrievals, its performance still depends on retrieval quality. In real-world applications, retrieved documents may be biased, outdated, or incorrect, which can affect the method's output. Careful source selection and monitoring remain important to avoid reinforcing misinformation or harmful biases.

The long-form evaluation pipeline relies partly on GPT-4o-based claim extraction and annotation. While we combine automatic annotation with targeted manual verification, some biases from the underlying model may still persist. Future work could explore more diverse annotation strategies, including broader human validation.

Better factuality estimation can support safer deployment of AI systems in knowledge-intensive domains such as education, healthcare, and law. Still, FRANQ should be viewed as a decision-support tool rather than a replacement for human fact-checking.

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, EMNLP '2024, pages 1419–1436, Miami, Florida, USA. 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 Proceedings of the 7th International Conference on Learning Representations, ICLR '19, New Orleans, LA, USA. OpenReview.net.

Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2024. Rowen: Adaptive retrieval-augmented generation for hallucination mitigation in LLMs. 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), ACL '2024, pages 5050–5063, Bangkok, Thailand. 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, Bangkok, Thailand. 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, EMNLP '2023, pages 446–461, Singapore. 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. ArXiv 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, ICML '2017, pages 1321–1330, Sydney, Australia. 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. 2022. Unsupervised dense information retrieval with contrastive learning. Transactions on Machine Learning Research.

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), ACL '2017, pages 1601–1611, Vancouver, Canada. 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), ACL '2025, 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 Proceedings of the Eleventh International Conference on Learning Representations, ICLR '23, Kigali, Rwanda. OpenReview.net.

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 Proceedings of the Advances in Neural Information Processing Systems, NeurIPS '2020, pages 9459–9474, Online. Curran Associates, Inc.

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), ACL '2023, pages 9802–9822, Toronto, Canada. 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, ACL '2020, 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, EMNLP '2023, pages 12076–12100, Singapore. 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), ACL '2025, pages 6355–6384, Vienna, Austria. Association for Computational Linguistics.

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), ACL '2024, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

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 Proceedings of the International Conference on Machine Learning, ICML '2023, pages 31210–31227, Honolulu, USA. 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: Empirical Methods in Natural Language Processing, EMNLP '21, pages 3784–3803, Punta Cana, Dominican Republic. 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 Proceedings of the Thirteenth International Conference on Learning Representations, ICLR '25, Singapore.

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 O Arik. 2025a. Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL '2025, pages 30553–30571, Vienna, Austria. Association for Computational Linguistics.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025b. Retrieval-augmented generation with conflicting evidence. In Second Conference on Language Modeling, COLM '2025, Montreal, Canada.

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. 2024. Factcheck-Bench: Fine-grained evaluation benchmark for automatic fact-checkers. In Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing, EMNLP '2024, pages 14199–14230, Miami, Florida, USA. 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 preprint, 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 Proceedings of the Advances in Neural Information Processing Systems, NeurIPS '2024, pages 80756–80827, Vancouver, Ontario, Canada. Curran Associates Inc.

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), ACL '2023, pages 11328–11348, Toronto, Canada. 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.

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.

A Additional Short-Form QA Results

In Table 3 of the main text, we reported aggregated results for short-form QA using mean values for ease of presentation and to facilitate direct comparison across methods and models in a concise and interpretable manner. 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), offering a more detailed breakdown of performance across datasets and highlighting dataset-specific trends and variations in method behavior, as well as enabling more fine-grained analysis of individual model performance and better understanding of method robustness across different question types.

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

For Falcon 3B Base, FRANQ condition-calibrated achieves top performance on TriviaQA and second-best on Natural Questions. It ranks among the top three methods on PopQA and 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 datasets.

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 4. 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 5. Extracted answers were decomposed into atomic claims using the prompt in Figure 6, and each claim was matched to its corresponding span in the original sentence using Figure 7. 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.3). 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 4: 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 5: 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 6: 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 7: 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.

Label	Instruction
True	Assign *True* if the claim is supported by reliable, verifiable sources under its most natural interpretation. Minor wording differences are acceptable as long as the factual content is preserved.
False	Assign *False* if the claim is contradicted by reliable sources or contains an incorrect factual statement.
Unverifiable	Assign *Unverifiable* if the claim cannot be reliably confirmed or refuted from available sources, or if it is too ambiguous, underspecified, subjective, or dependent on missing context.

Table 10: Protocol used for manual factuality verification of atomic claims.

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

Table 11: 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.

B.3 Long-Form Dataset Annotation

We annotate the long-form QA dataset using a two-stage pipeline. In the first stage, GPT-4o-search assigns faithfulness and factuality labels to extracted atomic claims using prompts. In the second stage, we manually verify the most difficult claims, namely those labeled as *False* or *Unverifiable*, and correct them when necessary. This design scales annotation while focusing human effort on cases where automatic labeling is least reliable.

Automatic First-Pass Annotation. Each claim is automatically annotated for both faithfulness and factuality. For faithfulness, GPT-4o-search assigns one of three labels: *faithful*, *unfaithful-contra*, or *unfaithful-neutral*. In the experiments, these labels are binarized as faithful $\rightarrow$ 1 and unfaithful-contra / unfaithful-neutral $\rightarrow$ 0, since the unfaithful-contra class constitutes less than 5% of the data.

For factuality, GPT-4o-search assigns one of three labels: *True*, *False*, or *Unverifiable*. Factuality is evaluated independently of the retrieved RAG context: the goal is to determine whether the claim is correct with respect to external knowledge rather than whether it is supported by the retrieved passages. For downstream evaluation, we retain only verifiable claims and binarize the labels as False $\rightarrow$ 1 and True $\rightarrow$ 0. The prompt used for automatic annotation is shown in Figure 8.

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 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 cannot be determined with current knowledge.
Return your answer in the exact format: ("faith. label", "factuality label")
Context Document: {retrievals}
Claim: {claim}
Label:

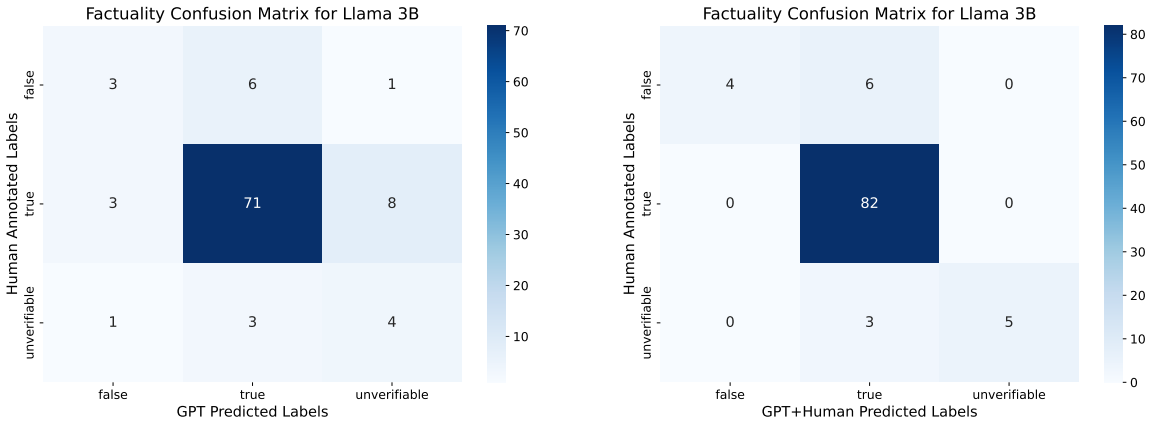
Figure 8: Prompt used with GPT-4o-search to automatically annotate claims for faithfulness and factuality in long-form QA benchmark.

Targeted Manual Verification. We manually validate the automatic annotations to assess their quality and to correct the most difficult cases. First, to estimate automatic annotation reliability and class balance, we compare automatic and human labels on randomly selected claims: 100 for Llama 3B Instruct and 76 for Falcon 3B Base. The resulting factuality comparisons are shown in Figure 9(a) and Figure 10(a); corresponding faithfulness comparisons are shown in Figure 11(a, b), providing a quantitative assessment of annotation accuracy and consistency across both evaluated models.

Our validation shows that factuality labels predicted as *False* or *Unverifiable* are substantially less reliable than those predicted as *True*. We therefore use a targeted second stage in which all claims initially labeled as *False* or *Unverifiable* are manually re-checked using reliable external sources identified through web search. This yields 359 manually reviewed claims for Llama 3B Instruct and 240 for Falcon 3B Base. The resulting corrected label distributions for the sampled subsets are shown in Figure 9(b) and Figure 10(b).

Six student annotators contributed to this manual verification stage, each spending about three hours on the task. Annotators followed the simple factuality protocol summarized in Table 10. All annotators volunteered and received no financial compensation.

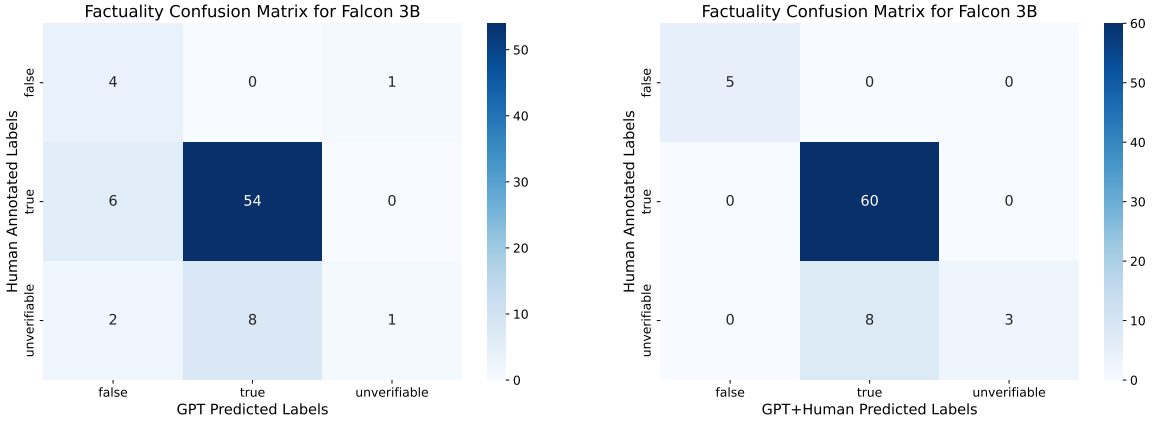
To evaluate annotation consistency, we additionally conduct an agreement analysis on the 100 Llama 3B Instruct claims, each independently reviewed by two annotators (Table 11). The results indicate generally strong agreement for factuality, suggesting that the annotation process is reliable and consistent across annotators, and that disagreements are relatively rare and limited in scope.



(a) Before manual enhancement of automatic annotation

(b) After manual enhancement of automatic annotation

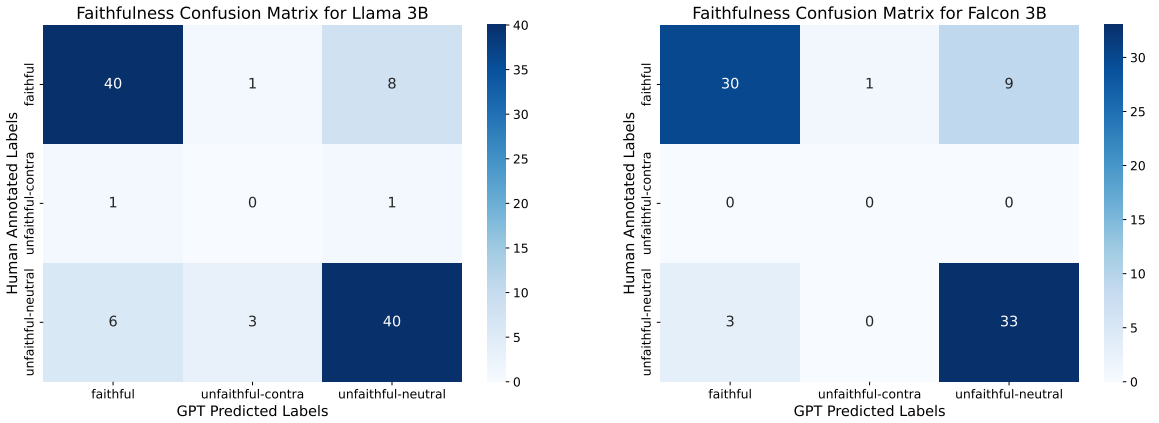
Figure 9: Balance of classes of factuality annotations for the Llama 3B Instruct model. Each matrix is based on 100 randomly selected claims, comparing annotations produced by the model with those from human annotators.



(a) Before manual enhancement of automatic annotation

(b) After manual enhancement of automatic annotation

Figure 10: 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 11: 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.

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

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, examining its performance and robustness across different datasets and settings.

C.1 Faithfulness Distribution and Calibration

We examine the empirical behavior of AlignScore when used to estimate claim-level faithfulness. Figure 13 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, providing insight into how faithfulness scores are distributed across different claims.

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, and can be effectively integrated with other uncertainty estimation components.

Figure 12 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 12a reports performance when faithfulness is treated as the target metric. All methods follow the same experimental setup used for the factuality evaluation, ensuring a fair and consistent comparison across approaches and settings, and isolating the contribution of each estimator to faithfulness prediction performance under controlled conditions and identical input configurations.

Among the compared approaches, AlignScore achieves the strongest performance across metrics, indicating its effectiveness in approximating claim-level faithfulness within the FRANQ decomposition and supporting its suitability as a core component of the framework for reliable faithfulness estimation, particularly in long-form generation settings with multiple claims, diverse evidence sources, and complex reasoning requirements.

C.3 Factuality Under Faithful and Unfaithful Conditions

We further analyze factuality estimation under faithful and unfaithful conditions. Table 12b 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, highlighting the importance of modeling internal knowledge when retrieval grounding is unreliable and external evidence may be misleading or partially incorrect in realistic scenarios.

Table 13 report results averaged across four QA datasets for Llama 3B Instruct, considering only claims with high and low AlignScore, respectively, to separate faithful and unfaithful regimes and analyze performance differences more precisely across datasets and evaluation conditions. 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, demonstrating the benefit of tailoring uncertainty estimation to distinct error sources and improving robustness across diverse retrieval conditions, dataset characteristics, varying levels of evidence quality, and different model behaviors, as well as across varying levels of task difficulty.

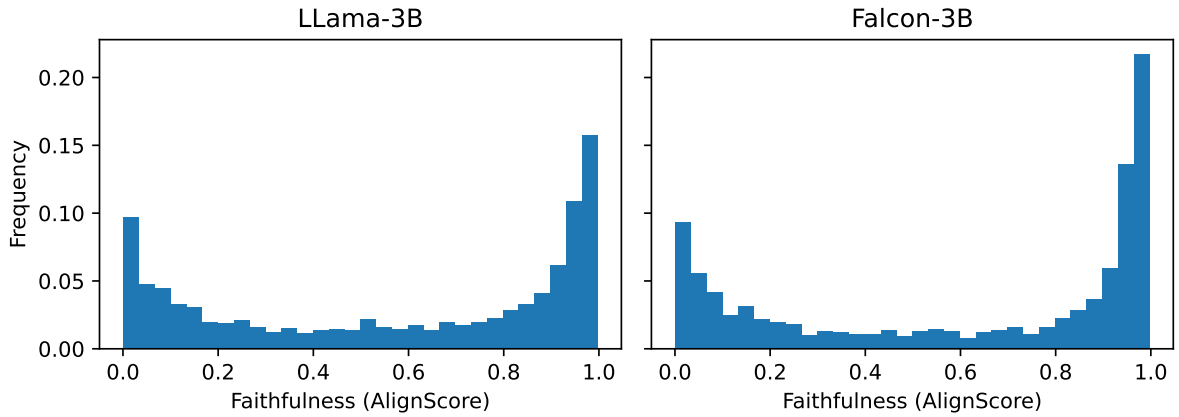


Figure 13: Distribution of AlignScore-based faithfulness estimates on the long-form QA benchmark for Llama 3B Instruct and Falcon 3B Base. A substantial mass lies in the intermediate region, and low ECE values indicate good calibration.

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 12: 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 13: Results averaged across 4 QA datasets for Llama 3B Instruct considering only claims with high and low AlignScore.

D Additional Ablation Studies

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.

D.1 FRANQ with Alternative Faithfulness Estimators

Table 15 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 $P(c \text{ is faithful to } \mathbf{r}) = \mathbb{1}(\text{AlignScore}(c) > T)$ 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, suggesting that preserving the full probabilistic signal of AlignScore is important for effective uncertainty estimation and leads to better integration within the FRANQ decomposition. In particular, thresholding discards intermediate confidence values that may capture partial grounding or uncertainty, which can be informative for downstream factuality estimation and improve robustness in borderline cases, especially when evidence is incomplete or partially relevant.

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, possibly due to the loss of useful ranking information during calibration and reduced sensitivity to fine-grained differences between claims.

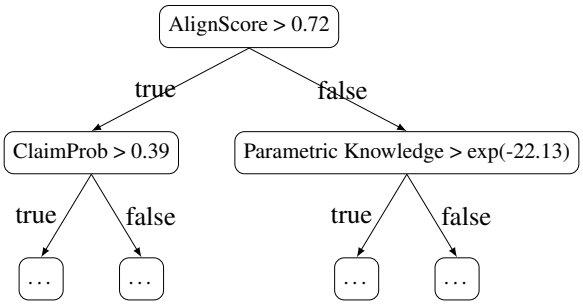


Figure 14: Top vertices of first XGBoost tree trained on FRANQ components (ClaimProb) for long-form QA Llama 3B Instruct behchmark.

D.2 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 and provides an interpretable approximation of the model's behavior.

Figure 14 presents the first several nodes in the 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 to FRANQ's decision process, highlighting alignment between learned and designed decision strategies.

D.3 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	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	.728	.553	.403
FRANQ no calibration T=0.5	.629	.105	.170	.782	.599	.524
FRANQ calibrated	**.653**	.103	**.256**	.797	.628	.537
FRANQ calibrated T=0.5	.607	.085	.084	.796	.566	.521
FRANQ condition-calibrated	.641	**.140**	.223	**.802**	**.631**	**.541**
FRANQ condition-calibrated T=0.5	.587	.111	.180	.793	.556	.515

Table 15: Comparison of FRANQ performance on Llama 3B Instruct benchmarks, when using AlignScore with and without threshold.

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

E Resource and Expenses

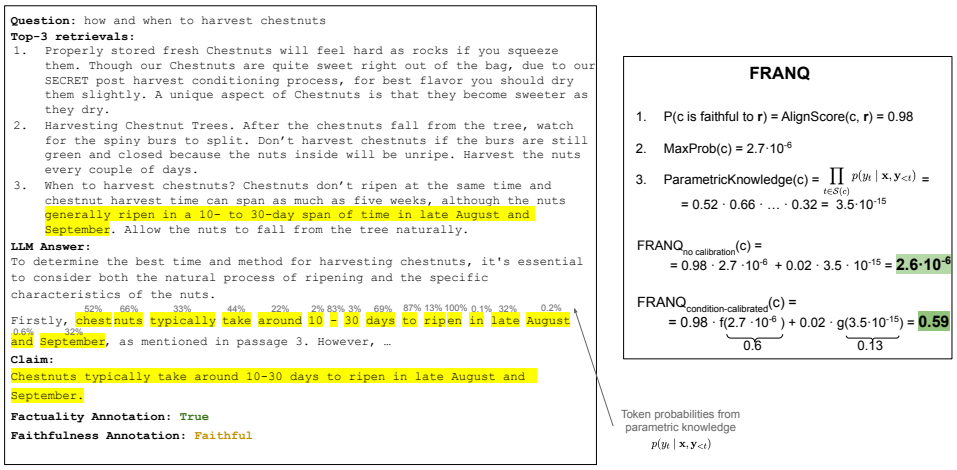
A full data-generation and UQ-baseline evaluation run required about 8 days of compute on an NVIDIA V100 32GB GPU for long-form QA, while short-form QA needed under one day. The OpenAI API was used for claim splitting, matching, and annotation, costing roughly $100 per model run (Llama 3B Instruct).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

F FRANQ Examples

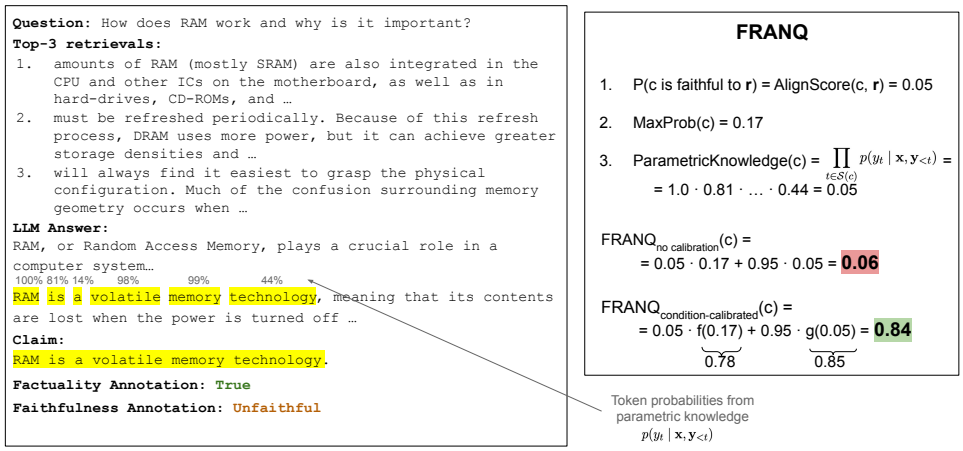
In Figure 15, we demonstrate the behavior of FRANQ using three examples from a long-form QA dataset evaluated with Llama 3B Instruct. We selected three representative claims and present their corresponding FRANQ scores for both the uncalibrated version and condition-calibrated version. The latest uses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

G The Usage of LLMs

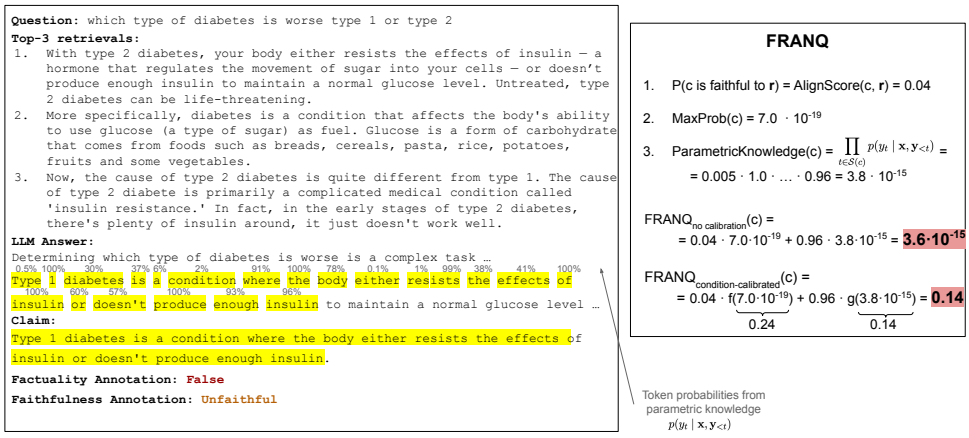
In this study, large language models are examined primarily as the object of analysis. For practical tasks such as programming and writing, we also make limited use of LLM-based assistants (e.g., ChatGPT) for grammar correction and code debugging, with all such use carefully supervised by human researchers.



(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.