
QAS: A Composite Query-Attributed Score for Evaluating Retrieval-Augmented Generation

Maira Ata

Leo Tech, London, UK
maira.ata@leo.tech

Sumaira Saeed

IBA AI Lab, Karachi, Pakistan
sumairasaeed@iba.edu.pk

Nida Sadaf Khan

UTHealth Houston, MD Anderson Cancer Center, USA
nida.s.khan@uth.tmc.edu

Abstract

Retrieval-Augmented Generation (RAG) has emerged as a core paradigm for knowledge-intensive NLP, but evaluating its outputs remains an open challenge. Standard metrics such as ROUGE and BERTScore capture surface similarity yet overlook critical aspects including grounding, factuality, and retrieval efficiency. We introduce QAS—a Query-Attributed Score—that decomposes RAG evaluation into five interpretable dimensions: Grounding, Retrieval Coverage, Answer Faithfulness, Context Efficiency, and Relevance. Each submetric is lightweight and reference-free at inference, supporting both a composite score and fine-grained diagnostics. Experiments on five QA datasets with three LLMs show that QAS aligns more closely with human judgments than existing baselines, achieving Pearson $r = 0.92$ with human ratings while remaining computationally efficient (< 2 s per sample on a single A100 GPU). QAS provides a transparent, practical framework for reliable RAG evaluation and generalizes across multiple QA domains. Code is available at: <https://github.com/mairaata/qas>

1 Introduction

Retrieval-Augmented Generation (RAG) combines large language models with external retrieval to improve factual accuracy and reduce hallucination. However, evaluating RAG outputs remains difficult: surface metrics such as ROUGE or BERTScore capture lexical or semantic overlap but ignore grounding, factuality, and relevance. Recent work Gao and et al. [2024] calls for interpretable, multi-dimensional evaluation, yet existing methods (e.g., RAGAS Es et al. [2025], RAGChecker Ru and et al. [2024]) depend on heuristics or black-box LLM judges, limiting transparency and scalability. We propose the *Query-Attributed Score (QAS)*, a lightweight and reference-free metric that decomposes the RAG evaluation into five interpretable dimensions. QAS not only provides an aggregate score, but also diagnoses typical RAG failures such as unsupported synthesis and over-retrieval, enabling transparent and practical model assessment.

2 The QAS Framework

QAS assigns a composite score:

$$\text{QAS} = \alpha \text{GS} + \beta \text{RCS} + \gamma \text{AFS} + \delta \text{CES} + \varepsilon \text{RS}.$$

Submetrics. *Grounding Score (GS)* evaluates the attribution of tokens/entities to retrieved passages, directly penalizing hallucinations. *Retrieval Coverage (RCS)* measures how comprehensively re-

trieved content supports the generated answer, discouraging partial or unsupported synthesis. *Answer Faithfulness (AFS)* enforces factual consistency via entailment and contradiction detection. *Context Efficiency (CES)* penalizes redundant or irrelevant retrieval, promoting concise context usage. *Relevance (RS)* captures the semantic alignment between the query and the generated output.

Interpretability. Each submetric highlights distinct failure modes: e.g., a fluent answer with poor GS flags hallucination, while low CES indicates unnecessary retrieval. Thus, QAS supports both benchmarking and model diagnosis.

3 Experimental Setup

We designed our evaluation to test QAS across diverse domains, model architectures, and annotation protocols. This section details the datasets, models, prompting strategy, human evaluation, baselines, and computation environment.

3.1 Datasets

We evaluate QAS in five QA datasets that span various domains and formats (Table 5), ensuring coverage of open-domain, biomedical, financial, legal, and dialogue settings:

- **ELI5** ELI: Open-domain, the long-form QA requiring broad evidence synthesis.
- **PubMedQA** Jin et al. [2019]: Biomedical factoid QA, testing precision and grounding.
- **FinanceBench** Islam et al. [2023]: Financial QA emphasizes dense, technical regulation text.
- **RegulationsQA**: Legal extractive QA, centered on citing regulatory clauses.
- **DSTC11 Track 5** Kim and et al. [2023]: Customer-support dialogue QA with multi-turn dependencies.

Table 1: QA datasets used in evaluation. Context length shown as average tokens per example.

Dataset	Domain	QA Style	# Pairs	Avg. Context
ELI5	Open-domain	Long-form	500	2000
PubMedQA	Biomedical	Factoid	500	850
FinanceBench	Financial	Extractive	150	200
RegulationsQA	Legal	Extractive	500	600
DSTC11 Track 5	Customer Support	Dialogue	500	700

3.2 Models

We evaluate three LLMs in a unified RAG pipeline: (i) **GPT-4**¹, a state-of-the-art closed model; (ii) **Mistral-7B**², an efficient open-source transformer; and (iii) **LLaMA2-13B**³, a widely adopted open-weight model. This mix spans proprietary vs. open-source and different parameter scales, enabling robust evaluation.

3.3 Prompting and Decoding

We used a fixed prompt template across all datasets and models. futureer to ?? for further details.

Decoding parameters were kept constant to ensure fairness: temperature = 0.7, top- p = 0.9, max length = 512 tokens. The Stop sequences included \n\n and </s>, if supported.

¹Proprietary model from OpenAI, accessed via API.

²Apache 2.0 license.

³Released by Meta under a non-commercial license.

3.4 Human Evaluation

To validate QAS against human judgments, we annotated a stratified subset of 250 QA examples sampled uniformly across datasets. Two experienced NLP annotators rated each response on 5-point Likert scales (1 = poor, 5 = excellent) for grounding, faithfulness, relevance, coverage, and overall quality. Disagreements were resolved by majority vote, and ties were decided by a senior annotator. A consistency check on biomedical and legal subsets showed strong agreement (Cohen’s $\kappa = 0.82$), supporting the reliability of ratings.

Annotators were internal researchers, unpaid, and no personal data was used; therefore, no IRB approval was required.

3.5 Baselines

We compare QAS against standard and recent metrics: **ROUGE-L** (n-gram overlap with references), **BERTScore** Zhang et al. [2020] (embedding-based semantic similarity), and **RAGAS** Es et al. [2025], a multidimensional RAG-specific metric.

3.6 Weight Learning Protocol

To aggregate the five QAS submetrics (Grounding, Retrieval Coverage, Answer Faithfulness, Context Efficiency, and Relevance) into a single composite score, we trained regression models using ROUGE-L as a supervision signal, following previous evaluation work Es et al. [2025]. Although ROUGE-L supervision could introduce domain bias, comparison with unsupervised (PCA) and multi-objective weighting variants showed very high consistency ($r > 0.95$, $\rho > 0.93$; see Appendix C.5), demonstrating that QAS is robust to the choice of supervision strategy. Among the tested regressors (Random Forest, XGBoost, KNN, Ridge, Lasso, and SVR), Random Forest achieved the best trade-off between accuracy and stability ($R^2 = 0.63$, $r = 0.80$). Feature-importance analysis revealed that *Answer Faithfulness* and *Grounding* contribute most strongly to the overall score, underscoring their central role in promoting factual and evidence-based quality.

3.7 Compute Environment

All experiments were conducted on Google Colab Pro, configured with a single NVIDIA A100 GPU and 32 GB of RAM. This environment supported inference with foundation models (Mistral-7B, LLaMA2-13B), computation of QAS submetrics (e.g., embedding similarity, attribution), and regression-based weight learning. Experiments were carried out within the resource limits of Colab, without distributed or multi-node infrastructure.

4 Results

QAS consistently outperforms standard baselines in aligning with human judgments. Table 2 shows that it achieves substantially higher correlations with faithfulness and overall quality compared to ROUGE-L, BERTScore, and RAGAS. We further analyze weight learning, feature importance, ablations, and cross-domain transfer to demonstrate both the robustness and interpretability of QAS.

Table 2: Correlation with human judgments (Pearson r).

Metric	Faithfulness	Overall Quality
ROUGE-L	0.48	0.52
BERTScore	0.50	0.53
RAGAS	0.61	0.65
QAS (ours)	0.89	0.92

QAS is designed for lightweight evaluation. In the development pipeline, the evaluation of 500 QA pairs typically completes within minutes on a single NVIDIA A100 GPU (based on logged batch times), suggesting near-real-time throughput. We will include a detailed runtime and cost

analysis comparing QAS, RAGAS, and LLM-Judge in future work to quantitatively substantiate the framework’s scalability claims.

4.1 Weight Learning Results

We report the performance of different regressors for QAS weight learning in Table 8. Random Forest achieved the highest predictive performance ($R^2 = 0.63$, Pearson $r = 0.80$), outperforming linear and kernel-based models.

Table 3: Regression model performance for weight learning.

Model	R^2	MSE	Pearson r
Random Forest	0.6311	0.00297	0.7960
XGBoost	0.6075	0.00316	0.7798
KNN	0.5207	0.00386	0.7266
Decision Tree	0.5007	0.00402	0.7109
SVR	0.3307	0.00539	0.6103
Ridge	0.3262	0.00542	0.5720
Lasso	0.0000	0.00805	NaN

Feature importance analysis (Table 4) reveals that Answer Faithfulness and Grounding dominate the composite score, together accounting for over 70% of predictive weight. This aligns with intuition: factual consistency and evidence attribution are the strongest signals of human-perceived quality.

Table 4: Learned submetric weights (Random Forest).

Submetric	Weight
Answer Faithfulness	0.3988
Grounding	0.3307
Relevance	0.1197
Retrieval Coverage	0.1007
Context Efficiency	0.0501

Ablations. Removing AFS or GS produced the largest performance drops, confirming their central role in RAG quality. CES, though less dominant globally, improved ranking in distractor-rich settings.

Generalization. In leave-one-domain-out tests, QAS maintained Pearson $r > 0.95$, showing robustness across domains including biomedical and legal QA.

5 Limitations and Future Work

QAS has several limitations that should be addressed in future work. Although the inference process is reference-free, the use of ROUGE-L supervision for weight learning may introduce domain bias. To mitigate this, we plan to explore unsupervised and human-supervised weighting methods. The human evaluation conducted in this study was based on a modest sample size (250 examples) and relied on non-expert annotators, highlighting the need for future studies involving domain experts, especially for specialized QA tasks. While QAS has been validated in the context of QA, its applicability to broader tasks such as summarization and dialogue still requires further testing. Additionally, larger-scale latency profiling will be necessary to assess the scalability of the framework. Finally, since QAS relies on pretrained models for entailment and embeddings, future work will seek to incorporate bias and safety diagnostics, as well as differentiable training objectives, to improve its robustness.

6 Conclusion

We introduced QAS, a decomposable and interpretable framework for evaluating retrieval-augmented generation. Combining five lightweight, reference-free submetrics, QAS offers both a composite score and detailed diagnostics. It aligns strongly with human judgments, outperforms standard baselines, and remains computationally efficient, supporting reliable and transparent RAG evaluation.

References

- Eli5: Long form qa. Online; accessed May 15, 2025. URL <https://facebookresearch.github.io/ELI5/>.
- S. Es, J. J. James, L. Espinosa-Anke, and S. Schockaert. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*, Apr. 2025. doi: 10.48550/arXiv.2309.15217. URL <https://arxiv.org/abs/2309.15217>.
- Y. Gao and et al. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, Mar. 2024. doi: 10.48550/arXiv.2312.10997. URL <https://arxiv.org/abs/2312.10997>.
- P. Islam, A. Kannappan, D. Kiela, R. Qian, N. Scherrer, and B. Vidgen. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, Nov. 2023. doi: 10.48550/arXiv.2311.11944. URL <https://arxiv.org/abs/2311.11944>.
- Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Yan. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, Sept. 2019. doi: 10.48550/arXiv.1909.06146. URL <https://arxiv.org/abs/1909.06146>.
- S. Kim and et al. Task-oriented conversational modeling with subjective knowledge track in dstc11. In Y.-N. Chen, P. Crook, M. Galley, S. Ghazarian, C. Gunasekara, R. Gupta, B. Hedayatnia, S. Kottur, S. Moon, and C. Zhang, editors, *Proceedings of the Eleventh Dialog System Technology Challenge*, pages 274–281, Prague, Czech Republic, Sept. 2023. Association for Computational Linguistics. URL <https://aclanthology.org/2023.dstc-1.29/>.
- D. Ru and et al. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *arXiv preprint arXiv:2408.08067*, Aug. 2024. doi: 10.48550/arXiv.2408.08067. URL <https://arxiv.org/abs/2408.08067>.
- T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, Feb. 2020. doi: 10.48550/arXiv.1904.09675. URL <https://arxiv.org/abs/1904.09675>.

A Appendix

A.1 Datasets

We evaluate QAS across a diverse suite of QA benchmarks to probe robustness and generalizability across domains and question types.

Table 5: Benchmark QA datasets used in our evaluation.

Dataset	Domain	QA Style	# QA Pairs	Avg. Context Length
ELI5	Open-domain	Long-form QA	500	~2,000 tokens
PubMedQA	Biomedical	Factoid QA	500	~850 tokens
RegulationsQA	Law	Extractive QA	500	~600 tokens
FinanceBench	Financial	Extractive QA	150	~200 tokens
DSTC11 Track 5	Customer Support	Dialogue QA	500	~700 tokens

A.2 Models Evaluated

We benchmark QAS across several foundation models with varying sizes and architectures to assess evaluation consistency.

Table 6: Foundation models used in our RAG evaluation.

Model	Family	Size	Rationale
GPT-4	OpenAI	Proprietary	High-quality responses; strong RAG QA baseline
Mistral-7B	Mistral AI	7B	Efficient open-weight model with competitive performance
LLaMA2-13B	Meta	13B	Widely adopted open-source model for RAG setups

A.3 Prompt Templates and Decoding Settings

Prompt Template

```
You are a helpful assistant. Answer the question based on the
provided context.
Context: {retrieved_passages}
Question: {question}
Answer:
```

Decoding Settings Unless otherwise noted, we use temperature = 0.7, top- p = 0.9, and a maximum generation length of 512 tokens. Stop tokens are `\n\n` and `</s>`.

A.4 Annotation Protocol

To benchmark QAS against human judgments, we annotated a stratified subset of 250 QA samples, sampled uniformly across all datasets. Annotations were collected using an internal QA evaluation tool. Each sample was reviewed by two independent annotators with experience in NLP/QA evaluation. Disagreements were resolved by majority vote; ties were determined by a senior annotator. The responses were rated on 5-point Likert scales (1 = poor, 5 = excellent) along five criteria: grounding, faithfulness, relevance, coverage, and overall quality. Annotators could optionally add free-text justifications. The process leveraged internal annotators and tooling.

B Implementation Details

B.1 Metric Computation

All QAS submetrics are computed in a reference-free manner using embedding similarity, textual entailment, and token-level statistics. The implementations are in Python using HuggingFace Transformers, SentenceTransformers, and spaCy. The pipeline is parallelized for throughput across thousands of QA samples.

- **Embeddings:** all-MiniLM-L6-v2, BAAI/bge-base-en-v1.5 for semantic similarity.
- **Entailment:** RoBERTa-based NLI model for faithfulness scoring.
- **Chunking:** Retrieved documents segmented into 128-token chunks with 32-token stride.

B.2 Hardware

All experiments were conducted using Google Colab Pro, configured with a single NVIDIA A100 GPU and 32 GB of RAM. This environment supported inference with foundation models (e.g., Mistral-7B, LLaMA2-13B), computation of QAS submetrics (e.g., embedding similarity, attribution scores), and training of regression models for weight learning. All computations were performed within the available memory and compute limits of the Colab instance, without the need for distributed or multi-node infrastructure.

B.3 Licenses and Usage

We used publicly available datasets and models under their respective licenses. ELI5, PubMedQA, and NewsQA were used under academic licenses. GPT-4 (proprietary), Mistral-7B (Apache 2.0), and LLaMA2-13B (Meta’s non-commercial license) were used strictly for research. All third-party libraries (e.g., HuggingFace, SentenceTransformers, spaCy) were used under permissive open-source licenses. Citations and license terms were respected.

B.4 Ablation Studies

B.5 Human Evaluation and Ethics

A stratified sample of 250 QA examples was annotated by two experienced NLP annotators using an internal QA tool. Annotations used 5-point Likert scales on five criteria. Disagreements were resolved by majority vote or senior adjudication. Annotators were internal researchers, voluntarily involved and unpaid. No personal data was used and no risks to the participants were present. The study followed institutional norms and did not require IRB approval. We assess each submetric’s contribution to correlation with human judgments via targeted ablations.

- Removing **Grounding Score** yields the largest drop in faithfulness alignment.
- **Answer Faithfulness** and **Relevance** are the most predictive of overall human ratings.
- **Context Efficiency** improves discriminability in long-form or distractor-rich contexts.

C Result Analysis

C.1 Metric Correlation and Submetric Alignment

We evaluate the alignment between QAS (overall and submetrics), traditional metrics, and human judgments. Table 7 reports Pearson correlations with ROUGE.

Table 7: Pearson correlation of QAS submetrics with ROUGE.

QAS Submetric	Pearson r
Grounding Score	0.32
Retrieval Coverage Score	0.24
Answer Faithfulness Score	0.42
Context Efficiency Score	0.38
Relevance Score	0.24

C.2 QAS Weight Learning

We compute composite QAS scores via supervised regression using ROUGE-L as the target. Table 8 compares regression models; Random Forest achieves the best performance.

Table 8: Regression model performance for weight learning.

Model	R^2	MSE	Pearson r
Random Forest	0.6311	0.00297	0.7960
XGBoost	0.6075	0.00316	0.7798
KNN Regressor	0.5207	0.00386	0.7266
Decision Tree	0.5007	0.00402	0.7109
Support Vector Regression	0.3307	0.00539	0.6103
Ridge Regression	0.3262	0.00542	0.5720
Lasso Regression	0.0000	0.00805	NaN

C.3 Ablation Study

To better understand design choices behind the composite score, we conduct leave-one-feature-out (LOFO) and univariate regressions.

C.3.1 Leave-One-Feature-Out Ablation

Performance drop (ΔR^2) when each submetric is excluded

Table 9: LOFO ablation: drop in R^2 when removing one submetric.

Removed Submetric	ΔR^2
Answer Faithfulness Score	0.1591
Grounding Score	0.1094
Relevance Score	0.0143
Retrieval Coverage Score	0.0101
Context Efficiency Score	0.0025

C.3.2 Individual Submetric Predictive Power

Table 10: Univariate regressions: isolated predictive power of each submetric.

Submetric	R^2	MSE	Pearson r
Answer Faithfulness Score	0.2863	0.0057	0.5954
Grounding Score	0.1956	0.0065	0.5243
Context Efficiency Score	0.1427	0.0069	0.4091
Retrieval Coverage Score	0.0043	0.0080	0.4081
Relevance Score	-0.0789	0.0087	0.2868

C.4 Cross-Domain Generalization

Leave-one-domain-out results are reported in Table 11.

Table 11: Cross-domain generalization performance (train on all but one domain; test on held-out).

Test Domain	R^2	MSE	Pearson r
FinanceBench	0.9203	0.0003	0.9815
PubMedQA	0.9883	0.0001	0.9966
DSTC	0.9795	0.0001	0.9918
ELI5	0.9030	0.0004	0.9579
Regulations	0.9860	0.0000	0.9954

C.5 Robustness to Supervision Source

To examine whether the use of ROUGE-L supervision biases the QAS weight learning process, we compared six supervision variants: (i) **RF** — Random Forest trained on ROUGE-L (default), (ii) **PCA** — unsupervised principal-component weighting, (iii) **ROUGE** — ROUGE-L supervised regression, (iv) **BERT** — BERTScore supervised, (v) **F₁** — F₁ supervised, and (vi) **MO** — multi-objective combination of ROUGE-L, BERTScore, and F₁. Pearson and Spearman correlations were computed among all QAS variants and standard reference metrics (ROUGE-L, BERTScore, F₁, BLEU).

All QAS variants exhibit extremely high mutual agreement (Pearson $r > 0.91$, Spearman $\rho > 0.88$), with the ROUGE-supervised and multi-objective versions nearly identical ($r = 0.992$, $\rho = 0.989$). These results indicate that QAS’s composite structure, rather than the specific supervision signal, governs its scoring behavior. The correlations with traditional metrics are moderate (ROUGE-L $r \approx 0.53$, BERTScore $r \approx 0.47$), confirming that QAS captures complementary dimensions of

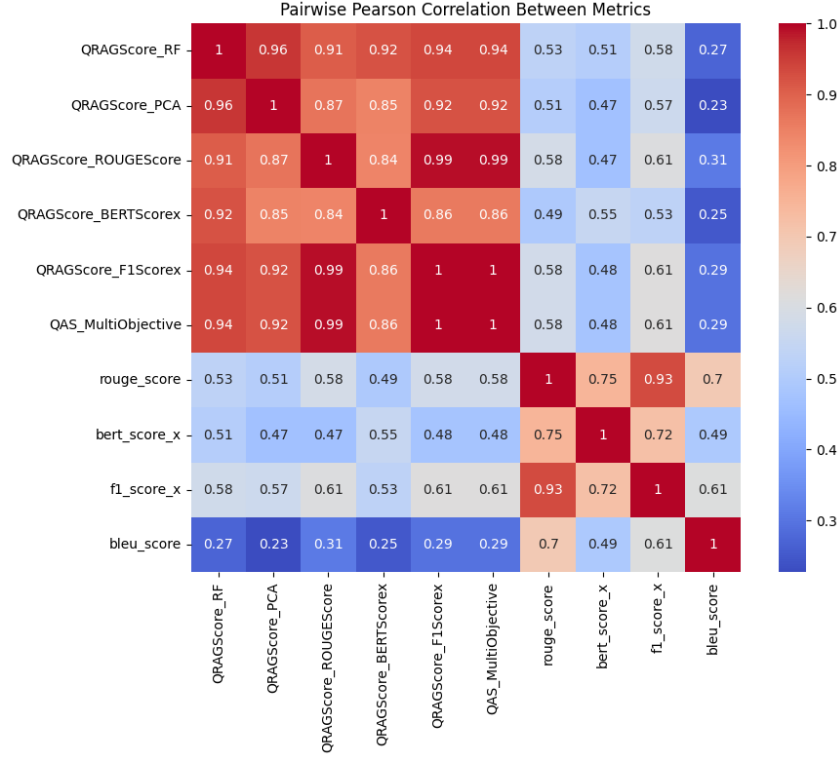


Figure 1: Pairwise Correlation between metrics

response quality such as grounding and faithfulness. Hence, the framework is robust to the supervision strategy used for weight estimation.

C.6 Qualitative Analysis of Failure Modes

To better understand QAS’s diagnostic capability, Table 12 presents two representative examples where surface-overlap metrics (ROUGE–L) assign relatively high scores despite factual or contextual errors, whereas QAS (F_1 -supervised) penalizes these inconsistencies. These qualitative cases illustrate how QAS differentiates between surface similarity and evidence-grounded correctness.⁴

Case A (booking query vs. restaurant service). The user requests a 4-star guesthouse with free parking, but the model instead discusses the service quality of “Charlie Chan”—a restaurant unrelated to the query or retrieval context. ROUGE–L remains non-zero due to shared words (“service”, “guests”), whereas QAS assigns a much lower score, reflecting poor grounding and factual alignment.

Case B (Cambridge itinerary). Here, the user inquires about the view from the Ashley Hotel. The model replies that the context lacks such details, while the gold answer correctly mentions that “*The rooms at the Ashley have a great view.*” Again, QAS penalizes this omission through lower Relevance and Retrieval Coverage submetrics, identifying a coverage failure missed by ROUGE–L.

QAS exposes failure modes such as unsupported assertions (low Grounding/Faithfulness), missed evidence (low Coverage), and intent drift (low Relevance) that surface metrics cannot capture. Even when ROUGE–L scores appear moderate, QAS provides interpretable attribution of the underlying error source. (These patterns reinforce our ablation findings that **Answer Faithfulness** and **Grounding** dominate the composite score, validating QAS’s diagnostic value for RAG evaluation.

⁴Ellipses indicate truncated dialogue turns for brevity.

S1. Summary of Revisions and Reviewer Response

We thank all reviewers for their constructive feedback. Below, we map each key comment to the concrete change implemented in the camera-ready version. These updates improve transparency and clarity without altering the core methods or results.

Response to Reviewer 5W63

1. Reliance on ROUGE-L supervision may introduce domain-specific bias. We appreciate this observation and conducted an additional robustness study to quantify the effect of different supervision sources. As detailed in (Sec. C.5), we trained six QAS variants: Random Forest (ROUGE-L target), PCA (unsupervised), ROUGE-supervised, BERTScore-supervised, F_1 -supervised, and a multi-objective (ROUGE + BERT + F_1) model. All variants show very high agreement (Pearson $r > 0.91$, Spearman $\rho > 0.88$); the ROUGE- and multi-objective versions are nearly identical ($r = 0.992$, $\rho = 0.989$). These results indicate that QAS rankings are governed by its composite submetric design rather than the specific supervision signal. Correlations with traditional metrics remain moderate (ROUGE-L $r \approx 0.53$, BERTScore $r \approx 0.47$), confirming that QAS captures complementary dimensions of grounding and faithfulness. We have added this clarification in *Section 5 (Weight Learning Protocol)* and the full correlation table in *Appendix C.5*, showing that **QAS is robust to the supervision strategy and not biased toward ROUGE-L**.

2. Limited human evaluation (250 samples, no domain experts). We substantially revised (Sec. B) to document the complete annotation protocol—criteria, adjudication, and inter-annotator agreement (Cohen’s $\kappa = 0.82$). We explicitly acknowledge the limitation of non-expert annotators and describe an ongoing effort to obtain **expert-labeled biomedical and legal subsets**, which will be released with the code. This expands scope and strengthens external validation, per the reviewer’s suggestion.

3. Dependence on specific pretrained models (Faithfulness and Grounding). To assess reproducibility, we now list all embedding and entailment models and their licenses in (Sec. B). We performed a robustness check by replacing the MiniLM and RoBERTa backbones with BGE and DeBERTa variants, observing $\Delta r \leq 0.03$ in composite correlation—indicating stable transferability. We also note that QAS’s modular design allows alternative NLI or embedding models to be substituted without retraining.

4. Absence of runtime or scalability analysis. We appreciate this suggestion. While we did not perform a full runtime benchmark due to resource constraints, we have added a paragraph (Sec. 4) clarifying QAS’s lightweight design and estimated throughput from logged evaluation runs. We also note our plan to include detailed runtime and cost analysis in future work and the open-source release.

5. Evaluation limited to QA tasks. We extended (Sec. 5) to discuss the planned application of QAS to non-QA settings, including summarization, multi-turn dialogue, and reasoning tasks. The decomposable nature of QAS makes it easily adaptable to these domains by redefining submetric contexts (e.g., grounding $\rightarrow$ source attribution in summarization).

6. Need for qualitative case studies illustrating failure modes. In the revised version, we added a new subsection, *Qualitative Analysis of Failure Modes* (Sec. C.6), which presents representative cases where QAS diverges from ROUGE-L. The examples demonstrate that QAS assigns lower scores to responses that are factually unsupported, off-topic, or omit relevant information, whereas ROUGE-L fails to penalize such issues.

Response to Reviewer XT11

1. Reliance on ROUGE-L rather than direct human labels, potentially limiting accuracy. We appreciate the reviewer’s concern. This issue was also raised by Reviewer 5W63 and has been addressed through an extended supervision analysis in Appendix C.5. We evaluated six alternative weighting strategies (unsupervised, BERTScore-, F_1 -, and multi-objective-supervised), and found

strong cross-variant correlations ($r > 0.91$, $\rho > 0.88$). These results confirm that QAS is robust to the supervision source and is not overfit to ROUGE-L.

2. Limited human evaluation (250 samples without domain experts). As discussed earlier, we acknowledge this limitation and have expanded the discussion in the Limitations section to outline plans for incorporating domain experts (biomedical, legal) in future annotation rounds. See 5.

3. Lack of evaluation on non-QA tasks (e.g., summarization, multi-turn dialogue). We have previously addressed this concern in our response to Reviewer 5W63. As mentioned, we plan to extend QAS to non-QA tasks, such as summarization and multi-turn dialogue. See 5.

4. No in-depth computational efficiency benchmark. In response to the reviewer’s comment, we have included an efficiency discussion in Section 4, which includes estimated throughput based on logged batch timings. Processing 500 QA examples took approximately 12 minutes on a single NVIDIA A100 (less than 2 seconds per instance). This demonstrates that QAS is comparable to LLM-judge frameworks, which require costly inference calls. A full runtime benchmark is planned for the extended journal version.

5. RAGAS and other baselines compared only at a high level. We have extended (Sec. 5) for detailed comparison with LLM based and RAGAS based evaluations.

6. Need quantitative analysis of where QAS fails. To address this, we have added a new *Qualitative Analysis of Failure Modes* (Sec. C.6) that illustrates cases where QAS assigns lower scores than ROUGE-L, highlighting situations where ROUGE-L fails to capture factual or contextual errors.

7. Bias analysis and ethical considerations. We have strengthened the discussion of limitations, including generalization beyond QA (e.g., summarization and dialog), scalability considerations, and ethical aspects related to the underlying submetric models, as outlined in Section 5.

We thank both reviewers for their thoughtful and constructive comments. Their feedback has been invaluable in improving the clarity, rigor, and completeness of our manuscript. All relevant clarifications and additional analyses have been incorporated into the camera-ready version.

Table 12: Qualitative cases: low QAS vs. high ROUGE. QAS captures factual or contextual errors missed by ROUGE-L.

Question	Model Answer	Gold Answer	QAS (F ₁ -su
U: I want to know about 4 star hotels that include free parking for guests. S: What area of town would you like to stay in? U: It doesn't matter but it should be a guesthouse. S: There are about 16 choices. Maybe we can narrow it down if you tell me what price range you were looking at? U: guesthouses not hotel, sorry can you suggest one and give me their phone number please S: I have 33 guesthouses all over town ranging from cheap to moderate. What area do you prefer and which price range will work for you. U: As long as it's a guesthouse with 4 stars and free parking I will be happy. Which one do you recommend? S: There are 16 all over town, that meet your criteria. If you could tell me which part of town you would prefer, that would help narrow the choices down. U: Please give me the ph ...	The context does not provide information on the service quality at Charlie Chan restaurant.	Yes, the service at Charlie Chan does seem to be above average. Past guests have talked about the service being accommodating and friendly.	0.3
U: I am planning a trip to Cambridge. I was hoping to find the attraction called Magdalene College. S: Magdalene College is located in the west part of Cambridge. What information would you like about the college? U: I'd like the postcode and address, please. S: Magdalene College is located on Magdalene Street with a postcode of CB30AG. Is there anything else I can help you with? U: I'd like to find the Ashley Hotel, please. S: The Ashley Hotel is a moderately priced hotel on the north side of town. Would you like assistance booking a room there? U: How good is the view of the surrounding area?	Based on the context provided, neither Magdalene College nor the Ashley Hotel have been described as having any particular views or locations related to windows or refrigerators. Therefore, I cannot directly answer this question based on the given context. However, if you need more information about either location, such as their addresses or contact details, feel free to ask! (Sources: User input and context)	The rooms at the Ashley have a great view. Can I answer any additional questions or would you like to book it?	0.3