Human-Anchored Factuality Evaluation with Strategic Annotation

Yu Wang
Amazon AGI
ywyu@amazon.com

Craig Erickson
Amazon AGI
craigeri@amazon.com

Kevin Small[†]
Amazon AGI
kevinsmall@gmail.com

Abstract

LLM-based factuality judges provide scalable evaluation signals, but their metrics are often systematically biased relative to human judgments. We study human-anchored factuality evaluation under limited annotation budgets, where judge predictions on the full dataset are combined with human labels on a small selectively sampled subset to obtain statistically valid estimates. The efficiency of this approach depends critically on which examples receive human annotation: in factuality evaluation, judge-human misalignment is not driven solely by low confidence, but also by structured failure modes such as incomplete evidence, temporal mismatch, unverifiable claims, and rubric misalignment. To exploit this structure, we introduce a factuality-specific annotation policy design pipeline that uses failure-space analysis (FSA) to derive diverse predictive signals for modeling human-judge misalignment. On an internal reference-based factuality evaluation system (AutoFA) and RAGTruth, where judge-predicted estimates substantially underestimate human-annotated factual accuracy, our FSA-guided policy improves annotation efficiency over uniform sampling and uncertainty-driven baselines, achieving effective-sample-size gains of 40.3% on AutoFA and 27.1% on RAGTruth.

1 Introduction

Factuality evaluation is a recurring concern when deploying large language models in user-facing systems. In production settings, teams rely on factual accuracy to measure system quality (e.g., LLM model, prompts, grounding retrieval), monitor regressions, and make release decisions from evaluation data that are both timely and statistically reliable. Human annotation remains the most trusted way to measure factual accuracy, but it is expensive and slow. Collecting sufficient labels to obtain narrow confidence intervals requires substantial annotation budget and introduces operational latency. This creates a practical tension between evaluation quality and evaluation efficiency.

LLM-as-a-judge (LLMaaJ) systems offer an attractive way to scale factuality evaluation (Min et al., 2023; Chen et al., 2023; Niu et al., 2024; Wei et al., 2024; Song et al., 2024). Given a query, a response, and reference evidence, the judge can decompose the response into claims, verify them against evidence, and produce factuality labels at much larger scale than human annotators. However, judge-predicted factuality estimates are not necessarily human-anchored (Zheng et al., 2023; Wang et al., 2024). In reference-based factuality evaluation, judge errors are often systematic rather than random. A judge may reject reasonable answers when references are incomplete, stale, conflicting, or indirectly supportive. It may also apply a more literal verification rubric than human annotators. As a result, treating judge predictions as ground truth can produce biased factual accuracy estimates, even when the judge is useful as a proxy signal.

This paper studies factuality evaluation under a budgeted annotation setting. Given a large set of evaluation items, an LLMaaJ provides proxy factuality predictions for all items, while only a small subset can receive human labels. The goal is to estimate the human-defined factual accuracy rate (FAR) with valid uncertainty quantification and lower variance than uniformly sampled human annotation alone. We adopt Active Statistical Inference (ASI) (Zrnic and Candes, 2024), which combines judge predictions with human corrections from a sampled subset of examples. ASI preserves the human-anchored target, but its efficiency depends critically on the annotation policy – human labels should be allocated to examples where the judge is most likely to disagree with humans.

Factuality residuals are often structured rather than uniformly distributed across examples. Exist-

[†]Now at Microsoft.

EMNLP 2026 Industry Track, pages 2218–2233
October 24-29, 2026 ©2026 Association for Computational Linguistics

ing confidence-based approaches (Gligorić et al., 2025) use the judge's expressed uncertainty as a sampling signal, but judge-human misalignment is not driven solely by low confidence. A judge can be confidently misaligned when evidence is incomplete, temporally inconsistent, or only indirectly supportive, and when responses require paraphrase, inference, or rubric tolerance that humans handle differently from the judge. To exploit this structure, we construct the annotation policy using *failure-space analysis* (FSA), a factuality-specific policy-design pipeline that derives predictive signals from judge-output, evidence quality, task and input strata, and answer-rubric alignment before converting them into ASI sampling policies.

We evaluate the framework on two benchmarks: an internal Automatic Factual Evaluation system (AutoFA) for virtual assistant responses, and RAGTruth, a public hallucination benchmark with human span-level annotations. In both settings, judge-predicted FAR substantially underestimates the human-annotated FAR and adopting FSA-derived policies preserve desired coverage while improving annotation efficiency over uniform sampling and confidence-based sampling. Aggregated across budgets, FSA-guided ASI achieves an average effective sample size (ESS) gain of 40.3% on AutoFA and 27.1% on RAGTruth.

To summarize, our contributions are threefold. First, we formulate reference-based factuality evaluation as a human-anchored, judge-assisted inference problem under limited annotation budget. Second, we introduce failure-space analysis as a practical policy-design layer for identifying factuality-specific judge-human residual risk. Third, we demonstrate on both internal and public factuality settings that FSA-guided ASI recovers the human-defined evaluation target while improving annotation efficiency over uniform and confidence-based alternatives. Together, these contributions enable reliable and cost-efficient factuality measurement in practical evaluation pipelines, where fully manual annotation is infeasible and judge-only metrics can be systematically biased.

2 Background

2.1 Problem Setup

We first formulate the problem of reference-based factuality evaluation for short-form query-response pairs, while noting the proposed methodology can be naturally extended to long form settings. Each evaluation item consists of a query q_i, a model response a_i, and the reference evidence e_i obtained from retrieval, curated sources, or an existing evaluation pipeline. We write the observable input as $X_i = \{q_i, a_i, e_i\}$ and denote the human-annotated factuality label of each X_i by Y_i. In practice, Y is usually not a direct verdict, but verified atomic claims, or detected hallucinated spans.

In this paper, we focus on *Factual Accuracy Rate* (FAR)[1], which is defined as:

$$\theta = \mathbb{E}[\phi(X, Y)], \tag{1}$$

where $\phi(X_i, Y_i)$ denotes the factual correctness for the i_{th} item.

2.2 Judge-Assisted Factual Evaluation under Budgeted Annotation

While FAR from human annotation is usually treated as the ground truth, acquiring human labels is expensive and slow, making it difficult to scale for drawing statistically significant conclusions. On the other hand, Automatic Factual Evaluation can produce judge predictions at scale but the obtained FAR is generally biased. This motivates us to study a budgeted, judge-assisted factual evaluation problem where we combine the abundant judge predictions with limited human labels to form a statistically valid and more efficient estimator.

Given N data points, an LLMaaJ provides predictions, denoted by $f(X)$, for all items, while only $n_{\text{bgt}} \ll N$ items can be sent for human annotation. We adopt Active Statistical Inference (ASI) (Zrnic and Candes, 2024), which selects the labeled subset from an annotation policy, denoted by $\pi(X)$.

Let $\xi_i \sim \text{Bernoulli}(\pi_i)$ indicate whether item i receives a human label with the inclusion probability π_i. ASI estimates FAR by:

$$\hat{\theta}^{\text{ASI}} = \frac{1}{N} \sum_{i=1}^{N} \left[\phi_i^f + \frac{\xi_i}{\pi_i} (\phi_i^h - \phi_i^f) \right],$$

where ϕ^f and ϕ^h denote the judge predicted and human annotated factual correctness respectively.

2.3 Statistical Efficiency

Drawing statistically meaningful conclusions requires not only targeting the correct human-defined metric, but also estimating it with low variance under a limited annotation budget. We compare three

[1] If the LLMaaJ's evidence is the grounding provided to the answer generator, FAR measures faithfulness rate.

Signal family	Representative signals
Judge-output artifacts (F1)	Claim/span structure, verification or hallucination fraction, unverifiable outputs, self-reflection or confidence.
Evidence and reference quality (F2)	Reference support strength, reference conflict, source freshness, direct vs. indirect support.
Task and input strata (F3)	Task or domain type, query category, time-sensitive or location context indicators.
Answer-rubric alignment (F4)	Claim importance, completeness vs. correctness, overprecision or auxiliary details, paraphrase/inference tolerance.

Table 1: Failure-space signal families for residual-risk policy design. The table lists representative pre-labeling signals rather than an exhaustive feature set.

estimators that are central to this paper. The *classical* estimator, $\hat{\theta}^h = \frac{1}{n_{\text{bgt}}} \sum_{i:\xi_i=1} \phi_i^h$, uses only the n_{bgt} human-labeled examples sampled uniformly at random. It is unbiased for the human-defined FAR, but can be noisy under small annotation budget. The *judge-predicted* estimator, $\hat{\theta}^f = \frac{1}{N} \sum_{i=1}^{N} \phi_i^f$, uses judge predictions on all N examples and therefore has low variance, but it is biased when the judge is systematically misaligned with human.

The ASI estimator remains human-anchored. However, its variance is policy dependent and scales with $\mathbb{E}[\Delta^2(1/\pi(X) - 1)]$, where $\Delta = \phi^h - \phi^f$ is the judge-human residual (Zrnic and Candes, 2024). Therefore, efficient annotation requires assigning higher sampling probabilities to examples with larger expected residuals, which motivates the failure-space analysis in the next section.

3 FSA-guided Policy Design

Efficient ASI requires allocating human labels to examples with large expected judge-human residuals. In the oracle case, the optimal policy is proportional to the conditional residual magnitude,

$$\pi^\star(X) \propto \sqrt{\mathbb{E}[\Delta^2 \mid X]}.$$

However, since ϕ^h is unavailable before annotation, the practical policy-design problem is to predict residual risk from pre-labeling information: the query q, response a, reference evidence e, and judge output $f(X)$.

Directly modeling residual risk from raw text is difficult, and generic uncertainty is often insufficient for factuality evaluation. We therefore introduce *failure-space analysis* (FSA), a factuality-specific procedure for deriving structured features $Z(X, f(X))$ that expose where judge-human misalignment is likely to occur. FSA organizes candidate policy signals into four families, summarized in Table 1: *judge-output artifacts*, *evidence and reference quality*, *task and input strata*, and *answer-rubric alignment*. Together, these signals capture not only whether the judge is uncertain, but also whether the evidence is incomplete or stale, whether the query belongs to a difficult slice, and whether the response requires inference or rubric tolerance that the judge may handle differently from humans.

Operationally, FSA uses a historical labeled set where both judge outputs and human annotations are available. We compute the residual target $|\Delta|$, extract numerical features from pre-labeling artifacts, and train a policy model to predict residual risk. The resulting scores are converted into sampling probabilities so that the annotation budget is concentrated on examples where the judge is most likely to disagree with humans. A pipeline overview is summarized and illustrated in Figure 1. We then provide a comprehensive demonstration of this pipeline in Section 4.2 and Section 4.3.

4 Experiments

4.1 Evaluation Settings

We study two factuality-judge settings with different judge architectures, data formats, and judge output structures.

Automatic Factual Accuracy (AutoFA) AutoFA is an internal reference-based factuality evaluation system for voice assistant responses. Given a user query and model response, three reference answers are produced by search-enabled LLMs. The judge decomposes the response into verifiable claims, fact-checks each claim against the references, and produces an overall factuality verdict, cf. SAFE (Wei et al., 2024) and VERISCORE (Song et al., 2024). Both claim-level and response-level outputs can be TRUE, FALSE, or UNKNOWN. We use $N = 2338$ responses and FAR is defined as the fraction of responses rated factually correct among

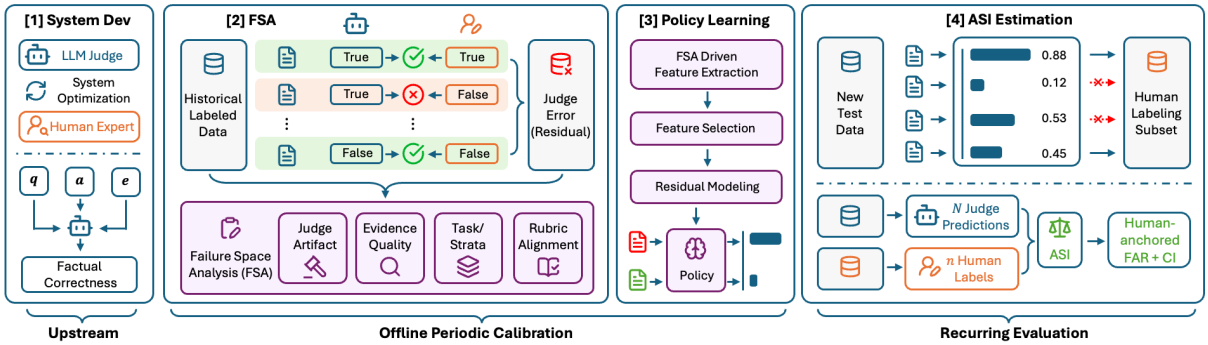

Figure 1: Overview of the pipeline of human anchored factual evaluation. [1] **System Development**: human experts design and iteratively optimize the LLM judge system, which takes the user query (q), model answer (a), retrieved evidence (e) and predicts the factual correctness. We consider this step as upstream and remains fixed once finished. [2] **Failure Space Analysis (FSA) on Calibration Data**: historical labeled data is leveraged to analyze the judge failure modes spreading across four main families. [3] **Policy Learning**: numerical features are extracted based on failure families and are selected with cross-validation. A regression model is trained to predict the risk of records based on the selected features, which is further converted to the sampling policy. High risk records are assigned higher probability to be sampled for human annotation. [4] **ASI estimation**: given new test dataset of size N, n (labeling budget) of them are sampled with the policy and sent for human annotations, while all data receives judge predictions. The ASI estimator combines both judge predictions and human labels and produces human anchored factual accuracy rate (FAR) and confidence interval (CI).

examples with a definitive verdict.

RAGTruth. RAGTruth (Niu et al., 2024) is a public hallucination benchmark with human span-level annotations. We use $N = 2937$ GPT-4-0613 responses across question answering, summarization, and data-to-text generation. We evaluate Claude Sonnet 4.6 as a faithfulness judge using the benchmark prompt, which asks the judge to identify hallucinated spans. For RAGTruth, FAR is the faithfulness rate: the fraction of responses without hallucination.

Self-Confidence Score. Following prior work on confidence-driven inference (Gligorić et al., 2025), we collect a self-reflected confidence score for each record by prompting the judge to assess its confidence in its own previous verdict, yielding a scalar uncertainty signal in $[0, 1]$.

4.2 FSA of Judge Residuals

We first compare judge FAR against human FAR to characterize the residuals that an annotation policy should target. On AutoFA, the judge underestimates human FAR by 13.3 percentage points. On RAGTruth, the gap is larger: the judge FAR is 55.6%, compared with a human FAR of 86.2%.

Typical judge failures. In both datasets, judge-human disagreements are dominated by judge over-rejection. On AutoFA, 90% of disagreements occur when the judge marks human-correct responses as FALSE (43%) or UNKNOWN (46%). Manual inspection identifies two common mechanisms: *evidence gaps*, where references do not directly state a reasonable answer that humans accept, and *overly-strict verification*, where the judge rejects an otherwise correct response due to unsupported peripheral details. On RAGTruth, 94% of disagreements occur when the judge flags hallucinations that humans do not annotate, mainly caused by *overly-literal grounding*: the judge penalizes valid paraphrases or reasonable inferences because they are not explicitly stated in the reference.

Dataset	Stratum	Prev.	$\mathbb{E}[\Delta^2]$	Lift
AutoFA	Partial Verification	0.32	0.41	2.58
	Time-sensitive Query	0.40	0.20	1.29
RAGTruth	Low Self-confidence	0.32	0.78	2.24
	Judge flagged output	0.43	0.70	2.02

Table 2: Residual concentration across observable strata. Prevalence (Prev.) is the fraction of records in the stratum. Lift > 1 indicates above-average residual density.

Non-uniform residual structure. Table 2 shows that residuals concentrate in observable strata derived from judge outputs and task metadata. We report each stratum's prevalence, average squared residual $\mathbb{E}[\Delta^2]$, and *lift*, defined as the ratio between stratum-level and population-level $\mathbb{E}[\Delta^2]$. On AutoFA, partial verification, i.e., cases where at least one extracted claim is marked UNKNOWN, has

Family	Feature	Spearman $\lvert\rho\rvert$ AutoFA	Spearman $\lvert\rho\rvert$ RAGTruth
Judge	self-confidence	0.31	**0.66**
	flag-reliability	–	$0.25^{\dagger}$
Evidence	ref-support-strength	**0.57**	0.22
	ref-temporal-consistency	0.36	–
Strata	time-sensitive / task	0.32	0.18
Rubric	inference-required	0.44	0.16
	qualitative-language	–	0.17

Table 3: Feature ranking by FSA family on absolute Spearman correlation ($|\rho|$) with $|\Delta|$. †Computed on the judge-flagged subset.

$2.58\times$ lift, while time-sensitive queries have $1.29\times$ lift. On RAGTruth, low self-confidence has $2.24\times$ lift, and judge-flagged outputs, i.e., cases where the judge returns one or more hallucination spans rather than an empty list, have $2.02\times$ lift. These concentrations show that judge-human residuals are structured rather than uniformly distributed, motivating the FSA-derived features in Section 4.3.

4.3 Residual-Risk Feature Modeling

Guided by Section 3 and the residual patterns in Section 4.2, we instantiate dataset-specific features from the artifacts available before human annotation. For both datasets, we use the judge's self-confidence score as a generic uncertainty signal.

For RAGTruth, where the judge may over-detect hallucinated spans, we additionally define *flag-reliability* to assess whether the judge-flagged spans are plausibly unsupported by the reference.

For AutoFA, the dominant residuals are tied to reference quality, so we extract *ref-support-strength*, which measures how directly the references support the response's core claims, and *ref-temporal-consistency*, which measures whether search-generated references agree on time-dependent facts. We also include task-level strata, such as time-sensitive queries in AutoFA and task type in RAGTruth.

Finally, we extract rubric-alignment features, including *inference-required*, which measures whether verification requires reasoning beyond literal evidence, and *qualitative-language*, which captures interpretive or subjective phrasing. LLM-extracted features are produced with structured prompts that return integer scores on a 0–3 scale. These feature-extraction calls are applied before human annotation and are substantially cheaper than human labeling in our setting. We provide more details in Appendix E.

Table 3 reports the absolute Spearman correlation between each feature and the residual magnitude $|\Delta|$. This provides a univariate diagnostic of whether each feature monotonically tracks residual risk. On AutoFA, *ref-support-strength* has the strongest correlation, consistent with evidence sufficiency being the dominant failure mode. On RAGTruth, *self-confidence* dominates, while *flag-reliability* provides an additional signal on the judge-flagged subset.

The correlations in Table 3 are diagnostic rather than the final policy criterion. A feature may correlate with $|\Delta|$ but still be less useful for sampling if it is sparse, redundant, or unstable under limited calibration data. We therefore split the calibration set into an inner training split for fitting residual-risk models and an inner validation split for comparing candidate feature configurations. This procedure selects *ref-support-strength* for AutoFA, and *self-confidence* together with *flag-reliability* for RAGTruth, reflecting the different residual structures of the two factuality settings.

4.4 Debiasing and Efficiency Evaluation

We evaluate the method along two dimensions: whether ASI recovers the human-anchored FAR despite judge bias, and whether FSA-derived policies improve annotation efficiency.

Policy learning. Each dataset is split 40/60 into calibration and evaluation sets. On the calibration set, we train a gradient-boosted tree (XGBoost) to predict $|\Delta|$ from the selected features, producing a residual-risk score for each record. The scores are converted into sampling probabilities by normalization with uniform mixing (Li et al., 2025; Zrnic and Candes, 2024), and we apply power-tuning (Angelopoulos et al., 2024) to set the judge contribution in the ASI estimator. Uniform-mixing and power-tuning parameters are estimated on the calibration set, with details in Appendix A. All reported results are evaluated on the evaluation split.

Estimators for comparison. We compare five estimators. *Classical* uniformly samples n_{bgt} human labels and estimates FAR with only human annotations. *Judge-predicted* reports the judge FAR over all N records without human correction. *Uniform ASI* applies the power-tuned ASI correction under a uniform policy. *Confidence ASI* uses only self-confidence as the sampling signal. *FSA ASI* uses the FSA-selected features from Section 4.3. The

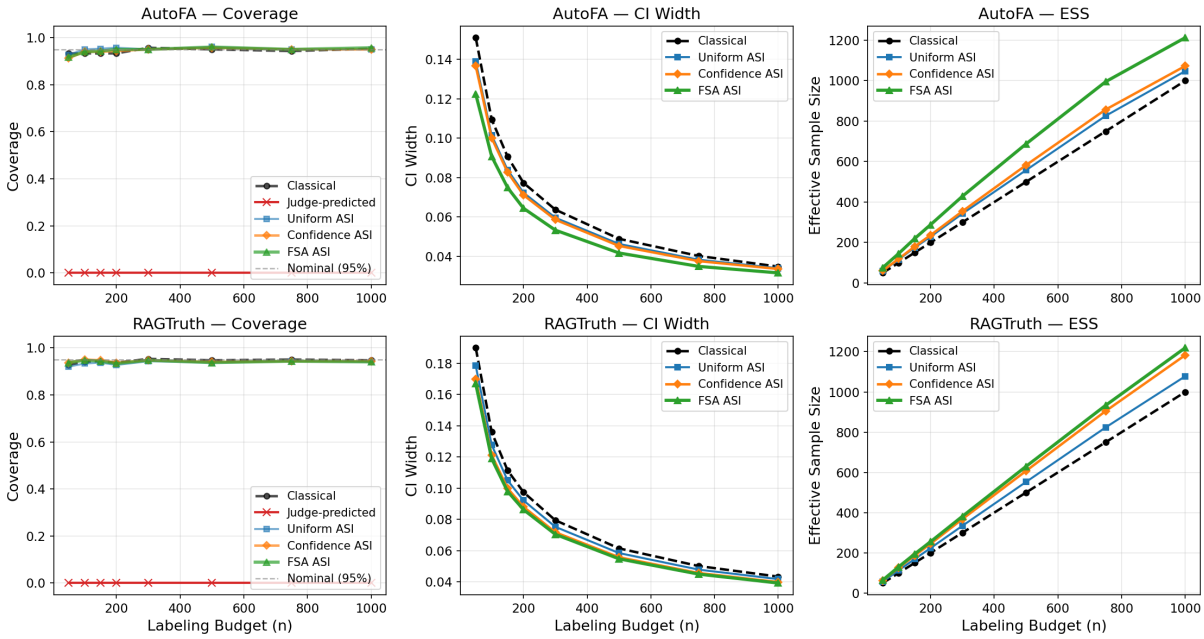

Figure 2: Coverage, CI width, and effective sample size across annotation budgets. FSA ASI maintains valid coverage while achieving the largest effective sample size (ESS) across both datasets.

Dataset	Policy	Cov.	ESS Gain (%)
AutoFA	Classical	$0.94_{\pm 0.01}$	—
	Judge-predicted	0.00	—
	Uniform ASI	$0.95_{\pm 0.01}$	$+13.3_{\pm 4.1}$
	Confidence ASI	$0.94_{\pm 0.01}$	$+17.0_{\pm 4.3}$
	FSA ASI	$0.95_{\pm 0.01}$	$+\mathbf{40.3}_{\pm 9.1}$
RAGTruth	Classical	$0.94_{\pm 0.01}$	—
	Judge-predicted	0.00	—
	Uniform ASI	$0.94_{\pm 0.01}$	$+11.3_{\pm 1.8}$
	Confidence ASI	$0.94_{\pm 0.00}$	$+22.6_{\pm 2.3}$
	FSA ASI	$0.94_{\pm 0.00}$	$+\mathbf{27.1}_{\pm 2.7}$

Table 4: CI simulation results aggregated across budgets $n_{\text{bgt}} \in \{50, 100, \ldots, 1000\}$ ($\alpha = 0.05$, 300 trials per budget). Values report mean $\pm$ standard deviation across budgets.

ASI variants differ only in how they allocate the human-labeling budget.

Metrics and simulation. We report coverage, confidence interval width, and effective sample size (ESS). Coverage is the fraction of confidence intervals containing the human FAR computed on the full evaluation set. ESS measures efficiency on the scale of uniform annotation: an ESS gain of $+40\%$ means that n policy-sampled labels achieve the same variance as $1.4n$ uniformly sampled labels. For each budget $n_{\text{bgt}} \in \{50, 100, \ldots, 1000\}$, we run 300 Monte Carlo trials on the evaluation split. In each trial, we bootstrap the evaluation set, sample human labels according to the corresponding policy, compute the estimator, and construct a Wald confidence interval at $\alpha = 0.05$.

Results. Table 4 and Figure 2 summarize the simulation results. The judge-predicted estimator has zero coverage on both datasets, confirming that judge FAR is biased relative to the human-anchored target. In contrast, all ASI variants maintain near-nominal coverage across budgets, showing that human-labeled correction successfully debiases the judge estimate.

Among the valid estimators, all ASI variants improve efficiency over the classical baseline. Uniform ASI already provides ESS gains of $+13.3\%$ on AutoFA and $+11.3\%$ on RAGTruth while Confidence ASI further improves efficiency to $+17.0\%$ and $+22.6\%$, respectively, by allocating more labels to low-confidence examples.

FSA ASI achieves the largest gains, improving ESS by $+40.3\%$ on AutoFA and $+27.1\%$ on RAGTruth when aggregated across budgets. The larger gap over Confidence ASI on AutoFA reflects that self-confidence is a weaker residual-risk signal in this setting, while FSA identifies reference support strength as a more informative evidence-quality feature. On RAGTruth, the smaller gap is consistent with self-confidence already capturing much of the residual structure, with flag-reliability providing an additional gain.

5 Related Work

Factuality evaluation is central to assessing LLM outputs when correctness must be verified against retrieved evidence, source documents, or external references. Prior work has developed automatic factuality and hallucination evaluation methods that decompose responses into claims, atomic facts, or hallucinated spans, and assess whether these units are supported by evidence (Min et al., 2023; Chen et al., 2023; Niu et al., 2024; Iqbal et al., 2024; Song et al., 2024; Wei et al., 2024). These methods improve scalability, but LLM judges can systematically diverge from human labels due to incomplete evidence, over-literal grounding, rubric mismatch, and other biases in the judging procedure (Zheng et al., 2023; Wang et al., 2024; Gu et al., 2025).

Another line of works including Prediction-Powered Inference and its variants studies how to combine abundant machine predictions with limited human labels while preserving statistically valid inference (Angelopoulos et al., 2023, 2024). Active Statistical Inference extends this idea to budgeted annotation, where examples are sampled according to a policy designed to reduce estimator variance (Zrnic and Candes, 2024). Recent work adapts these ideas to LLM evaluation, showing how LLMaaJ metrics can be reported with valid uncertainty quantification (Gligorić et al., 2025; Wu et al., 2026; Lee et al., 2025). Cost optimal AI evaluation (Angelopoulos et al., 2025) extends ASI by incorporating the cost of human annotator and LLMaaJ into the optimization objective to derive cost-aware annotation policy. Meanwhile, MultiPPI (Cowen-Breen et al., 2026) and AM-PPI (Brawand et al., 2026) explore settings where multiple predictors are available and how to route between predictors to balance the cost-performance trade-off.

Closest to our work, confidence-driven inference (Gligorić et al., 2025) uses verbalized LLM confidence to guide human annotation, which provides a strong generic sampling signal. However, factuality residuals are not always explained by uncertainty alone: judges can be confidently misaligned when evidence is incomplete or only indirectly supportive. Our work adds a factuality-specific policy-design layer: FSA derives residual-risk features from judge artifacts, evidence quality, task strata, and rubric alignment, improving annotation efficiency while preserving the human-anchored ASI target.

6 Limitations

Our proposed framework is effective but is constrained by a few limitations:

First, FSA-guided ASI relies on residual structure learned from historical or calibration data. Its efficiency depends on this structure remaining stable across the evaluation population, i.e., model updates, retrieval or grounding changes, shifts in user queries, or new failure modes can degrade the policy. Silent high-confidence errors that are not captured under distribution shift will be under-sampled by the policy, thus inflate inverse-probability weights and increase variance. Uniform mixing can mitigate but does not remove the need for distribution-shift monitoring and periodic policy recalibration in production deployment. Specifically, a small uniformly sampled audit stream can test whether recent LLM–human residuals remain concentrated in the regions predicted by the policy. If this relationship degrades, the system can increase uniform mixing or revert temporarily to Uniform ASI, while acquiring additional human annotations to recalibrate the policy.

Second, while the four failure families provide useful pre-annotation signals for predicting disagreement in the studied settings, they do not exhaustively cover all error categories. Rare but complex cases, for example, failures in multi-hop reasoning across multiple pieces of evidence, may still occur and may not be captured by FSA. Depending on the application setting, the FSA rubric may therefore need to be further adapted or extended.

Third, our formulation also simplifies the annotation process. We treat the final adjudicated human label as the target and do not explicitly model annotator-level noise or disagreement. We also focus on a single human-annotator rather than a multi-annotator setting with different costs and reliabilities, such as cheap annotators, expert annotators, or multiple judge models. Extending FSA-guided ASI to jointly allocate budget across heterogeneous annotators is an important direction for future work.

Finally, ASI improves efficiency for a given annotation budget, but it does not determine the budget required for a particular production decision. In practice, the choice of n_{bgt} depends on cost, latency, acceptable uncertainty, and release-risk tolerance. Our results quantify the variance reduction achieved by FSA-guided sampling, while budget selection itself remains an operational decision.

References

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. 2023. Prediction-Powered Inference. *Science*, 382(6671):669–674.

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. 2024. PPI++: Efficient Prediction-Powered Inference. *Preprint*, arXiv:2311.01453.

Anastasios N. Angelopoulos, Jacob Eisenstein, Jonathan Berant, Alekh Agarwal, and Adam Fisch. 2025. Cost-Optimal Active AI Model Evaluation. *Preprint*, arXiv:2506.07949.

Nicholas Brawand, Nima Leclerc, Anhthy Ngo, Matthew Peterson, Sriram Vishwanath, Laith Alhussein, and Ben Wellner. 2026. Active Multiple-Prediction-Powered Inference. *Preprint*, arXiv:2605.08429.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. FELM: Benchmarking Factuality Evaluation of Large Language Models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794. ACM.

Charlie Cowen-Breen, Alekh Agarwal, Stephen Bates, William W. Cohen, Jacob Eisenstein, Amir Globerson, and Adam Fisch. 2026. Multiple-Prediction-Powered Inference. *Preprint*, arXiv:2603.27414.

Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel Candès, and Dan Jurafsky. 2025. Can Unconfident LLM Annotations Be Used for Confident Conclusions? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3514–3533, Albuquerque, New Mexico. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A Survey on LLM-as-a-Judge. *Preprint*, arXiv:2411.15594.

Hasan Iqbal, Yuxia Wang, Minghan Wang, Georgi Nenkov Georgiev, Jiahui Geng, Iryna Gurevych, and Preslav Nakov. 2024. OpenFactCheck: A Unified Framework for Factuality Evaluation of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 219–229, Miami, Florida, USA. Association for Computational Linguistics.

Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy-yong Sohn, and Kangwook Lee. 2025. How to Correctly Report LLM-as-a-Judge Evaluations.

Puheng Li, Tijana Zrnic, and Emmanuel Candès. 2025. Robust Sampling for Active Statistical Inference. *Preprint*, arXiv:2511.08991.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

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, Singapore. 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)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. VeriScore: Evaluating the Factuality of Verifiable Claims in Long-form Text Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474, Miami, Florida, USA. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form Factuality in Large Language Models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Skyler Wu, Yash Nair, and Emmanuel J. Candès. 2026. Efficient Evaluation of LLM Performance with Statistical Guarantees. *Preprint*, arXiv:2601.20251.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang,

Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Tijana Zrnic and Emmanuel Candes. 2024. Active Statistical Inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 62993–63010. PMLR.

A Policy Learning Details

This appendix provides experimental details on implementing FSA-policy and running simulations.

Regression target. The policy learning objective is to predict the residual magnitude $|\Delta_i| = |\phi_i^h - \phi_i^f|$ from pre-labeling features. On both datasets, this target is binary: $|\Delta| = 1$ when the judge and human disagree, and $|\Delta| = 0$ when they agree. Figure 3 shows the class distribution. The disagreement rate is 28.2% on AutoFA and 34.7% on RAGTruth.

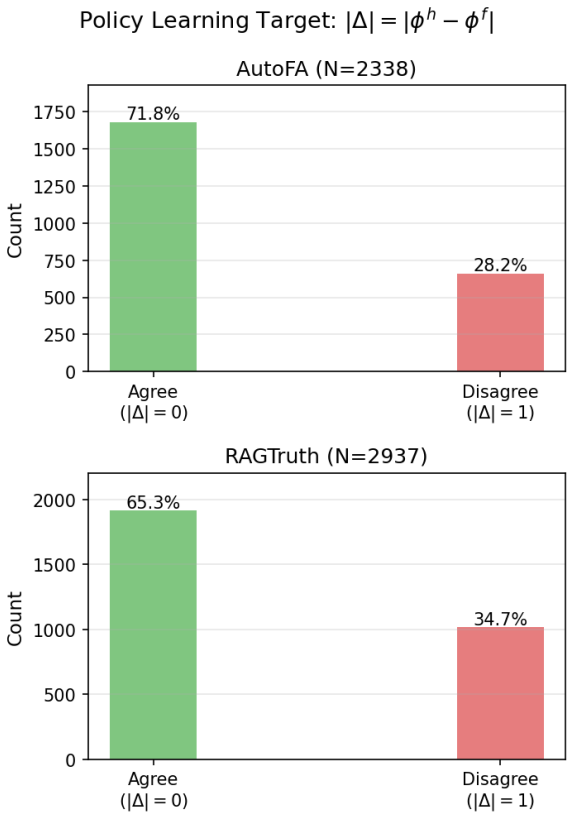

Figure 3: Distribution of the policy learning target $|\Delta| = |\phi^h - \phi^f|$. Both datasets have an imbalanced binary target, with agreement (green) as the majority class.

Score model. We train an XGBoost regressor (Chen and Guestrin, 2016) with `objective=reg:squarederror` to predict $|\Delta|$ from the selected features. On a binary target, this is equivalent to learning $P(\Delta \neq 0 \mid X)$, producing a residual-risk score in $[0, 1]$. We use 200 boosting rounds, learning rate 0.05, and maximum tree depth 3. The model is trained on the 40% calibration split and its output serves as the raw score $s(X_i)$ for policy construction.

Score normalization and uniform mixing. Following the oracle ASI policy in Section 3, we first transform the predicted residual risk as $r(X_i) = \sqrt{s(X_i)}$, and then convert it into inclusion probabilities:

$$\pi_i^{\text{raw}} = r(X_i) \cdot \frac{n_{\text{bgt}}}{\sum_j r(X_j)},$$

with water-filling to ensure $\pi_i \leq 1$. To prevent extreme weights that destabilize the ASI variance estimate, we also apply uniform mixing (Li et al., 2025):

$$\pi_i = (1 - \tau) \cdot \pi_i^{\text{raw}} + \tau \cdot \frac{n_{\text{bgt}}}{N},$$

where $\tau \in [0, 1]$ interpolates between the learned policy ($\tau = 0$) and uniform sampling ($\tau = 1$). Larger τ improves stability at the cost of policy concentration.

Mixing coefficient selection. We select τ by grid search over $\{0.05, 0.10, \ldots, 1.0\}$, evaluating the ASI variance on the calibration split for each candidate. Figure 4 shows that both datasets exhibit a concave ESS-gain curve: too-small τ creates extreme inverse-probability weights that inflate variance, while too-large τ dilutes the policy signal. The selected values are $\tau = 0.40$ for AutoFA and $\tau = 0.55$ for RAGTruth. The higher τ on RAGTruth reflects its larger $\mathbb{E}[\Delta^2]/\mathrm{Var}(\phi^h)$ ratio, which requires a larger mixing coefficient for stability.

Power tuning. The ASI estimator uses a power-tuning coefficient $\lambda \in [0, 1]$ that controls how strongly the judge contribution enters the correction (Angelopoulos et al., 2024). At $\lambda = 1$, the estimator uses the full judge signal; at $\lambda = 0$, it ignores the judge entirely. Power-tuning selects λ to minimize the policy-dependent variance term:

$$J(\lambda) = \mathbb{E}\left[(\phi^h - \lambda\phi^f)^2 \cdot \left(\frac{1}{\pi(X)} - 1\right)\right].$$

Since $J(\lambda)$ is quadratic in λ, the optimum has a closed-form solution:

$$\lambda^* = \frac{\mathbb{E}[\phi^h \phi^f \cdot w]}{\mathbb{E}[(\phi^f)^2 \cdot w]},$$

where $w_i = 1/\pi(X_i) - 1$. In practice, we estimate λ^* from the calibration split using sample averages. Because $J(\lambda)$ is convex, the tuned estimator always has variance no greater than the untuned ($\lambda = 1$) estimator.

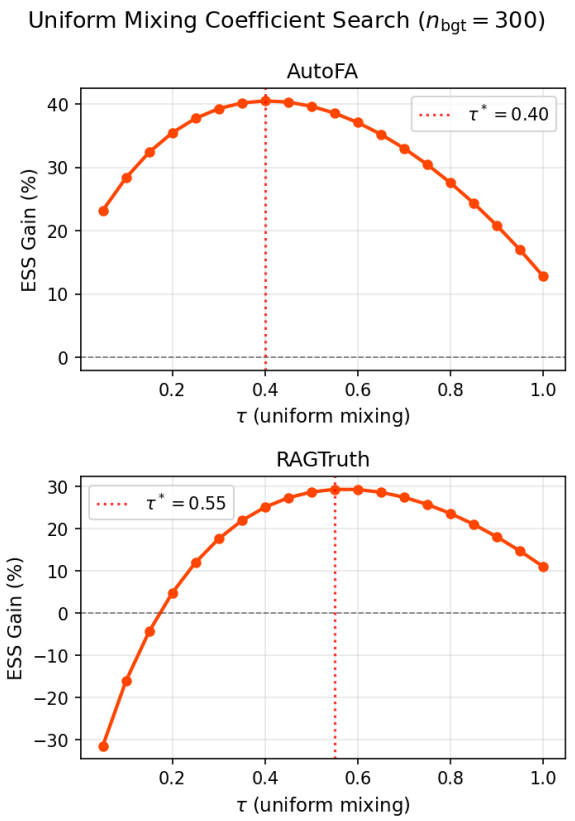

Figure 4: Uniform mixing coefficient (τ) search. ESS gain is evaluated on the calibration split at $n_{\text{bgt}} = 300$. Both curves are concave: small τ overfits, large τ under-concentrates. Red dotted line marks the selected τ^*.

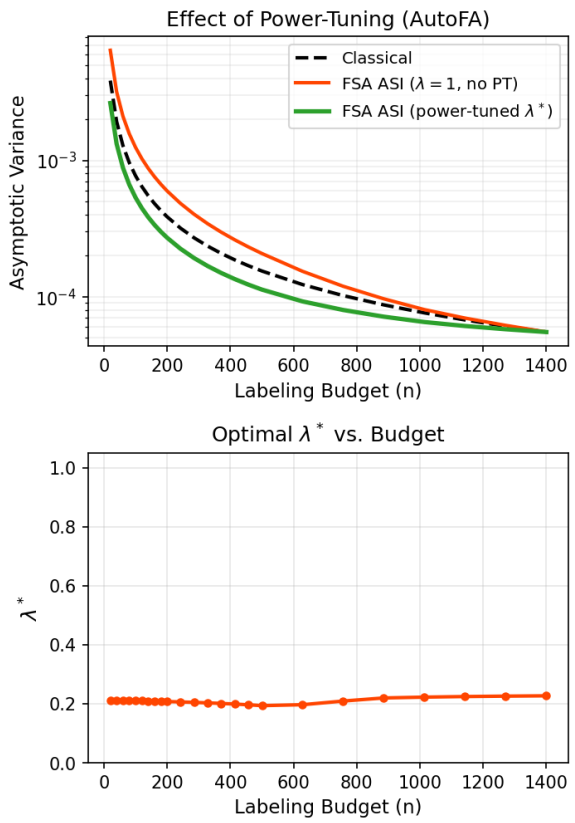

Figure 5: Effect of power-tuning on AutoFA. Left: variance comparison showing that $\lambda = 1$ (untuned) can be worse than baseline, while λ^* (tuned) consistently improves. Right: the optimal $\lambda^* \approx 0.2$ is stable across budgets, indicating the judge should be used at ∼20% strength.

Figure 5 illustrates the effect on AutoFA: without power-tuning ($\lambda = 1$), the ASI variance exceeds the labeled baseline at small budgets because the judge residual is large. Power-tuning ($\lambda^* \approx 0.2$) shrinks the judge contribution to the level most appropriate for the residual structure, consistently reducing variance below the baseline.

B Feature Selection Details

This appendix describes the feature selection procedure and provides post-hoc validation.

Selection procedure. Feature candidates are derived from the qualitative FSA in Section 4.2. We evaluate each candidate's residual-ranking quality via 5-fold cross-validated Spearman correlation on the calibration set. We then select the final feature set using two criteria: (1) high CV Spearman, and (2) parsimony: we prefer a smaller feature set because, with only $N \approx 1000$ calibration samples, using many features can introduce spurious correlations and cause the score model to overfit.

Joint feature importance. Figure 6 shows SHAP values (Lundberg and Lee, 2017) for the score model trained on all candidate features, confirming which features drive the model's predictions. On AutoFA, ref-support-strength dominates. On RAGTruth, self-confidence and flag-reliability jointly dominate. Features from other families contribute marginally.

Post-hoc policy validation. To further evaluate how the selected features translate into effective policies, we compare held-out ESS gain across feature configurations in Table 5. On AutoFA, ref-support-strength ($|\rho| = 0.50$) alone outperforms self-confidence ($|\rho| = 0.12$) alone by a large margin, validating that FSA identifies a stronger signal. On RAGTruth, self-confidence ($|\rho| = 0.63$) is already strong; the addition of flag-reliability provides a modest improvement ($|\rho| = 0.69$). In both cases, adding more features beyond the selected set does not improve ESS, consistent with the observation of overfitting at limited calibration

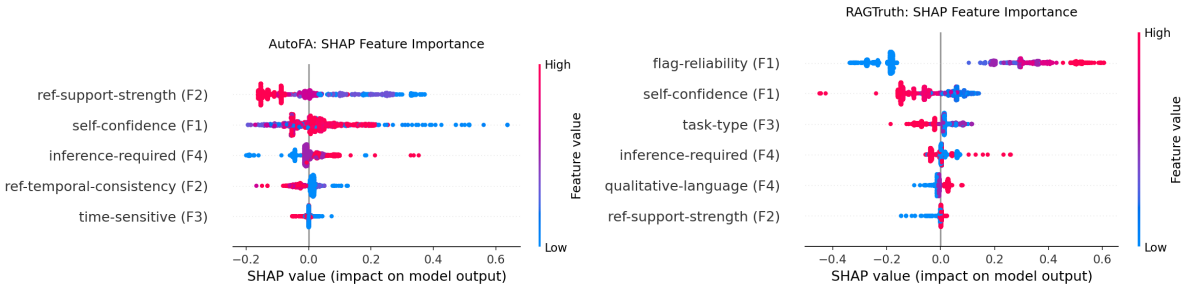

Figure 6: SHAP feature importance for the residual-risk score model trained on all candidate features. Left: AutoFA. Right: RAGTruth.

size.

Dataset	Feature configuration	CV ρ	ESS gain
AutoFA	Self-confidence only	0.12	+16.6%
	Ref-support-strength	**0.50**	**+40.5%**
	All candidates (5)	0.52	+24.7%
RAGTruth	Self-confidence only	0.63	+17.8%
	Self-confidence + flag-reliability	**0.69**	**+19.7%**
	All candidates (6)	0.71	+9.6%

Table 5: Feature selection validation. CV ρ: 5-fold cross-validated Spearman on calibration set. ESS gain: post-hoc evaluation on held-out split at $n_{\text{bgt}} = 300$. Bold: selected configuration. Adding more features improves CV correlation but can reduce ESS due to overfitting.

The key observation is that higher univariate or multivariate correlation does not guarantee better policy performance. On AutoFA, the "all features" model achieves CV $\rho = 0.52$ but only $+24.7\%$ ESS, while the single-feature model achieves $+40.5\%$ ESS. This occurs because the multi-feature model requires more aggressive uniform mixing ($\tau = 0.70$) to stabilize extreme weights, which dilutes the policy concentration. The single feature provides smoother, full-population ranking that translates more effectively into sampling probabilities.

C Human Annotation Details

Both datasets leverage human annotators to acquire reference labels and we summarize the details as follows: RAGTruth (Niu et al., 2024) reports two independent annotators per response and a third review for substantial disagreements, and consistency rates of 91.8% at the response level and 78.8% at the span level. Annotators were recruited through a professional vendor and compensated at \$25 per hour. For AutoFA, all records received a human annotation, and 90% received a second, independent annotation. For records where two annotators disagree with each other on the final response factual verdict, a third human moderator was involved to adjudicate and made the final judgment. The human annotators are instructed to extract core, verifiable claims from the model responses and search the web to judge the factual correctness of the claims and final overall response level factual correctness.

D Qualitative Analysis

We provide two qualitative examples from RAGTruth illustrating LLM judge errors and how FSA identifies them.

In record 168, the response states that Walter Scott was "shot whilst running from Officer Michael Slager." The reference states that Scott was "fatally shot in the back by a police officer" and separately describes video showing Scott running from the officer while the officer fires eight shots. The response therefore combines two closely connected facts into a concise, well-supported statement, but the LLM judge flags "shot whilst running" as hallucinated, apparently requiring the temporal relation to be stated verbatim. Importantly, the LLM judge assigns 0.85 self-confidence, so confidence-based sampling would not identify this as a high-risk record. In contrast, the FSA feature flag-reliability receives a score of 2 (higher means LLM-judge is more likely to be wrong), indicating elevated over-detection risk and allowing the residual-risk policy to prioritize the record for human annotation.

In record 5718, the structured reference lists the same hours, 12:00–17:00, for every day from Monday through Sunday. The response accurately summarizes this as operating "seven days a week from 12:00 to 17:00," yet the LLM judge flags the statement, apparently failing to aggregate the seven

JSON entries. The LLM judge reports a low self-confidence score of 0.30 while the FSA assigns the maximum unreliability (flag-reliability=3).

E Feature Extraction Prompts

This appendix lists the prompts used for self-confidence scoring and FSA feature extraction. All prompts are written as Jinja2 templates: variables enclosed by `{{ }}` are replaced with record-specific content at inference time. The prompts are used only to extract auxiliary features or confidence scores, not to determine the final human label.

E.1 Self-Confidence Prompt

We use the prompt in Figure 7 to elicit the judge's confidence in its own previous verdict. The same structure is used for both datasets, with minor wording changes depending on whether the underlying task is factuality verification or hallucination detection.

E.2 AutoFA Feature Extraction Prompt

The prompt in Figure 8 extracts five FSA features for AutoFA. The judge assesses the verification difficulty of a response given the query and three reference answers. It is explicitly instructed not to make the final factuality decision.

E.3 RAGTruth Feature Extraction Prompt

The prompt in Figure 9 extracts general verification-difficulty features for RAGTruth records. The general verification-difficulty features characterize the response–reference relationship, while *flag_reliability* additionally assesses the judge's flagged spans.

Self-Confidence Score: AutoFA Variant

You previously evaluated whether a model response is factually correct by comparing it against reference answers, and concluded that the response is `{factually correct / factually incorrect}`.

Inputs

- User query: `{{ query }}`
- Model response: `{{ response }}`
- Reference answer 1: `{{ ref_1 }}`
- Reference answer 2: `{{ ref_2 }}`
- Reference answer 3: `{{ ref_3 }}`
- Previous verdict: `{The response is factually correct / incorrect.}`

Instruction
How likely is it that your previous verdict is correct?
Output only a single number between 0 and 1 representing the probability. Do not provide an explanation or reasoning.

Output
`Probability:`

Figure 7: Prompt used for generating self-confidence score.

AutoFA: Evidence and Rubric Features

You are an experienced evaluator in factuality assessment for voice assistant responses.
Your job is not to decide the final label. Your job is to extract a compact set of interpretable features that characterize the difficulty of verifying whether the response is factually correct.
Use only the provided information. Return valid JSON only. Do not provide chain-of-thought. Use integer scores in `{0,1,2,3}`, where higher values indicate more of the corresponding feature.

Feature definitions

1. `reference_temporal_consistency`

   How consistent the three references are on time-dependent facts.

   `0` = fully consistent; `3` = major conflicts on the core answer.

2. `reference_support_strength`

   How strongly the references support the response's core claims.

   `0` = no support; `3` = all references confirm the core answer.

3. `inference_required`

   How much inference is needed to verify the response against the references.

   `0` = pure extractive matching; `3` = substantial reasoning required.

4. `claim_importance_variance`

   How much the claims vary in importance to the core question.

   `0` = all claims are equally important; `3` = core answer plus many peripheral claims.

5. `answer_completeness_vs_correctness`

   Whether the response appears correct but potentially incomplete.

   `0` = comprehensive; `3` = answers only a narrow slice.

Output format

```
{
  "feature_name": {
    "score": int,
    "justification": str
  },
  ...
}
```

Inputs

- User query: `{{ query }}`
- Model response: `{{ response }}`
- Reference answer 1: `{{ ref_1 }}`
- Reference answer 2: `{{ ref_2 }}`
- Reference answer 3: `{{ ref_3 }}`

Figure 8: Prompt used for feature extraction for AutoFA.

RAGTruth: Evidence and Rubric Features

You are an experienced evaluator in groundedness assessment.
Your task is to extract features that characterize the difficulty of verifying whether the response is faithful to the reference. If hallucination spans are provided, also assess whether those flagged spans are plausible.
Use only the provided information. Return valid JSON only. Do not provide chain-of-thought. Use integer scores in `{0,1,2,3}`, where higher values indicate more of the corresponding feature.

Feature definitions

1. `reference_support_strength`: how strongly the reference supports the response's claims. `0` = very weak support; `3` = strong and direct support.
2. `inference_required`: how much inference beyond verbatim extraction is needed to verify the response. `0` = purely extractive; `3` = substantial interpretation.
3. `qualitative_language`: degree of subjective or interpretive language in the response. `0` = purely factual; `3` = heavily qualitative.
4. `flag_reliability`: if hallucination spans are provided, how likely the judge is correct about those flags. `0` = very likely correct; `3` = likely incorrect due to over-detection. If no spans are provided, set this score to `0` and state that no flagged spans were provided.

Output format
Return a JSON object of the following form:

```
{"feature_name": {"score": int, "justification": str}, ...}
```

Inputs

- Task type: `{{ task_type }}`
- Reference: `{{ reference }}`
- Response: `{{ response }}`
- Judge's flagged spans: `{{ spans }}`

Figure 9: Prompt used for feature extraction for RAGTruth