

Question-Answer Generation-based Evaluation Framework

Simon Loi Baks

sbaks20@student.aau.dk

Aalborg University

Computer Engineering - AI, Vision, and Sound

June 4, 2025



Use of Generative AI

Generative AI tools were used during the thesis process for idea development, literature exploration, and assistance in locating relevant academic sources. Additionally, they supported language refinement and rephrasing of author-written content. No sections were generated and submitted without critical review and rewriting by the author.

GitHub Copilot was used as a supportive tool during implementation, offering quick suggestions and code completions. All code was critically reviewed, tested, and adapted by the author to ensure correctness and alignment with project goals.

All final content, including analyses, experiments, and conclusions, reflects the author's own work. This use complies with AAU's guidelines on generative AI disclosure and is provided here for full transparency.



AALBORG UNIVERSITY

STUDENT REPORT

Title:

Question-Answer Generation-based
Evaluation Framework

Theme:

Master's Thesis: Computer
Engineering - AI, Vision, and
Sound

Project Period:

02/01/2025 - 06/04/2025

Project Group:

1041

Participants:

Simon Loi Baks

Supervisor:

Jesper Rindom Jensen

Co-supervisor:

Simon Dahl Jepsen

Copies:

1

Page Numbers:

72

Date of Completion:

June 04, 2025

Abstract:

Evaluating the quality of automatically generated summaries remains a central challenge in natural language processing. Standard metrics like ROUGE focus on lexical overlap and often fail to capture deeper qualities such as factual consistency, fluency, or relevance. Recent QA-based approaches, such as UniEval, offer multi-dimensional evaluation but often act as black boxes, providing binary outputs without interpretability or reasoning.

This thesis introduces QAG-Eval, a modular framework that combines question generation, answer reasoning, and scalar scoring to evaluate summaries across four quality dimensions: coherence, consistency, fluency, and relevance. By generating natural language justifications, QAG-Eval provides transparent, interpretable evaluations instead of opaque scalar scores.

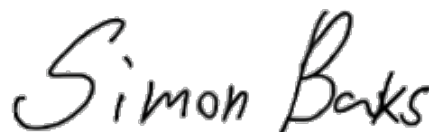
The framework is evaluated against ROUGE and re-trained UniEval models on human-annotated datasets. Results show that QAG-Eval offers strong alignment with human judgments and captures subtle mid-range quality distinctions more effectively. The thesis also analyzes score distributions, justification quality, and performance trade-offs between multi-task and continual learning setups.

By integrating reasoning and scoring in a transparent pipeline, QAG-Eval contributes toward more interpretable, modular, and human-aligned evaluation methods applicable across diverse summarization tasks.

Preface

This thesis was written as part of my Master's degree in Computer Engineering at Aalborg University 2025. I would like to formally express my gratitude for the external collaboration with XCI A/S. During my internship, the company not only welcomed me into their team, but also gave me valuable insight into the role of a Machine Learning Engineer. Their mentorship helped shape my practical skills and professional development.

Thank you for providing the opportunity to explore an exciting and highly relevant research area. Special thanks also go to my supervisors, Jesper Rindom Jensen and Simon Dahl Jepsen, for their continued guidance and for patiently listening to my ramblings during our meetings.

A handwritten signature in black ink that reads "Simon Baks". The signature is fluid and cursive, with the first letter of each word being capitalized and prominent.

Simon Loi Baks
sbaks20@student.aau.dk

Table of Contents

1 Introduction	1
2 Problem Analysis	2
2.1 Motivation and Context	2
2.2 Existing Summarization Techniques	3
2.2.1 Extractive Summarization	3
2.2.2 Abstractive Summarization	3
2.3 Evaluation Metrics for Summarization	4
2.3.1 Key Quality Evaluation Dimensions	4
2.3.2 ROUGE and Traditional Overlap-Based Metrics	5
2.3.3 Other Automatic Evaluation Metrics	6
2.4 Towards QA-Based Evaluation	6
2.4.1 Overview of the UniEval Framework	6
2.4.2 Limitations of the UniEval Framework	7
2.5 Problem Formulation	8
2.6 Research Questions and Evaluation Objectives	9
2.6.1 RQ1: Framework Interpretability	10
2.6.2 RQ2: Scoring Accuracy & Granularity	10
3 Question-Answering Based Evaluation Framework	11
3.1 Framework Motivation	11
3.1.1 Design Objectives and Contributions	11
3.2 Evaluation Dimensions and Metrics	12
3.2.1 Evaluation Criteria	12
3.2.2 Scoring Scale and Label Representation	12
3.2.3 Metric-Based Evaluation	13
3.2.4 Correlation Granularities	13
3.3 Core Components	15
3.3.1 Question-Answer Generation Module	16
3.3.2 Scoring Evaluator	21
3.4 Dataset Creation	23
3.4.1 Intermediate Training Dataset	23
3.4.2 Dimension-Specific QA Dataset	26
3.4.3 Scoring Evaluator Dataset	29
3.5 Training Procedure	30
3.5.1 QAG Module Training	30
3.5.2 Scoring Evaluator Training	32
4 Experimentation	35
4.1 Evaluation Overview	35
4.1.1 Datasets	35
4.1.2 Baselines and Comparative models	37
4.1.3 Evaluation Metrics	38
4.2 Framework Interpretability (RQ1)	39
4.2.1 Motivation	39
4.2.2 Evaluation Objectives	40
4.2.3 Evaluation Procedure	40

4.3	Scoring Accuracy and Granularity (RQ2)	40
4.3.1	Motivation	40
4.3.2	Evaluation Objectives	41
4.3.3	Evaluation Procedure	42
4.4	Experiment Results	42
4.4.1	RQ1: Interpretability Results	42
4.4.2	RQ2: Scoring Accuracy & Granularity Results	45
5	Discussion	51
5.1	Framework Interpretability	51
5.1.1	Evaluation Outcomes	51
5.2	Scoring Accuracy and Granularity	52
5.2.1	Correlation Findings	52
5.2.2	Score Distribution Insights	53
5.3	Comparative analysis	54
5.3.1	UniEval Architectural Limitations	54
5.3.2	Interpretability Gaps	54
5.4	Strengths and Limitations	54
5.5	Future Works	56
6	Conclusion	58
	Bibliography	59
	Appendix A	64
A.1.	Metric Definitions	64
A.1.1.	Spearman Rank Correlation (ρ)	64
A.1.2.	Kendall Rank Correlation Coefficient (τ)	64
	Appendix B	65
B.1.	Score Distribution per Dimension	65
B.1.1.	Training Score Distribution	65
B.1.2.	Validation Score Distribution	66
B.2.	Qualitative Reasoning Trace Showcase	66
B.2.1.	Showcase Example 1: Low-Quality Summary	66
B.2.2.	Showcase Example 2: Mid-Quality Summary	68
B.2.3.	Showcase Example 3: High-Quality Summary	69
	Appendix C	71
C.1.	Tables of Correlation Levels RQ2	71
C.1.1.	Sample-Level Correlation	71
C.1.2.	Summary-Level Correlation	71
C.1.3.	System-Level Correlation	72

List of Figures

Figure 1	Example of extractive summarization using sentence selection	3
Figure 2	Example of abstractive summarization	4
Figure 3	Overview of the framework of UniEval	7
Figure 4	Overview of the QAG-Eval Framework	16
Figure 5	Overview of the FLAN-T5 framework	17
Figure 6	Instruction-based prompting for question and answer generation during intermediate training	19
Figure 7	Dimension-specific prompting for question and answer generation during fine-tuning	21
Figure 8	Scoring Evaluator – Model Input and Output Format	22
Figure 9	Intermediate dataset: Sample distribution by dataset	25
Figure 10	Intermediate dataset: Sample distribution by domain	25
Figure 11	Dimension-specific dataset: Sample distribution by dataset	28
Figure 12	Dimension-specific dataset: Sample distribution by domain	28
Figure 13	Evaluator dataset: Sample distribution	30
Figure 14	Score Distribution per Dimension for Test Data	36
Figure 15	Label distribution across UniEval dimensions	37
Figure 16	Example of a fluency QA pair generated from the QAG module	39
Figure 17	Instruction-based prompting for question and answer generation during intermediate training	41
Figure 18	Grouped histograms comparing score distributions for each model .	49
Figure 19	Histogram of raw UniEval softmax confidence scores	50
Figure 20	Score Distribution per Dimension for Training Data	65
Figure 21	Score Distribution per Dimension for Validation Data	66

List of Tables

Table 1	Granularity levels for metric correlation	14
Table 2	Family of T5 models instruction finetuned	17
Table 3	Family of DeBERTa-V3 models	22
Table 4	Datasets used for intermediate multi-task learning.	24
Table 5	Datasets used for dimension-specific QA generation	27
Table 6	Human-annotated datasets for evaluator training	30
Table 7	Training configuration for Stage 1	31
Table 8	Training configuration for Stage 2	32
Table 9	Training configuration for multi-task scoring evaluator	33
Table 10	Training configuration for continual scoring evaluator	34
Table 11	Average answer length by dimension	43
Table 12	Tokenization of short QA-style responses	43
Table 13	Justification content in QAG-Eval answers	44
Table 14	Sample-Level Correlation of Automatic Metrics for RQ2	71
Table 15	Summary-Level Correlation of Automatic Metrics for RQ2	71
Table 16	System-Level Correlation of Automatic Metrics for RQ2	72

Acronyms

AG - Answer Generation	16
BERT - Bidirectional Encoder Representations from Transformers	21
DeBERTa - Decoding-enhanced BERT with disentangled attention	21
LLM - Large Language Model	1, 2
LM - Language Model	17
MSE - Mean Squared Error	23, 32
NLG - Natural Language Generation	6, 7, 9, 11
NLI - Natural Language Inference	7, 18, 23, 24
NLP - Natural Language Processing	1, 2, 4, 43, 51
QA - Question-Answer	1, 6, 7, 8, 9, 11, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 37, 38, 40, 42, 44, 66
QAG - Question-Answer Generation	15, 16, 23, 29, 30, 32
QAG-Eval - Question-Answer Generation-based Evaluation Framework	1, 11, 12, 13, 15, 16, 29, 30, 32, 33, 35, 36, 37, 38, 42, 43, 44
QG - Question Generation	16
ROUGE - Recall-Oriented Understudy for Gisting Evaluation	1, 5, 38
T5 - Text-to-Text Transfer Transformer	16

1 Introduction

Text summarization refers to the task of producing short, coherent summaries that capture the essential information of a longer document. As one of the foundational challenges in Natural Language Processing (NLP), summarization has evolved significantly, from early rule-based and statistical systems to modern neural models and Large Language Models (LLMs) capable of generating fluent, human-like summaries. While these advances have led to impressive generative performance, they also bring new challenges in reliably evaluating the quality of automatically generated summaries.

Traditionally, evaluation relied on surface-level lexical overlap metrics such as Recall-Oriented Understudy for Gisting Evaluation (ROUGE), which compare model-generated summaries against human-written references. Although widely adopted, these metrics are limited in their ability to assess semantics, factual consistency, or content relevance, especially in abstractive summarization, where multiple valid summaries may differ lexically. More recent learned and Question-Answer (QA)-based evaluation methods attempt to capture deeper quality signals but often suffer from a lack of interpretability, oversimplified scoring schemes, or opaque decision-making.

This thesis explores a new direction for summary evaluation that aims to improve interpretability, scoring granularity, and modularity. The proposed framework, Question-Answer Generation-based Evaluation Framework (QAG-Eval), leverages instruction-tuned models to generate dimension-specific question-answer pairs that reason explicitly about summary quality across fluency, coherence, consistency, and relevance. These reasoning traces are then mapped to human-aligned quality scores, allowing both transparent and fine-grained evaluation.

The remainder of this thesis is structured as follows: Chapter 2 provides a detailed problem analysis, covering existing summarization techniques, evaluation metrics, and recent QA-based approaches. It concludes with a problem formulation and a set of guiding research questions. Chapter 3 introduces the proposed QAG-Eval framework, outlining its evaluation dimensions, core components, dataset construction process, and training procedures. Chapter 4 describes the experimental setup and presents results addressing the defined research questions. Chapter 5 analyzes the empirical findings in greater depth, comparing the proposed framework against existing metrics and models, and concludes with directions for future work. Chapter 6 offers final conclusions and reflections on the contributions of the thesis.

2 Problem Analysis

This chapter analyzes the problem space surrounding automatic summarization and its evaluation. It begins by introducing the motivation and historical development of summarization techniques, followed by a discussion of the current state of evaluation methods used to assess summary quality. Both traditional metrics and more recent learned approaches are considered, along with the quality dimensions they aim to capture. Special attention is given to the emerging use of question-answering as a paradigm for evaluation, including its potential and current limitations. Finally, the chapter concludes by formulating the research problem and objectives that guide the remainder of this thesis.

2.1 Motivation and Context

Text summarization refers to the creation of short, accurate, and fluent summaries of longer documents. As one of the most challenging tasks in NLP, it has seen a steady evolution—from early rule-based systems in the 1950s [1], to unsupervised feature-based methods in the 1990s and 2000s [2–4], and more recently, to neural models trained on large-scale datasets [5–8]. The field has advanced further with the rise of such as BERT [9] and T5 [10] has shown significant performance leap has been amplified by LLMs like GPT-4 [11]; [12], which can produce high-quality summaries that rival those written by humans [13].

Despite these advancements, evaluating the quality of generated summaries remains a complex issue. Traditional automatic metrics like ROUGE [14] and BLEU [15], which rely on n-gram overlap with reference summaries, often fail to capture the semantic adequacy, factual correctness, and overall quality of summaries, especially when multiple valid summaries exist for a single source document. This limitation is particularly pronounced in domains where summaries are expected to capture nuanced information, such as scientific articles, legal documents, and conversational data.

Recent research has consistently highlighted the limitations of current evaluation metrics across a wide range of summarization tasks. Metrics like ROUGE and BLEU, though widely used, often fail to align with human judgment, particularly in settings where semantic adequacy, factual grounding, and contextual relevance are crucial [16, 17]. These issues are amplified in domains such as scientific, legal, or conversational summarization, where the input structure is more complex and the expectations for summary content vary significantly. For example, dialogue summarization presents unique challenges: it requires models to interpret informal, often disfluent language, resolve speaker references, preserve discourse flow, and surface salient points from loosely structured exchanges. These difficulties reflect broader evaluation gaps that persist across summarization domains. As outlined by Kirstein et al. (2024) [18], robust evaluation must account for latent document structure, factual consistency, pragmatic inference, and topic salience, features that are typically invisible to surface-level metrics. Yet, even modern neural or LLM-based evaluation methods often lack transparency or rely on binary decisions that obscure nuanced judgment.

These challenges underscore the pressing need for evaluation frameworks that go beyond token-level matching or opaque scoring systems. Instead, future evaluation methods must offer fine-grained, interpretable, and dimension-specific insights into summary quality. Addressing this need is the central motivation behind the framework proposed in this thesis.

2.2 Existing Summarization Techniques

Modern summarization methods are generally categorized into two main types: extractive and abstractive summarization. The following sections briefly outline the distinctions between them and their respective advantages and limitations.

2.2.1 Extractive Summarization

Extractive summarization, which dates back to early work by Rau et al. in 1989 [19], was among the first approaches to automatic summarization. It involves selecting and concatenating sentences or phrases directly from the source text to form a summary. This method relies on identifying the most informative or salient units, typically using statistical heuristics, graph-based algorithms, or supervised models trained to rank sentence importance.

Extractive methods tend to be more robust in terms of factual accuracy and grammaticality, as they avoid generating new content. However, they often suffer from redundancy, incoherence, and poor discourse structure, particularly when summarizing long or loosely organized source documents. An example of extractive summarization can be seen in Figure 1.

Source:

The city council voted on Tuesday to approve a new green initiative aimed at reducing carbon emissions by 30% over the next decade. The plan includes expanding bike lanes, investing in public transportation, and offering tax incentives for electric vehicle owners. Some council members expressed concerns about the budget implications of the initiative. However, the majority agreed that the long-term environmental benefits outweigh the short-term costs. The initiative is set to begin implementation in early 2025.

Summary:

The city council voted on Tuesday to approve a new green initiative aimed at reducing carbon emissions by 30% over the next decade. The plan includes expanding bike lanes, investing in public transportation, and offering tax incentives for electric vehicle owners. The initiative is set to begin implementation in early 2025.

Figure 1: An illustrative example of extractive summarization. The summary is formed by selecting full sentences directly from the source document without rephrasing or compression.

2.2.2 Abstractive Summarization

Abstractive summarization, by contrast, seeks to generate novel sentences that paraphrase, compress, or reorganize content from the source. This approach aims to mimic how humans write summaries by rephrasing information instead of copying it verbatim [20, 21].

Early abstractive systems employed rule-based or statistical generation techniques, later evolving into neural models using recurrent architectures with attention mechanisms (Rush et al., 2015 [22]). Recent progress has been driven by transformer-based models such as BART (Lewis et al., 2020 [23]), T5 (Raffel et al., 2020 [10]), and other large pre-trained language models. These models, when fine-tuned on summarization datasets, can produce fluent and coherent summaries that often outperform extractive systems in readability and informativeness.

However, this increased generative flexibility also introduces the risk of hallucinations, where generated content is fluent but factually incorrect, along with semantic drift or inconsistencies (Maynez et al., 2020 [24]). These issues are particularly pronounced in complex domains such as multi-document or dialogue summarization, making faithful and interpretable evaluation essential in abstractive pipelines. An example of abstractive summarization can be seen in Figure 2.

Source:

The city council voted on Tuesday to approve a new green initiative aimed at reducing carbon emissions by 30% over the next decade. The plan includes expanding bike lanes, investing in public transportation, and offering tax incentives for electric vehicle owners. Some council members expressed concerns about the budget implications of the initiative. However, the majority agreed that the long-term environmental benefits outweigh the short-term costs. The initiative is set to begin implementation in early 2025.

Summary:

The city council approved a green initiative set to launch in 2025, aiming to cut emissions by 30% through eco-friendly transport investments. Despite budget concerns, most members supported the plan for its long-term environmental impact.

Figure 2: Unlike extractive summarization, which copies full sentences from the source, abstractive summarization generates novel phrasing. This example shows how a model can compress and reword content to produce a more concise and human-like summary.

2.3 Evaluation Metrics for Summarization

Evaluating the quality of generated summaries is a critical yet unresolved challenge in NLP. A wide variety of metrics have been proposed, ranging from simple surface-level comparisons to complex learned models. This section outlines the major categories of evaluation metrics, highlights their limitations, and introduces the key quality dimensions used in modern summarization evaluation.

2.3.1 Key Quality Evaluation Dimensions

Before introducing common evaluation metrics, it is important to clarify the underlying quality dimensions they are intended to assess. Following Kryściński et al. (2019) [25], modern summarization evaluation typically considers four core dimensions, each capturing a distinct aspect of summary quality:

- **Coherence:** Assesses the overall structure and logical flow of the summary. A coherent summary reads naturally and presents information in a sensible order.
- **Consistency (or factual consistency):** Evaluates whether the summary remains faithful to the content of the source text, avoiding hallucinations or unsupported claims.
- **Fluency:** Refers to the grammaticality and readability of individual sentences. A fluent summary is well-written and free from language errors.
- **Relevance:** Measures whether the summary captures the most important and salient information from the source, omitting redundant or trivial details.

These dimensions are not always entirely independent, but they provide a structured basis for evaluating summary quality beyond surface-level similarity. Many existing automatic metrics conflate or overlook these distinct criteria, highlighting the need for more granular and interpretable evaluation methods.

2.3.2 ROUGE and Traditional Overlap-Based Metrics

ROUGE [14] is the most widely used automatic evaluation metric in summarization. It quantifies the lexical overlap between a generated summary and one or more human-written reference summaries, typically using precision, recall, and F1 variants over n-grams or subsequences.

The most common variants include:

- **ROUGE-N**: Measures n-gram overlap between candidate and reference summaries.
 - **ROUGE-1** focuses on unigrams (single words).
 - **ROUGE-2** captures bigram overlap, offering slightly more sensitivity to local phrase structure.
- **ROUGE-L**: Measures the longest common subsequence (LCS), accounting for word order while allowing gaps. This variant is more flexible than strict n-gram matching and emphasizes sequence-level alignment.

The ROUGE-N recall is computed as:

$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{reference summaries}\}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \{\text{reference summaries}\}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)}$$

Where n is the n-gram size, and $\text{Count}_{\text{match}}(\text{gram}_n)$ is the maximum number of overlapping n-grams between the candidate and reference summaries.

ROUGE-L uses the longest common subsequence between reference (X) and candidate (Y) texts:

$$\begin{aligned} R_{\text{lcs}} &= \frac{\text{LCS}(X, Y)}{m} \\ P_{\text{lcs}} &= \frac{\text{LCS}(X, Y)}{n} \\ F_{\text{lcs}} &= \frac{(1 + \beta^2) R_{\text{lcs}} P_{\text{lcs}}}{R_{\text{lcs}} + \beta^2 P_{\text{lcs}}} \end{aligned}$$

Where m and n are the lengths of the reference and candidate summaries, respectively. β is typically set to balance recall and precision, often as $\beta = \frac{P_{\text{lcs}}}{R_{\text{lcs}}}$.

Due to its simplicity, language independence, and ease of use, ROUGE has become a *de facto standard* in summarization benchmarks. However, it suffers from several well-known limitations:

- It fails to account for semantic equivalence, penalizing valid paraphrases that do not match at the surface level.
- It is insensitive to factual errors or hallucinations.
- It provides no explanatory insight—only numerical similarity scores.

These limitations are especially problematic for **abstractive summarization**, where correct summaries may differ significantly in phrasing from reference texts. Nonetheless, ROUGE remains widely used due to its efficiency, reproducibility, and historical precedent. For consistency with prior work, ROUGE is included as a baseline in this thesis.

2.3.3 Other Automatic Evaluation Metrics

In response to ROUGE’s limitations, several alternative metrics have been proposed to better capture meaning, factual accuracy, and human judgment alignment.

- Semantic similarity metrics such as BERTScore [26] and MoverScore [27], attempt to measure deeper contextual alignment between summaries and references using pre-trained language models.
- Learned scoring models like BARTScore [28] generate likelihood-based scores directly from transformer models conditioned on source and/or reference inputs.
- QA-based metrics, including FEQA [29] and QuestEval [30], attempt to assess factual consistency by generating and answering questions derived from the summary or source content.

While these approaches address some of ROUGE’s weaknesses, particularly in handling paraphrasing and factual accuracy, they often suffer from limited transparency and interpretability. Most produce a single overall score without distinguishing between specific quality dimensions such as fluency, coherence, or consistency. As a result, it is often unclear which aspects of summary quality are being evaluated or what types of errors contribute to the score.

2.4 Towards QA-Based Evaluation

QA has emerged as a promising paradigm for evaluating summaries, especially in contexts where factual consistency or content inclusion must be assessed. Unlike traditional metrics that rely on string overlap, QA-based methods evaluate summaries by asking whether specific questions can be answered using only the summary, or whether the summary answers key questions about the source.

A number of QA-style metrics have been introduced, including SummaQA [31], FEQA [29], and QuestEval [30]. While these metrics help identify factual omissions or hallucinations, they are often limited in scope, apply only to factual consistency, and offer no insight into other quality dimensions such as fluency or coherence.

More recently, UniEval [32] has extended QA-based evaluation to multiple dimensions. The following subsections provide a detailed overview of the UniEval framework and its limitations.

2.4.1 Overview of the UniEval Framework

UniEval [32] is a unified, automatic evaluation framework designed to assess Natural Language Generation (NLG) outputs, such as summaries or dialogue responses across multiple quality dimensions including **coherence**, **consistency**, **fluency**, and **relevance**. It reframes multi-dimensional evaluation as a **Boolean QA** task, enabling a single language model to handle all evaluation dimensions via simple, natural language prompts. An overview of the framework can be seen in Figure 3.

At its core, UniEval uses a fine-tuned T5 [33] model to answer dimension-specific yes/no questions about the source text. For example, to assess coherence and relevance, the model is asked:

“Is this a coherent summary to the document?”

“Is this summary relevant to the reference?”

The model outputs either “Yes” or “No,” and the confidence in the “Yes” answer is used to compute a continuous score:

$$s_i = \frac{P(\text{Yes}|x, y, c, q_i)}{P(\text{Yes}|x, y, c, q_i) + P(\text{No}|x, y, c, q_i)}$$

Here, s_i is the score for dimension i , and $P(\cdot)$ denotes the model’s probability of generating a specific output token. By leveraging the softmax-normalized probability over “Yes” and “No,” UniEval turns binary classification into a graded scoring system. This formulation allows the model to reflect uncertainty in its judgment, producing real-valued scores aligned with human quality.

To prepare the model for this task, UniEval undergoes a two-stage training process:

1. **Intermediate multi-task pretraining**, using generic QA, Natural Language Inference (NLI), linguistics-related tasks, and Opening Sentence Prediction as a self-supervised task to develop general-purpose reasoning skills.
2. **Unsupervised fine-tuning on evaluation tasks**, using pseudo data generated via rule-based transformations (e.g., deleting words to break fluency, swapping sentences to break coherence).

The final model is capable of **zero-shot or few-shot evaluation**, using only question prompts to assess quality along different axes, without requiring separate evaluators for each task or human-annotated labels.

However, this design choice also introduces several trade-offs, particularly in **interpretability**, **reasoning traceability**, and **content-level fidelity**. Because the model generates a final decision without exposing the underlying content relationships it considered, it behaves as a black-box evaluator. Additionally, while the binary QA format is efficient, it provides no explicit evidence of what information was preserved, omitted, or hallucinated in the generated summary.

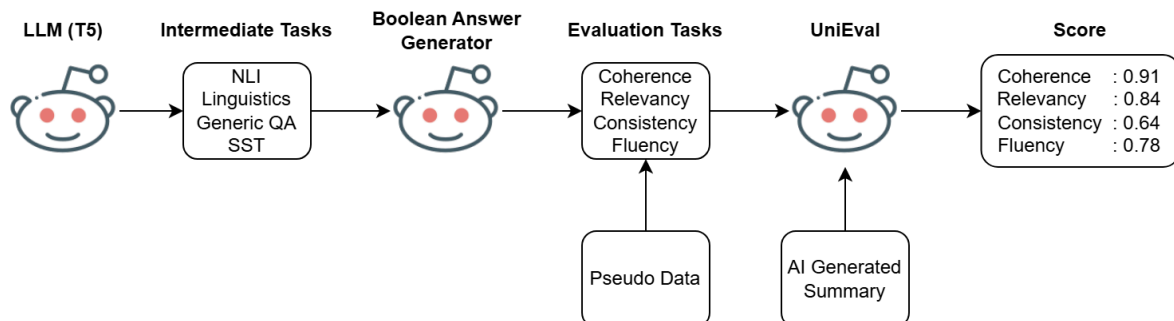


Figure 3: An overview of the overall framework of UniEval.

2.4.2 Limitations of the UniEval Framework

Evaluating the quality of NLG outputs, especially in complex domains such as dialogue summarization, presents ongoing challenges [18]. Traditional metrics based on surface similarity (e.g., ROUGE, BERTScore) fall short in capturing deeper semantic qualities like factual correctness, coherence, and informativeness. To address these shortcomings, the UniEval framework proposes a unified, multi-dimensional evaluation approach that reformulates evaluation as a Boolean QA task. This framework enables a single model to assess multiple quality dimensions, such as coherence, consistency, fluency, and relevance, by prompting the model with dimension-specific yes/no questions.

While UniEval represents a significant step toward more comprehensive and flexible evaluation, several architectural and methodological limitations merit closer analysis. These limitations,

detailed below, identify potential blind spots in UniEval’s design and suggest avenues for further improvement in the development of evaluation metrics.

2.4.2.1 Lack of Interpretability and Explainability

UniEval relies on a pre-trained language model to produce binary judgments (“Yes”/“No”) in response to evaluation questions. However, the model offers no explicit reasoning or evidence behind its predictions. This black-box design limits transparency and makes it difficult for users to verify or understand why a given summary is rated as incoherent or irrelevant. As evaluation tools are increasingly deployed in high-stakes or safety-critical contexts (e.g., medical or legal summarization), the lack of interpretability poses a significant limitation. The authors of UniEval themselves acknowledge this limitation, explicitly stating that although neural evaluators can correlate well with human judgments, it remains unclear how exactly the model arrives at its evaluation scores.

2.4.2.2 Binary Scoring Formulation May Oversimplify Complex Judgments

UniEval quantifies the quality of generated text by calculating the model’s confidence in a “Yes” answer, which is interpreted as a continuous score. While this enables straightforward scoring across dimensions, it inherently reduces nuanced quality assessments to a probabilistic binary decision. Many aspects of summary quality, such as partial factual overlap or borderline coherence, may not be well captured by a binary framing. This could lead to the underrepresentation of ambiguous or intermediate cases in evaluation results.

2.4.2.3 Synthetic Pseudo Data: Quality and Realism Concerns

A key feature of UniEval is its use of rule-based pseudo data for unsupervised training on evaluation tasks. Negative examples are constructed through transformations such as sentence swapping (to break coherence) or entity substitution (to break consistency). Although this eliminates the need for costly human annotations, it introduces a dependency on handcrafted corruption rules that may not accurately reflect real-world model outputs. As a result, the model may overfit to specific error patterns introduced during pseudo data construction and underperform on naturally occurring errors found in real summaries. The authors of UniEval explicitly acknowledge concerns about noise in their synthetic pseudo data. For example, when constructing negative samples for fluency, they remove text spans without considering their importance, which may not affect sentence fluency. Nevertheless, these altered sentences are still consistently labeled as negative examples. This indicates potential gaps in the realism and quality of the pseudo data used to train the evaluator.

2.4.2.4 Implicit Modeling of Informational Content and Faithfulness

UniEval’s QA-based design evaluates global qualities of text (e.g., “Is this summary consistent with the document?”) without explicitly modeling what factual or informational content is preserved. There is no built-in mechanism to assess whether key pieces of information from the source text are present, omitted, or hallucinated in the generated output. This limits the system’s sensitivity to fine-grained content-level discrepancies and may obscure critical omissions or hallucinations that negatively affect summary utility.

2.5 Problem Formulation

Automatic evaluation of text summaries remains a challenging task, particularly for complex domains such as dialogue, where traditional metrics (e.g., ROUGE, BLEU) often fail to capture the nuanced quality judgments made by humans. Recent frameworks such as UniEval have

introduced instruction-based, multi-dimensional evaluation methods to approximate human judgment across dimensions like coherence, consistency, fluency, and relevance.

However, UniEval faces notable limitations, including:

1. Lack of explicit reasoning traces, making it difficult to interpret or trust the model’s scoring decisions.
2. Oversimplification of complex quality dimensions due to a binary scoring formulation.
3. Reliance on synthetic or pseudo-labeled training data, raising concerns about realism and alignment with human preferences.
4. A boolean QA generator that yields yes/no judgments, but does not verify whether key information is preserved, omitted, or hallucinated.

This thesis investigates whether a question generation and answering-based evaluation strategy can address these limitations by:

- Introducing interpretable, dimension-specific QA probes that explicitly model quality dimensions.
- Leveraging instruction-tuning on diverse, multi-domain datasets to improve generalization.
- Training a separate scoring module to simulate human judgment on a discrete 1–5 scale, rather than a binary formulation.
- Incorporating human-annotated and grounded datasets to ensure human alignment and scoring fidelity.

Ultimately, this thesis aims to develop a transparent, modular, and generalizable evaluation framework that can be applied across diverse NLG tasks. As such, the central research question guiding this thesis is:

How can a summarization evaluation framework be designed to be transparent, modular, and generalizable across multiple NLG domains, while producing human-aligned and interpretable quality scores?

2.6 Research Questions and Evaluation Objectives

As discussed in the previous sections, the UniEval framework presents a notable advancement in automatic multi-dimensional evaluation across different NLG domains. However, as the authors also stated, its design still suffers from key limitations. These include its reliance on binary QA outputs without explicit reasoning traces, limited capacity for capturing borderline or ambiguous quality judgments due to its probabilistic scoring formulation, its dependency on synthetic pseudo data that may not reflect realistic error patterns, and finally, the lack of explicit modeling of preserved or hallucinated content.

To address these limitations, this thesis proposes a novel, QA-based evaluation framework designed to produce interpretable, human-aligned scores across multiple quality dimensions using modular components. The framework incorporates reasoning-aware QA generation, and a regression-based scoring model trained on real human-annotated data. A thorough and detailed explanation of the framework will be given in Chapter 3.

To systematically investigate the effectiveness and practical value of this approach, the following research questions guide the experimental evaluation. Each of these research questions will be explored through targeted experiments, described in more detail in Chapter 4. The experiments include cross-domain evaluations, score correlation analysis, and qualitative comparisons of

model outputs to test both predictive performance and the transparency of the evaluation process.

2.6.1 RQ1: Framework Interpretability

The following research question aims to investigate whether the QA-based reasoning approach, where dimension-specific questions are generated and explicitly answered, provides meaningful explanatory traces that both improve trust and transparency, but also provide meaningful context for improved evaluation scoring. While direct human validation is outside the scope of this thesis, the proposed framework’s interpretability is assessed using a combination of structural and behavioral criteria. Specifically, interpretability is defined as the transparency and traceability of the evaluation process, evidenced by the presence of explicit dimension-specific questions, natural language answers, and their alignment with summary content. In addition to qualitative examples, the thesis includes quantitative proxy indicators such as average explanation length, presence of content-specific justifications, and full trace coverage for each score.

RQ1: To what extent does the proposed evaluation framework provide interpretable and transparent quality judgments compared to black-box models such as UniEval?

2.6.2 RQ2: Scoring Accuracy & Granularity

The following research question aims to investigate how closely the scores predicted by the proposed framework align with human-labeled quality judgments, using correlation metrics such as Spearman and Kendall. In addition, an analysis will be conducted on the score distributions between UniEval and the proposed framework, with particular attention to intermediate human ratings. This is intended to assess whether the regression-based design of the proposed approach captures subtle differences in summary quality more faithfully than UniEval’s soft binary output. To ensure a fair comparison, the UniEval model is retrained on the same training data as the proposed framework.

Beyond comparing QAG-Eval with UniEval, the experimental setup also includes the traditional summarization metric, ROUGE, following the benchmarking protocol established in prior works. These reference-based metrics have long served as standard tools for summary evaluation and offer a surface-level approximation of quality based on lexical or semantic overlap with human-written references. Including them provides a more comprehensive evaluation landscape, enabling a clearer understanding of the trade-offs between interpretability, semantic sensitivity, and alignment with human judgment.

RQ2: How well do the scores predicted by the proposed framework correlate with human annotations, and how effectively does it distinguish between intermediate levels of summary quality compared to UniEval and standard metrics?

3 Question-Answering Based Evaluation Framework

In this chapter, QAG-Eval will be introduced, a novel framework designed to evaluate summaries across multiple quality dimensions using a *question-answering-based approach*. Traditional evaluation metrics such as ROUGE often fall short in capturing deeper quality aspects of a summary, such as factual consistency, relevance to the source, linguistic fluency, and structural coherence. This is especially true when applied to abstractive or context-rich content.

QAG-Eval reframes summary evaluation as a structured reasoning task. Each quality dimension is transformed into a set of targeted questions that probe specific properties of the summary and optionally the context. This includes whether it faithfully represents the source content, includes key information pieces, and maintains readability and logical flow. The system then generates answers to these questions based on either the summary alone or the source-summary pair, and uses the results to infer quality scores through a separate scoring model.

This chapter presents a comprehensive breakdown of the QAG-Eval pipeline. Section 3.1 outlines the underlying motivation and the challenges that inspired the framework. Section 3.2 defines the evaluation dimensions used and explains how each is operationalized within the QA paradigm. Section 3.3 details the core components of the architecture, including instruction-tuned modules for question and answer generation, as well as the scoring evaluator. Section 3.4 discusses the design and construction of datasets used across the training stages. Finally, Section 3.5 describes the training procedures employed, including general QA training, scoring model learning, and optional domain-specific fine-tuning.

3.1 Framework Motivation

As detailed in the previous chapter, existing summarization evaluation frameworks suffer from critical limitations. Traditional metrics such as ROUGE fail to capture deeper semantic and contextual nuances. Meanwhile, recent multi-dimensional frameworks, such as UniEval, introduce promising multi-dimensional assessment, but remain constrained by interpretability issues, oversimplified binary judgments, synthetic training data, and inadequate modeling of fine-grained informational content.

To explicitly address these limitations, this thesis proposes the QAG-Eval, which fundamentally reframes summary evaluation as an interpretable, structured QA task. Specifically, QAG-Eval introduces explicit reasoning traces via dimension-specific QA probes, moves beyond binary scoring toward fine-grained 1–5 human-aligned scores, and leverages diverse, human-annotated datasets to ensure robustness and generalizability. By embedding explicit interpretability and human-like reasoning directly into the evaluation pipeline, QAG-Eval bridges existing methodological gaps and provides a transparent, modular, and flexible framework, suitable for evaluating summaries across a wide range of NLG tasks.

3.1.1 Design Objectives and Contributions

The QAG-Eval framework is designed to address several limitations in existing evaluation systems, such as UniEval and traditional metrics like ROUGE. The key design objectives and corresponding contributions of this thesis are:

- **Interpretability:** The framework generates explicit reasoning traces in the form of natural language question–answer pairs per evaluation dimension. This allows human reviewers to verify and understand why a score was assigned.
- **Dimension-Specific Evaluation:** Each core quality dimension (fluency, coherence, consistency, relevance) is modeled independently, using both tailored prompts and separate scores.

- **Modularity:** The pipeline separates reasoning (QA generation) and scoring (regression), enabling flexible adaptation and independent improvement of each module.
- **Human-Aligned Scoring:** A regression-based scoring model trained on 1–5 human labels (rather than binary decisions) captures nuanced human preferences and reduces mid-score compression.
- **Multi-Domain Robustness:** The use of multi-task intermediate learning and domain-varied fine-tuning datasets improves generalization across dialogue, news, and open-domain summarization tasks.

Together, these contributions provide a flexible, interpretable, and human-aligned alternative to black-box or reference-only evaluation metrics.

3.2 Evaluation Dimensions and Metrics

Evaluating the quality of abstractive summaries requires attention to multiple quality dimensions that collectively reflect human preferences and expectations. Unlike extractive summaries, which often preserve factual consistency and relevance by construction, abstractive summaries must be judged for how well they retain key information, maintain coherence, and fluently convey ideas using paraphrased expressions derived from the source content.

3.2.1 Evaluation Criteria

The evaluation criteria used in this framework are based on the four key quality dimensions introduced by Kryściński et al. (2019) [25], which have become a standard in summarization evaluation and research.

- **Consistency:** Assesses factual alignment between the summary and the source. A summary is consistent if all of its claims are directly supported or entailed by the source content.
- **Relevance:** Measures the importance of the information selected from the source. A relevant summary highlights key content while omitting trivial or off-topic details.
- **Fluency:** Evaluates sentence-level quality in terms of grammar, readability, and stylistic naturalness. It focuses on the summary’s adherence to expected linguistic norms.
- **Coherence:** Captures how well the summary flows as a whole, considering sentence order, topical progression, and discourse-level structure.

3.2.2 Scoring Scale and Label Representation

While human annotators typically provide ratings on each of these dimensions using a discrete 1-5 Likert scale, originally introduced by Rensis Likert [34], ranging from:

- 1 = Very poor
- 2 = Poor
- 3 = Acceptable
- 4 = Good
- 5 = Excellent

QAG-Eval, however, models scoring as a continuous regression task. Rather than predicting discrete categories, the model is trained to output a floating-point score in the 1.0–5.0 range, using the mean of multiple annotator ratings as the target label. This allows the evaluator to capture subtle gradations in summary quality, such as distinguishing between outputs that would otherwise be assigned the same categorical label but differ in perceived quality (e.g., 3.7 vs. 4.2).

This approach aligns well with datasets such as SummEval [35] and DialSummEval [36], in which each summary is independently scored by multiple human annotators across all dimensions. By leveraging aggregated human judgments, QAG-Eval produces calibrated, human-aligned predictions and avoids the brittleness of hard class boundaries. The continuous scoring scheme thus supports more nuanced and realistic evaluation, both during training and in downstream applications.

3.2.3 Metric-Based Evaluation

In addition to qualitative definitions of the evaluation dimensions, QAG-Eval adopts a set of quantitative metrics to assess and monitor model performance during training and evaluation. These metrics are chosen to reflect both prediction accuracy and alignment with human judgment, particularly in regression-based scoring settings. This section summarizes the key metrics used throughout the framework. All formal definitions and formulations are included in Appendix A.1, while this section provides a usage-oriented overview.

3.2.3.1 Correlation-Based Metrics

These metrics assess the degree to which the model’s predicted scores align with human ratings, focusing on both relative and absolute agreement.

- **Spearman Rank Correlation (ρ)**

Measures monotonic agreement by comparing the ranked order of scores. It assumes no particular distribution and is robust to nonlinear but monotonic trends. Spearman is used as the primary early-stopping criterion, since relative quality ranking is often more meaningful than absolute score accuracy [37].

- **Kendall’s Tau (τ)**

Provides an alternative rank correlation metric, placing greater emphasis on pairwise concordance. It is typically more conservative than Spearman and useful for validating ranking robustness [38].

These metrics are reported per dimension (consistency, relevance, fluency, coherence) and aggregated where needed. This allows for fine-grained analysis of strengths and weaknesses in the evaluator’s ability to model human judgment.

3.2.4 Correlation Granularities

Automatic–human correlation can be measured at different granularities, each capturing distinct facets of evaluation metric performance. Prior work in summarization meta-evaluation, notably by Bhandari et al. (2020) [39], formally introduced two primary correlation strategies: summary-level and system-level, which evaluate how well automatic metrics align with human judgments at varying degrees of aggregation. The UniEval framework (Zhong et al., 2022) later introduced an additional sample-level correlation, focusing on per-summary agreement. The QAG-Eval framework adopts all three levels to provide a comprehensive analysis of metric quality. A summary of all three correlation levels can be seen in Table 1.

The correlation levels presented below follow a shared computational setup, inspired by the formulation in Bhandari et al. (2020). A consistent notation is introduced first to standardize the definitions and facilitate comparison across levels:

Let each document in the dataset be denoted as d_i where $i \in \{1, \dots, n\}$, and let each document have J system outputs. Denote the j^{th} summary of the i^{th} document as s_{ij} , where $j \in \{1, \dots, J\}$. Let m represent a scoring function (e.g., an automatic metric or human annotation), and K a correlation measure (e.g., Spearman or Kendall Tau).

Level	Unit of Analysis	Computation	Insights
Sample	Single summary	Correlation of predicted \leftrightarrow mean human scores per summary	Captures per-summary alignment and outliers
Summary	All systems' summaries for one source document	Mean correlation of human \leftrightarrow metric rankings across systems per document	Highlights whether metrics preserve local ranking
System	All summaries from one system	Correlation between mean predicted \leftrightarrow mean human scores per system	Stable estimate of leaderboard agreement

Table 1: Three correlation levels used in QAG-Eval. Each level captures a different view of metric-human agreement: local accuracy, document-level ranking, and system-wide alignment.

3.2.4.1 Sample-Level Correlation

Definition: Measures the correlation between predicted metric scores and the mean human score for each individual summary across the dataset. This approach, commonly referred to as sample-level correlation in UniEval, was not part of the original Bhandari et al. (2020) formulation.

Purpose: Assesses whether the metric produces scores that align with human ratings at the level of individual summaries.

Example: Given three distinct summaries:

- Summary A: human = 4.5, predicted = 4.3
- Summary B: human = 3.0, predicted = 3.1
- Summary C: human = 2.0, predicted = 2.5

The correlation is computed across these points: $\text{correlation}([4.5, 3.0, 2.0], [4.3, 3.1, 2.5])$.

Insight: This level reveals instance-specific alignment, offering a fine-grained view of how well a metric tracks human perception on a per-summary basis.

3.2.4.2 Summary-Level Correlation

Definition: For each source document, the correlation is calculated between human-assigned and metric-predicted rankings of all system-generated summaries. The final score is the average correlation across all documents.

Formula: Let d_i represent a document with system outputs s_{i1}, \dots, s_{iJ} . The summary-level correlation is calculated as:

$$K_{m_1 m_2}^{\text{sum}} = \frac{1}{n} \sum_{i=0}^n (K([m_1(s_{i1}), \dots, m_1(s_{iJ})], [m_2(s_{i1}), \dots, m_2(s_{iJ})]))$$

Purpose: Evaluates the metric's ability to preserve the relative ranking of system outputs for a given source document.

Example: Suppose three systems generate outputs for the same document. If human scores rank the outputs as $A > B > C$ and metric scores rank them as $A > C > B$, a partial rank correlation is computed for this document. This process is repeated across all documents, and the correlations are averaged.

Insight: This level highlights document-level variation and whether a metric is consistent with human judgment in ranking systems for the same input.

3.2.4.3 System-Level Correlation

Definition: Computes the average predicted score and average human score for each system across all documents, then correlates these vectors of system-level means.

Formula: Let J denote the number of systems evaluated on n documents. The system-level correlation is calculated as:

$$K_{m_1 m_2}^{\text{sys}} = K \left(\left[\frac{1}{n} \sum_{i=0}^n (m_1(s_{i1}) \dots m_1(s_{iJ})) \right], \left[\frac{1}{n} \sum_{i=0}^n (m_2(s_{i1}) \dots m_2(s_{iJ})) \right] \right)$$

Purpose: Determines whether the metric produces system-wide rankings that align with human assessments, averaged over the entire test set.

Example: Consider three summarization systems:

- System A: human mean = 4.1, predicted mean = 4.0
- System B: human mean = 3.5, predicted mean = 3.6
- System C: human mean = 2.9, predicted mean = 3.1

The correlation is computed across the three system-level pairs: correlation([4.1, 3.5, 2.9], [4.0, 3.6, 3.1]).

Insight: This level provides the most stable indication of overall system performance, smoothing out per-sample variance and offering a leaderboard-style evaluation perspective.

3.3 Core Components

The QAG-Eval framework is built around a two-stage pipeline that transforms summary evaluation into a structured, interpretable reasoning task. As illustrated in Figure 4, the system first generates dimension-specific QA pairs based on the summary and optionally its source content. These QA pairs are then fed into a dedicated scoring evaluator, which produces a fine-grained quality score on a continuous 1.0-5.0 scale.

The architecture supports three core aspects of QAG-Eval. First, it promotes transparency by generating QA pairs that show how the model evaluates each dimension. These reasoning steps can be reviewed by humans to understand the basis for each score. Second, it offers modularity, allowing the reasoning and scoring components to be developed and improved upon independently. Third, it enables generalizability, since the same components can be reused across different domains and types of summaries. At each stage, the framework is designed to be interpretable, making it easier to diagnose model behavior and understand why a particular score was assigned.

This section describes the two core components that make up the QAG-Eval pipeline: the Question-Answer Generation (QAG) module and the Scoring Evaluator. The QAG module is responsible for producing interpretable, dimension-specific reasoning traces by generating natural-language question-answer pairs that probe summary quality. This component is

implemented using a unified FLAN-T5 model trained in two stages: an intermediate multi-task learning stage and a dimension-specific fine-tuning stage. The Scoring Evaluator then maps these QA traces to continuous quality scores using a DeBERTa-based regression model. Together, these components form a modular and interpretable evaluation system capable of approximating human-like summary assessment across multiple quality dimensions. Figure 4 provides an overview of the framework.

All architectural figures in this section and those that follow use a consistent color-coding scheme to enhance readability: blue boxes represent input data, green denotes model outputs, purple indicates instruction prompts, red highlights QAG-Eval modules, and magenta is used for scoring components. This convention is followed consistently throughout the thesis.

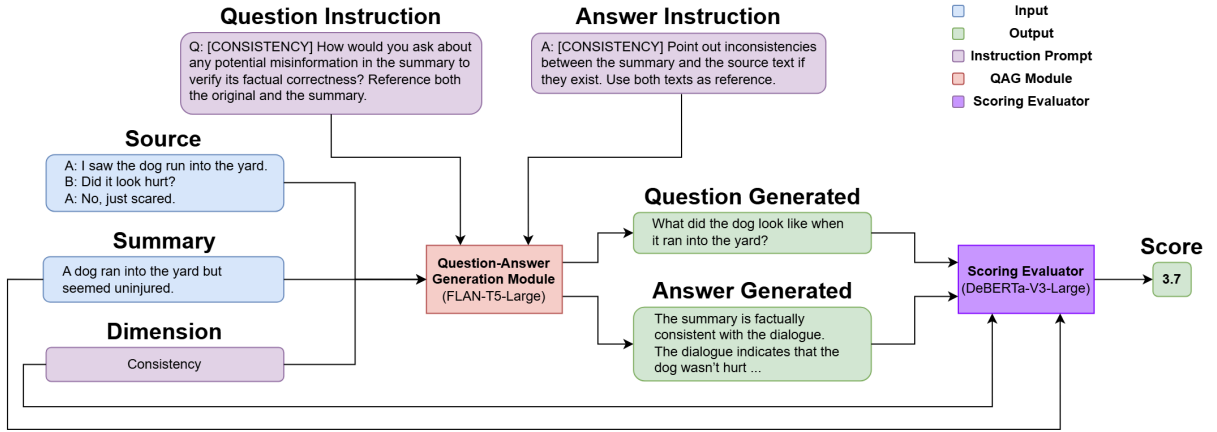


Figure 4: Overview of the QAG-Eval framework. Given a source, a generated summary, and a target evaluation dimension, the FLAN-T5-based QAG module produces a dimension-specific question and answer pair. This reasoning trace is then passed to a DeBERTa-based scoring evaluator, which predicts a continuous quality score between 1.0 and 5.0. The architecture supports interpretability and modular fine-tuning across evaluation dimensions.

3.3.1 Question-Answer Generation Module

The core of the QAG-Eval framework is a unified FLAN-T5-Large [40] model trained to handle both Question Generation (QG) and Answer Generation (AG) tasks. Rather than relying on two separate components, the framework employs a single model to sequentially produce both elements of each QA pair, which allows for a shared contextual understanding and consistent reasoning across steps.

The module serves as the foundation of QAG-Eval’s reasoning-based evaluation strategy: it transforms summaries and source inputs into interpretable, dimension-specific QA traces that reflect the model’s understanding of summary quality. These traces serve not only as internal evidence for the scoring evaluator, but also as optional explanations that can be reviewed or visualized by humans.

The QAG module is trained in two stages: an initial **intermediate multi-task learning stage**, designed to teach the model general-purpose reasoning skills and generate QA pairs, and a **dimension-specific fine-tuning stage**, that specializes the model for generating dimension-specific questions and answers.

3.3.1.1 Model Architecture and Prompt Design

The Question-Answer Generation (QAG) module is implemented using a unified FLAN-T5-Large model [40], a transformer-based encoder-decoder architecture pre-trained for instruction

following. In the paper they instruction finetuned various different models, most interestingly the T5 family. See Table 2 for the family of T5 models that were instruction finetuned. FLAN-T5 builds on the original Text-to-Text Transfer Transformer (T5) model [41] by combining the original span-corruption pre-training objective with large-scale instruction tuning. This tuning stage incorporates a wide variety of tasks, including question answering, summarization, entailment, and classification, expressed in natural language prompts. As a result, FLAN-T5 demonstrates strong generalization to unseen tasks and domains, making it particularly suitable for zero-shot and few-shot scenarios.

Model	Params	Architecture
Flan-T5-Small	80 M	Encoder-Decoder
Flan-T5-Base	250 M	Encoder-Decoder
Flan-T5-Large	780 M	Encoder-Decoder
Flan-T5-XL	3 B	Encoder-Decoder
Flan-T5-XLL	11 B	Encoder-Decoder

Table 2: This table lists the five encoder-decoder variants of the FLAN-T5 family, all of which share the same Transformer architecture and have been instruction fine-tuned on a diverse mixture of tasks to enable strong zero-shot and few-shot generalization.

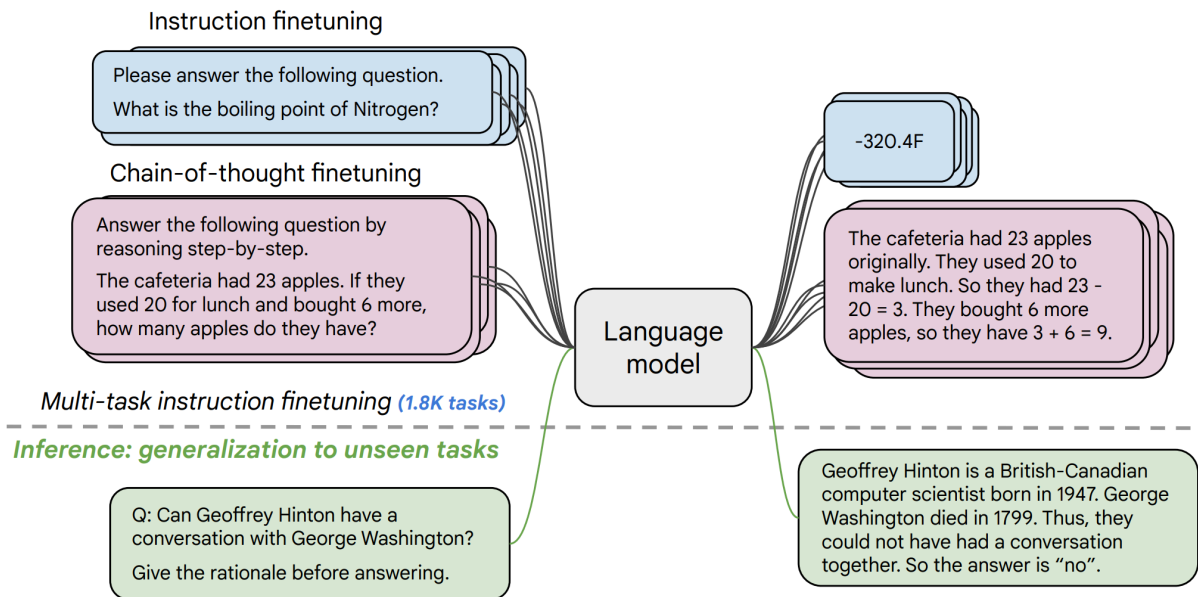


Figure 5: An overview of the FLAN-T5 framework used to finetune various LMs 1.800 tasks, each phrased as instructions [40].

For the proposed evaluation framework the *Large* variant is used and contains approximately 780 million parameters and inherits the full encoder-decoder structure of T5, enabling both conditional generation and multi-step reasoning. The design choice of using this model was primarily guided and inspired by UniEval. In the UniEval framework they used the original T5-Large variant which showcased strong performance across multiple domains. As such, it was only natural to use the descending model family FLAN-T5. The Large variant was chosen as a tradeoff between performance and hardware constraints.

FLAN-T5’s instruction-tuned nature makes it ideal for modular, prompt-driven pipelines. In QAG-Eval, a single instance of the model is used to perform both question generation (QG) and

answer generation (AG), rather than splitting these steps across separate models. This unified approach ensures shared parameterization and consistent reasoning across QA pairs, which is critical when modeling structured reasoning traces for summary evaluation.

Inputs to the model are formatted using **natural-language instructions**, consistent with the FLAN-style prompting paradigm. Instructions explicitly describe the generation task and the required context. During intermediate training, each instruction is prefixed with a “Q:” or “A:” tag to help the model distinguish between question and answer generation tasks. Example prompts from this stage include:

- *Q: Rewrite the follow-up question into a standalone question using the context.*
- *A: Answer the standalone question using the context.*

During dimension-specific fine-tuning, the prompt format is augmented with explicit **evaluation dimension tags**, such as [CONSISTENCY], [RELEVANCE], [FLUENCY], and [COHERENCE]. These tags condition the model to focus its generation behavior on specific quality dimensions, enabling a flexible and interpretable evaluation process. For example:

- *Q: [CONSISTENCY] Write a question that could reveal hallucinated or unsupported claims in the summary.*
- *A: [CONSISTENCY] Compare the information in the summary with the source to identify any mismatches.*

This unified prompt framework ensures that the model learns to associate different generation behaviors with distinct tasks and quality dimensions while maintaining flexibility in input phrasing. Prompts are paraphrased across training examples to prevent instruction overfitting and to promote generalization across domains.

Each question and answer is generated independently by the same model, using the same encoder-decoder structure. This modular generation behavior allows the QAG module to serve as a reusable and adaptable reasoning engine within the broader evaluation pipeline.

3.3.1.2 Intermediate Multi-Task Learning

After initializing the QAG module with a general-purpose, instruction-following model (FLAN-T5-Large), the first training stage focuses on teaching foundational QA and reasoning skills across diverse NLP tasks. This **intermediate multi-task learning** stage is not yet aligned with summarization evaluation but instead builds the broad reasoning abilities that underpin the QAG module’s later specialization.

A wide range of datasets is used during this stage to promote generalization and expose the model to different question and answer formats. Each dataset is selected to support a particular reasoning Multicompetency, for example, the Multi-Genre NLI [42] and the Stanford NLI [43] dataset, are primarily used for teaching the model entailment and contradiction detection. The Question Rewriting in Conversational Context (QReCC) [44] dataset is used for teaching the model paraphrased standalone question generation from conversation contexts. A table of all datasets used in this stage can be found in Section 3.4.1, Table 4.

Each dataset is converted into instruction-style prompts using the FLAN prompting format. Natural language instructions explicitly define the task behavior. These prompts are prepended with a “Q:” or “A:” tag to signal whether the model is expected to generate a question or an answer. For instance:

QReCC:

- “Q: Rewrite the follow-up question into a standalone question using the context.”
- “A: Answer the standalone question using the context.”

MNLI:

- “A: Determine the logical relationship between the following two statements.”

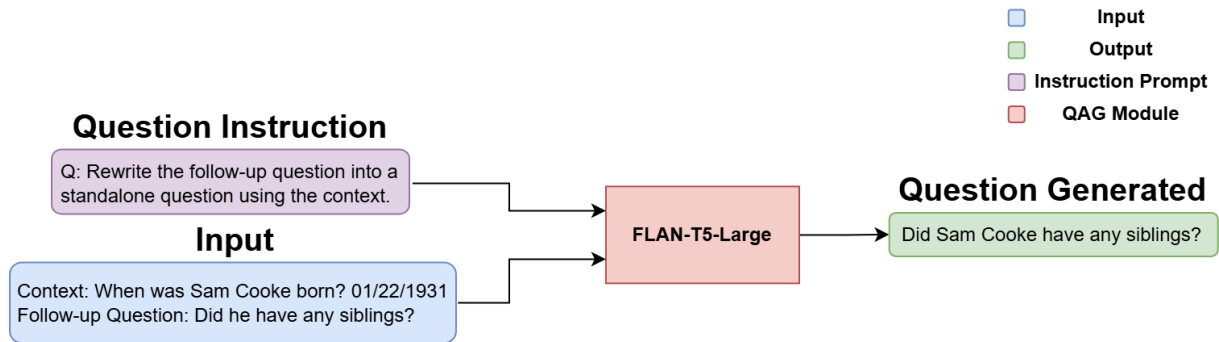
Quoref:

- “Q: Generate a question that requires resolving coreference in the text.”

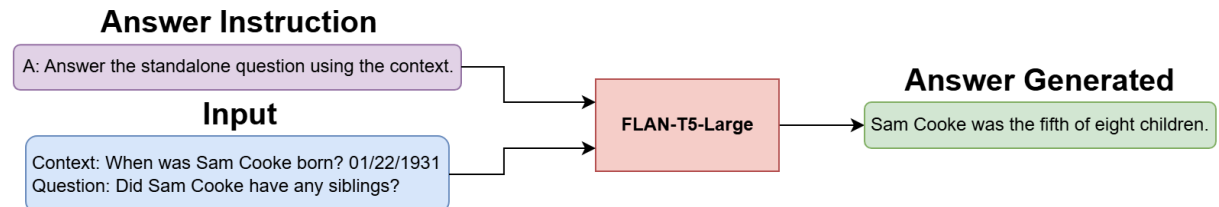
CNN/DailyMail:

- “Q: Generate a question that reflects the information captured in the summary.”

Each instruction is treated as an independent training instance. Some datasets support both QG and AG instructions, while others contribute only one. This decoupled design allows the model to generalize across tasks and promotes flexible adaptation in the next training stage.



(a) Instruction-based prompt for question generation, with an example illustrating the process using a sample from the QReCC dataset.



(b) Instruction-based prompt for answer generation, with an example illustrating the process using a sample from the QReCC dataset.

Figure 6: Examples of instruction-based training prompts from the intermediate multi-task learning stage. Each prompt is constructed from the same QReCC sample and used independently for question or answer generation.

3.3.1.3 Fine-Tuning for Dimension-Specific QA Generation

Building on the model’s multi-task pretraining, the second training stage fine-tunes the QAG module for the specific task of evaluating summary quality. This stage introduces **dimension-specific supervision**, teaching the model to generate reasoning traces that correspond to the four evaluation criteria described in Section 3.2: consistency, relevance, fluency, and coherence.

Each training instance consists of a summary and, where applicable, the source input. The model is first prompted to generate a question that targets a specific quality dimension. Then, in a separate instruction, the model generates an answer to that question based on the same input. Unlike the intermediate stage, these prompts include explicit dimension tags (e.g., [CONSISTENCY]) to guide the model’s reasoning. Example prompt pairs include:

- Consistency (summary + source):

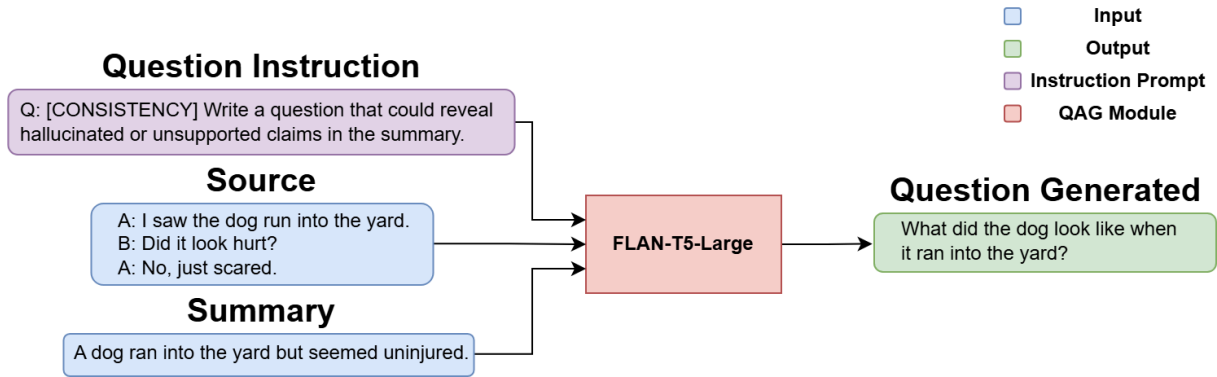
- ▶ “Q: [**CONSISTENCY**] Write a question that could reveal hallucinated or unsupported claims in the summary.”
- ▶ “A: [**CONSISTENCY**] Check whether the information in the summary matches the original source facts.”
- Relevance (summary + source):
 - ▶ “Q: [**RELEVANCE**] How would you ask to ensure the summary addresses the key aspects of the source material?”
 - ▶ “A: [**RELEVANCE**] Confirm if the summary addresses each major topic or event from the source.”
- Fluency (summary only):
 - ▶ “Q: [**FLUENCY**] Generate a question that tests how easy the summary is to read aloud or silently.”
 - ▶ “A: [**FLUENCY**] Point out grammatical errors or awkward phrasing if present.”
- Coherence (summary only):
 - ▶ “Q: [**COHERENCE**] Generate a question that checks for clarity in how the summary moves from one idea to the next.”
 - ▶ “A: [**COHERENCE**] Determine if there are any abrupt transitions or confusing jumps in the summary.”

To improve the model’s robustness and avoid overfitting to rigid templates, each instruction was paraphrased into several semantically equivalent forms. For example, the consistency question could alternatively be phrased as:

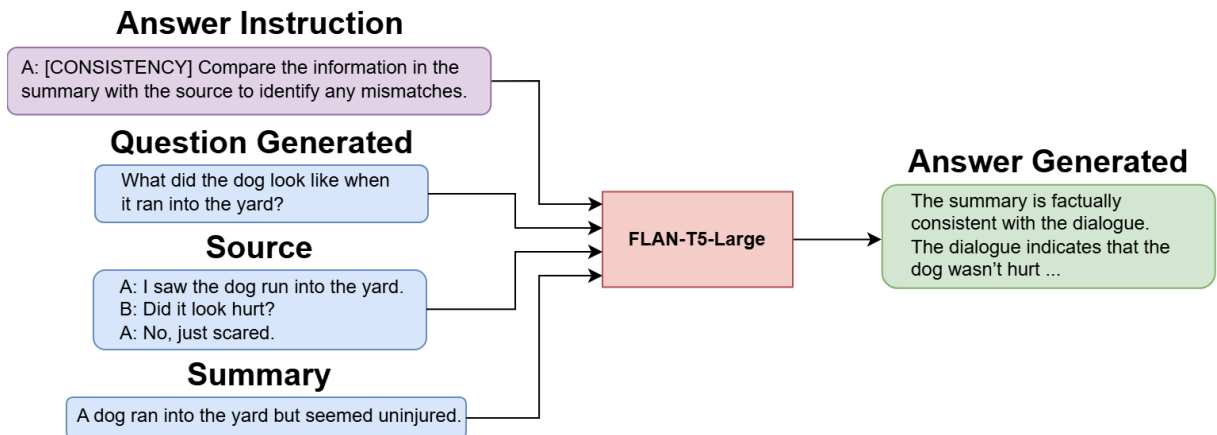
- “Formulate a question to verify the summary’s factual accuracy.”
- “How would you test whether the summary aligns with the source?”

The dataset used for this fine-tuning was synthetically constructed using GPT-4 prompting. A curated set of summarization datasets from diverse domains was selected to ensure coverage of different writing styles and structures. For example, Extreme Summarization (XSum) [45] dataset is based on news articles and consists of short abstractive summaries. Samsung (SAMSum) [46] dataset consists of very short abstractive summaries centered around dialogue conversations. A table of all datasets used can be found in Section 3.4.2, Table 5.

This stage completes the transformation of the QAG module from a general-purpose reasoning engine into a dimension-aware assistant capable of producing interpretable justifications for summary quality evaluation.



(a) Instruction prompt for dimension-specific question generation. The instruction targets the consistency dimension.



(b) Instruction prompt for dimension-specific answer generation. The model responds to a consistency-focused question using summary and source context.

Figure 7: Examples of dimension-specific QA generation prompts used during fine-tuning. Each prompt is constructed from the same sample and targets the consistency dimension. Question and answer generation are handled independently by the FLAN-T5 model, guided by explicit instruction tags.

3.3.2 Scoring Evaluator

The scoring evaluator constitutes the second major component of the QAG-Eval pipeline. Its primary function is to translate the structured reasoning traces generated by the QAG module’s dimension-specific QA pairs, into scalar quality scores on a continuous 1.0–5.0 scale. Unlike conventional evaluators that operate directly on raw summaries, QAG-Eval simulate the evaluative reasoning process a human annotator might apply, promoting both interpretability and modularity. This section details the model architecture, input formatting, and evaluation principles that govern this module’s design.

3.3.2.1 Model Architecture and Input Format

The scoring evaluator is built on top of a pre-trained DeBERTa-Large model, a transformer-based encoder architecture known for its strong performance on language understanding tasks. Compared to earlier models like Bidirectional Encoder Representations from Transformers (BERT) [47] and [48], Decoding-enhanced BERT with disentangled attention (DeBERTa) [49] introduces improvements in how it represents word meaning and sentence structure, which help it better capture subtle differences in language. These strengths make it a strong foundation for evaluating dimension-specific QA explanations that reflect human judgment.

Model	Params	Architecture
DeBERTa-V3-XSmall	22 M	Encoder
DeBERTa-V3-Small	44 M	Encoder
DeBERTa-V3-Base	86 M	Encoder
DeBERTa-V3-Large	304 M	Encoder

Table 3: This table lists selected models from the DeBERTa V3 family, which are based on an encoder-only transformer architecture. Each variant differs in parameter size, enabling trade-offs between computational efficiency and performance.

The core model architecture employed in this module is based on the DeBERTa-v3 encoder, chosen for its strong semantic representation capabilities and competitive performance on sentence-level regression tasks. A regression head is appended to the final layer to produce a scalar score. While encoder-decoder variants such as FLAN-T5 were also explored, the encoder-only setup was ultimately preferred for its efficiency and alignment with the input format.

Each input sample is flattened into a natural-language prompt comprising:

- the **dimension tag** (e.g., [CONSISTENCY]),
- the **generated question**,
- the **generated answer**, and
- the original **summary**.

This format enables the model to condition its prediction on a complete reasoning trace, rather than raw summary content alone. A representative example:

- “**Q:** [CONSISTENCY] What key claims in the summary might not be supported by the source?”
- “**A:** [CONSISTENCY] The claim about the minister’s resignation is not corroborated in the source.”
- “**S:** [SUMMARY] The article reports that the finance minister resigned amidst public pressure...”

This structured representation encourages the model to ground its prediction in explicit, human-readable logic.

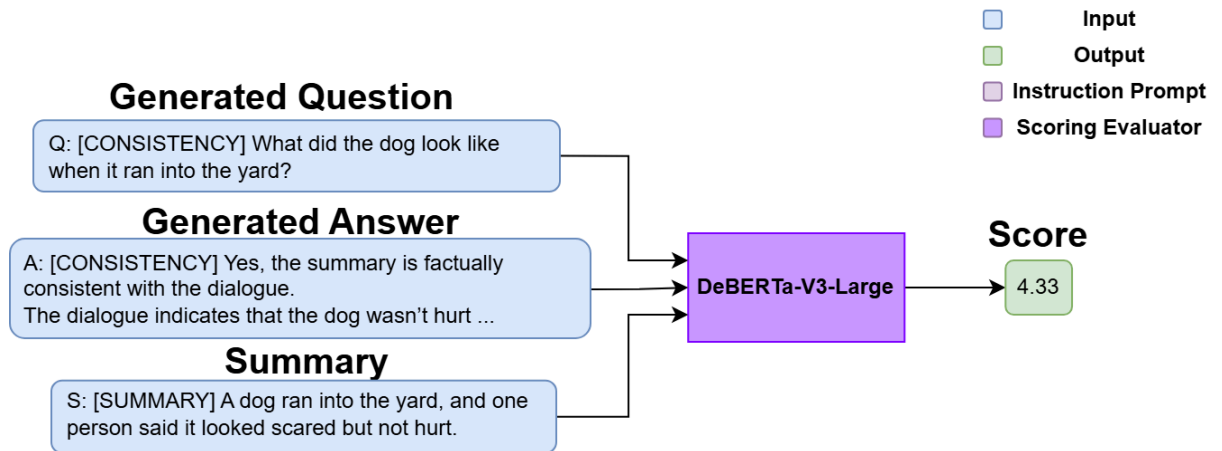


Figure 8: Overview of the scoring evaluator. The model processes structured QA traces and predicts a continuous quality score per dimension. Input is flattened into a concatenated prompt that includes the question, answer, and original summary.

3.3.2.2 Scoring Objective and Prediction Behavior

The evaluator is trained as a regression model that predicts continuous quality scores between 1.0 and 5.0 for each evaluation dimension, using structured QA traces as input. Each trace is paired with a human-annotated score, representing the average judgment across multiple annotators. The model learns to map linguistic reasoning artifacts, such as questions probing summary quality and their corresponding answers, to scalar ratings that reflect consensus human opinion. To encourage the model to rank summaries in the same order humans prefer, Spearman rank correlation is used as the early stopping criterion, ensuring that the model prioritizes the correct ordering of summaries even if score magnitudes differ. The regression loss itself is computed using Mean Squared Error (MSE).

At inference time, the model takes as input a QA trace and summary for a given dimension and outputs a real-valued prediction. Because each prediction is grounded in a natural-language reasoning process, model outputs are inherently interpretable: low scores can be traced back to problematic answers or misleading questions, offering insight into evaluation failures.

Only a single regression head is used across all dimensions, with the active dimension explicitly embedded in the prompt. This design maintains architectural simplicity while enabling flexible control via input formatting.

The model is trained on two human-annotated datasets, SummEval [35] and DialSummEval [36], covering both news and dialogue domains. These datasets offer real-world summaries scored across fluency, coherence, consistency, and relevance, enabling supervised learning with reliable evaluation targets.

3.4 Dataset Creation

This chapter details the construction and curation of datasets used throughout the QAG-Eval pipeline, spanning the intermediate multi-task training stage, the fine-tuning of the QA generation module, and the supervision of the scoring evaluator. Each section provides both the rationale behind dataset selection and an explanation of how each dataset contributes to the targeted evaluation dimensions. Limitations and domain coverage are also discussed to clarify the trade-offs made during dataset creation.

3.4.1 Intermediate Training Dataset

The first stage of training the QAG module involves an intermediate multi-task learning stage. This stage introduces the model to a broad range of reasoning types and input-output formats to instill the necessary competencies for summary evaluation. Unlike dimension-specific fine-tuning, the objective here is not to align the model with specific evaluation criteria but to equip it with general-purpose skills like entailment, contextual inference, paraphrasing, and both extractive and abstractive QA capabilities.

The datasets selected for this stage span a variety of domains and tasks, each contributing distinct reasoning challenges. As shown in Table 4, they include span-based QA datasets like SQuAD v2 and NewsQA, multi-hop inference tasks like HotpotQA, paraphrasing datasets like Quora and QReCC, and dialogue-centric datasets like CoQA and SAMSum. Most of the dataset choices were inspired by the UniEval paper, which used similar datasets for question-answering, NLI, opening sentence prediction, and linguistically related tasks.

Dataset	Domain	Purpose
CNN/DailyMail [50]	Long summarization	Teaches answer location across large spans of narrative text.
CoQA [51]	Conversational QA	Builds dialogue-aware reasoning and contextual chaining across turns.
HotpotQA [52]	Multi-hop QA	Enhances multi-step logical reasoning across multiple documents.
MNLI / SNLI [42, 43]	Natural Language Inference	Develops entailment and contradiction detection, critical for faithfulness probing.
NewsQA [53]	Span-based QA	Trains contextual answer grounding in a news domain.
QReCC [44]	Question rewriting	Teaches paraphrased standalone question generation from conversation context.
Quora [54]	Paraphrasing	Enhances lexical and syntactic diversity in question generation.
Quoref [55]	Coreference QA	Strengthens resolution of entity references and pronouns.
SAMSum [46]	Dialogue + Summary	Bridges dialogue summarization with question generation training.
SQuAD v2 [56]	QA (span + no-answer)	Instills span-based extraction and answerability uncertainty modeling.
WikiQA [57]	Sentence-level QA	Introduces binary QA decision-making and sentence-level entailment.
XSum [45]	Abstractive summarization	Encourages highly compressed summary abstraction and generation diversity.

Table 4: Datasets used in the intermediate multi-task learning stage of QAG-Eval. Each dataset contributes a specific reasoning or QA capability foundational for later fine-tuning.

To support general-purpose reasoning and robust instruction-following behavior, the intermediate training dataset was constructed from 13 diverse QA-style corpora spanning domains such as news, dialogue, NLI, coreference resolution, paraphrasing, and multi-hop question answering. Each dataset was selected to target specific cognitive or linguistic capabilities relevant to QA generation, such as entailment detection, discourse coherence, answer span extraction, and question rewriting. While the raw dataset pool contained substantially more entries, only a filtered subset of 29,000 instruction-formatted samples was retained. This filtering step excluded samples with input sequences that exceeded the context window of the FLAN-T5 model, ensuring compatibility with the model’s architecture. Dataset contributions were kept relatively balanced to avoid overrepresentation by any single domain or task. This stage served to prepare the model for downstream dimension-specific fine-tuning by exposing it to a wide spectrum of input-output formats and reasoning styles. Figure 9 and Figure 10 visualize the sample distribution by dataset and domain.

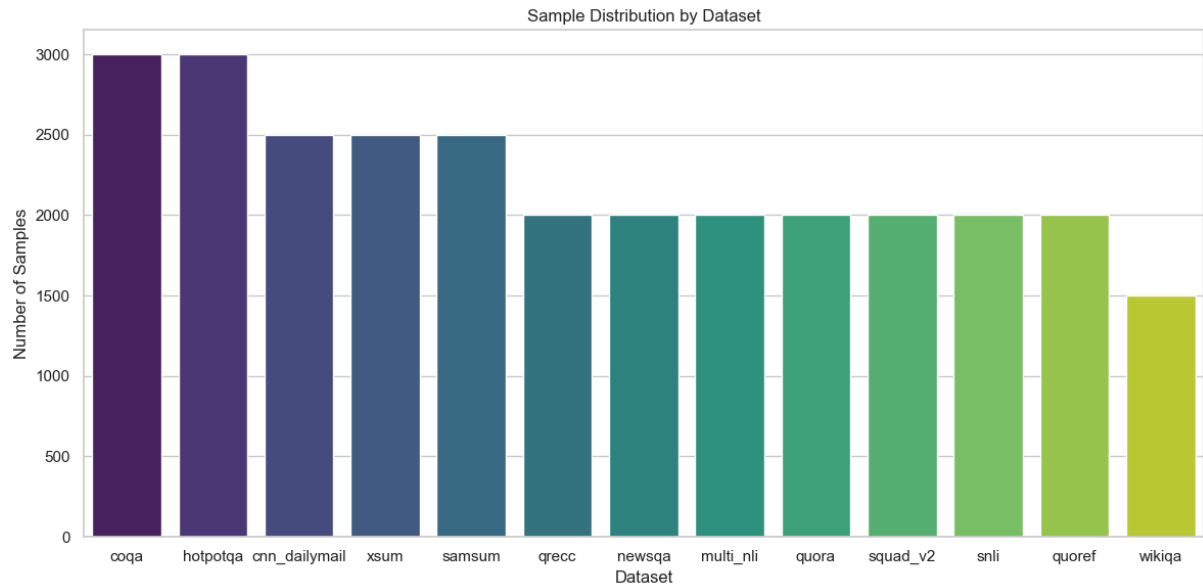


Figure 9: Bar chart showing the number of training samples included from each dataset used in the intermediate multi-task training stage.

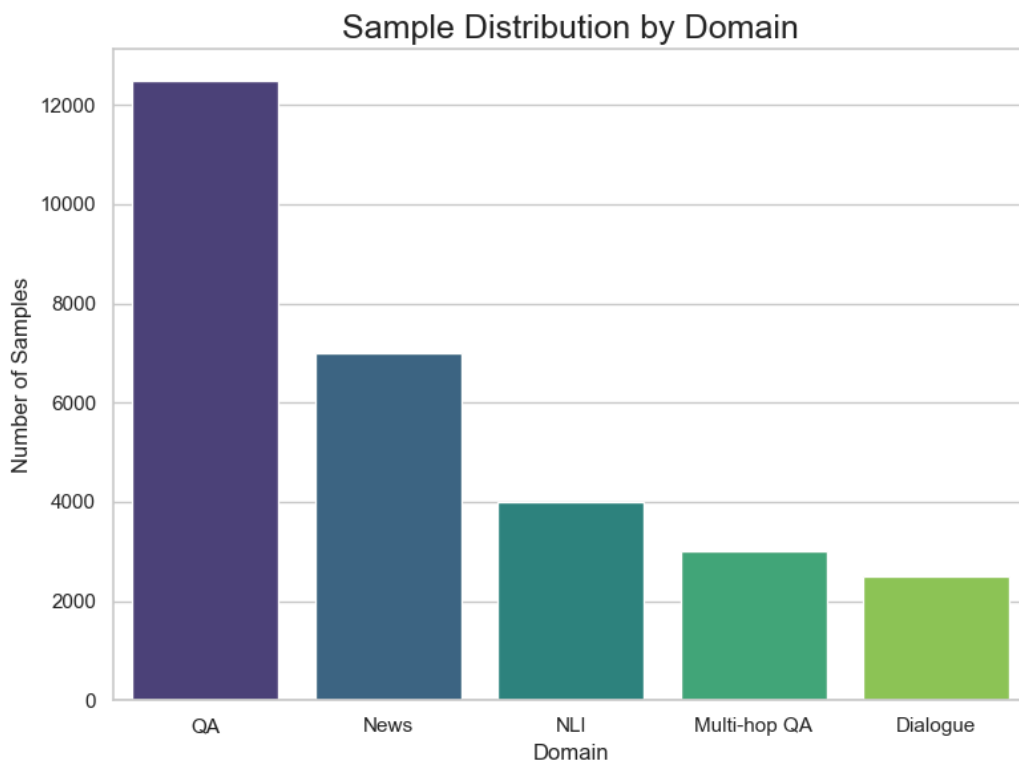


Figure 10: Bar chart showing the distribution of training samples by domain in the intermediate multi-task learning stage.

This stage also serves a secondary purpose: familiarizing the model with a wide range of input representations. All datasets are converted into instruction-style prompts in the FLAN format, marked with either a “Q:” (question) or “A:” (answer) prefix depending on the objective. For example, a span-based QA task might be formatted as:

- “Q: Generate a question that asks about a key detail from the passage.”
- “A: Extract the correct answer from the passage.”

By varying task types and instruction formulations, the model becomes robust to diverse input-output schemas, a feature critical for generalization across domains and summary types.

The intermediate stage is not supervised using summary quality scores. Instead, it lays the groundwork for the model’s ability to generate meaningful QA traces across contexts, which will be explicitly directed toward quality dimensions in the next fine-tuning stage.

3.4.2 Dimension-Specific QA Dataset

Following intermediate multi-task training, the unified FLAN-T5 model was fine-tuned on a curated dataset of dimension-specific QA pairs. This dataset forms the backbone of the QAG module’s reasoning behavior, bridging general-purpose QA skills with dimension-specific summary evaluation. Each entry in this dataset consists of a generated question and answer tailored to one of the four quality dimensions, using a summary and, when necessary, its corresponding source.

Each dataset contributes a fixed number of samples based on its relevance to specific dimensions. For example:

- News datasets (e.g., XSum, CNN/DailyMail) support relevance and fluency assessments,
- Biomedical/scientific datasets (e.g., PubMed, arXiv) enable hallucination and consistency probing,
- Dialogue datasets (e.g., SAMSum, DialogSum) provide strong signals for coherence and fluency,
- Legal/governmental summaries support consistency and structured relevance.

This sampling strategy ensures representation of:

- Extractive vs. abstractive styles
- Short-form vs. long-form summaries
- Formal vs. informal domains

The complete breakdown of dataset contributions is provided in Table 5.

Dataset	Domain	Purpose
XSum [45]	News (abstractive)	Short, focused summaries useful for fluency and coherence assessment
CNN/DailyMail [50]	News (longform)	Longer summaries ideal for evaluating relevance and factual consistency
SAMSum [46]	Dialogue	Tests coherence and fluency in conversational summarization
DialogSum [58]	Dialogue	Complementary dialogue domain with more grounded summaries
Reddit TIFU [59]	Informal/social	Paraphrased, casual writing style ideal for fluency robustness
WikiHow [57]	Instructional	Tests logical flow and step-based coherence
GovReport [60]	Government	Dense and factual — useful for faithfulness and detail accuracy
BillSum [61]	Legal	Structured language ideal for checking relevance and factual alignment
PubMed [62]	Biomedical	Fact-heavy domain for hallucination detection and consistency
arXiv [63]	Scientific	Long-form abstracts suited for relevance and factuality

Table 5: Datasets used for constructing the dimension-specific QA instruction dataset. Each summary was paired with a synthetic question and answer targeting a specific evaluation dimension. The datasets span a wide range of domains to encourage generalization across content types and writing styles.

To ensure robust generalization across evaluation domains, the dataset was constructed from summaries sourced across 10 public summarization corpora spanning news, dialogue, instructional, scientific, legal, and informal domains. While the source texts and summaries originate from real-world datasets, the dimension-specific questions and answers were entirely synthetically generated using GPT-4.1. For each summary–dimension pair, the model was prompted with a carefully designed instruction to produce a targeted question and corresponding answer, followed by manual filtering and paraphrasing to ensure variation and quality. After filtering for input length constraints, approximately 15,000 unique summaries were retained. Each summary was then expanded into four QA pairs, one per evaluation dimension, resulting in a total of ~61,000 training samples. This setup promoted balanced domain exposure and dataset scale while avoiding overfitting to specific content types or phrasing styles. Figure 11 and Figure 12 visualize the sample distribution by dataset and domain.

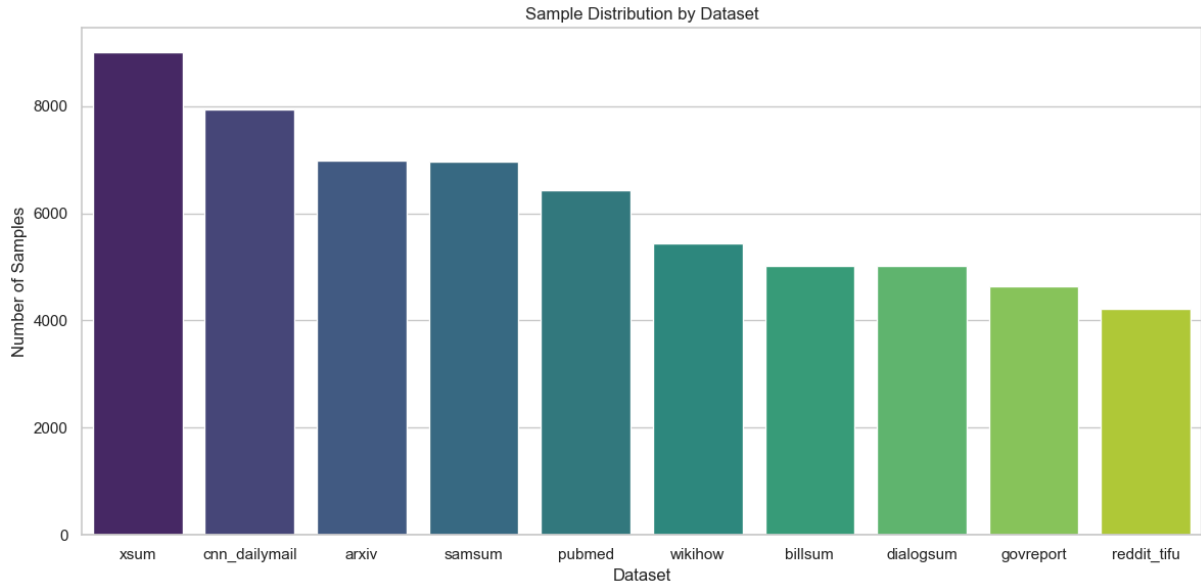


Figure 11: Bar chart showing the number of training samples included from each dataset used in the dimension-specific QA fine-tuning stage.

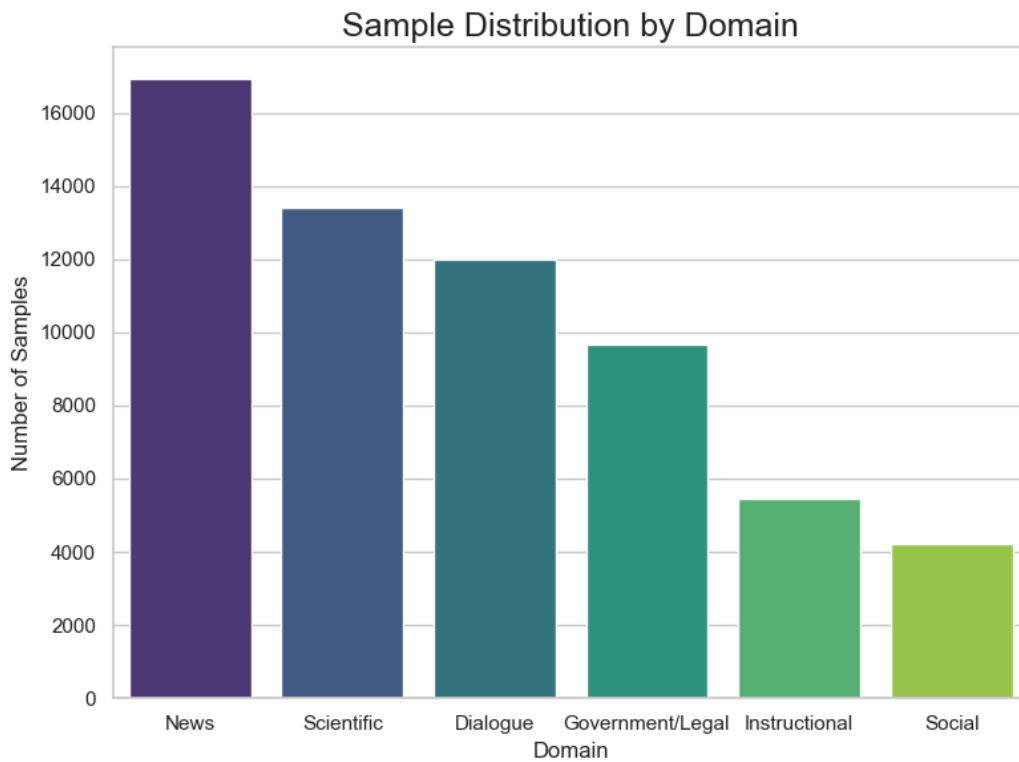


Figure 12: Bar chart showing the distribution of training samples by domain in the dimension-specific QA fine-tuning stage.

Each summary was paired with one QA pair per evaluation dimension. Prompts were generated using GPT-based instruction scaffolds and later paraphrased to ensure instruction variation. These QA pairs serve as fine-tuning signals, guiding the QAG model to approximate human-like reasoning when judging summaries along specific dimensions.

This dataset plays a key role in transitioning the model from general QA ability to structured, interpretable, and dimension-specific quality assessment.

3.4.3 Scoring Evaluator Dataset

The final dataset used in the QAG-Eval pipeline is designed to supervise the scoring evaluator, the component responsible for assigning real-valued quality scores to summaries based on the QAG module’s dimension-specific reasoning traces. Unlike previous phases that focus on generative capabilities, this stage teaches the model to align its scoring behavior with human judgments.

The dataset is derived from two widely-used human-annotated benchmarks:

- SummEval: A general-domain summarization benchmark where summaries are rated across four dimensions: consistency, relevance, fluency, and coherence, by multiple annotators.
- DialSummEval: A dialogue-specific dataset that mirrors SummEval’s annotation framework but evaluates summaries grounded in conversational inputs, supporting adaptation to spoken-language summaries.

Both datasets contain a small number of unique source documents (100 each), paired with 14–16 system-generated summaries per input, resulting in a limited set of annotated samples. Each summary is rated by multiple annotators (3 for SummEval and 8 for DialSummEval), providing averaged dimension-specific scores.

To convert these annotations into training data for the scoring evaluator, the QAG module is applied to each summary to generate one QA pair per evaluation dimension. These dimension-aligned QA traces serve as input to the scoring evaluator, while the mean human rating acts as the regression target.

After filtering for sequence length and formatting compatibility, the number of usable samples was reduced to ~1,300 from SummEval and ~1,500 from DialSummEval. After combining the datasets approximately ~2,800 samples was available, and after further splitting the dataset (80%/10%/10%) into separate training, validation, and test datasets, the total number of samples available for the training was ~2,300. Each sample was then exploded into four dimension-specific entries, yielding approximately ~5,200 and ~6,000 samples, respectively.

To mitigate this size limitation and improve generalization, further augmentation was applied to the dataset by generating multiple paraphrased QA traces per summary–dimension pair. This was achieved by re-prompting the QAG module with varied phrasings to create alternative yet semantically equivalent questions and answers for each dimension. This approach yielded a final dataset of ~140,000 samples, and the distribution can be seen in Figure 13.

This augmentation strategy serves two purposes:

1. It diversifies the linguistic inputs used for each summary, reducing overfitting to specific QA phrasing.
2. It teaches the model score invariance, i.e., that different reasoning formulations describing the same quality issue should receive the same score.

As a result, the scoring evaluator becomes more robust to natural variation in phrasing and better equipped to generalize beyond its limited supervision base. This dataset ultimately closes the training loop for QAG-Eval by anchoring structured QA-based reasoning in real-world human evaluation standards.

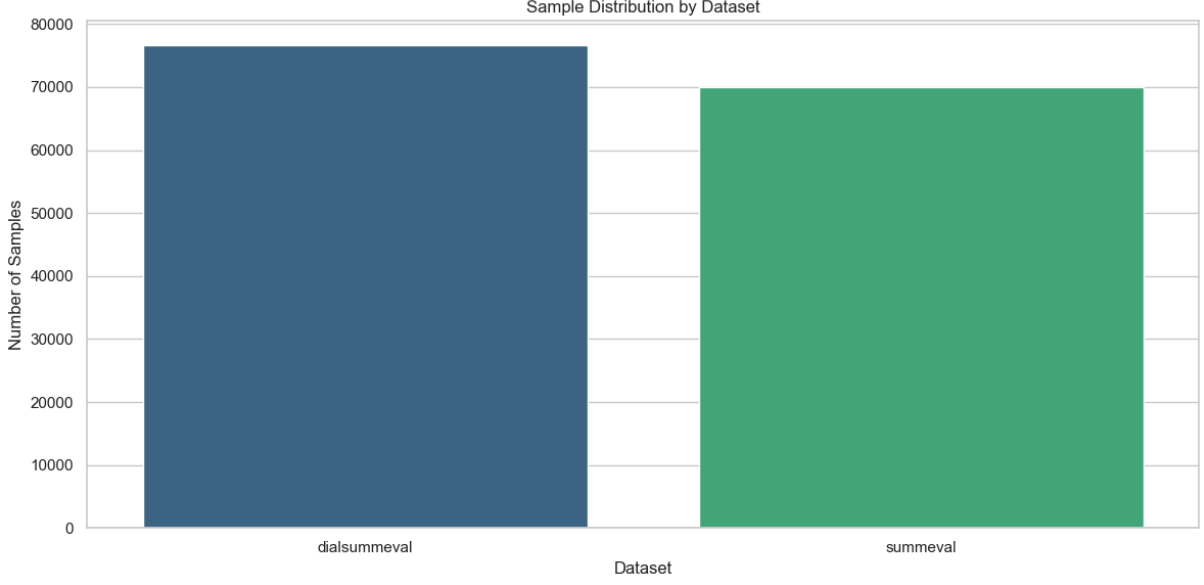


Figure 13: Bar chart showing the number of training samples derived from each dataset (SummEval and DialSummEval) after QAG alignment and augmentation.

Dataset	Domain	Purpose
SummEval [35]	News	Summaries annotated by multiple human raters across fluency, coherence, consistency, and relevance.
DialSummEval [36]	Dialogue	Dialogue summaries rated on the same four dimensions by crowd annotators, supporting domain adaptation.

Table 6: Datasets used to supervise the scoring evaluator. Each dataset contains summaries annotated across the four quality dimensions. These ratings serve as regression targets for training the evaluator.

3.5 Training Procedure

This chapter outlines the training procedure applied to the QAG-Eval pipeline across its three core stages. Each stage builds upon the previous one, gradually specializing the model from general-purpose reasoning to dimension-specific evaluation and ultimately to scoring grounded in human judgment.

3.5.1 QAG Module Training

The QAG module was trained in two consecutive stages using the FLAN-T5-Large architecture. The goal was to first equip the model with a broad foundation in instruction-based question answering and then fine-tune this ability toward generating dimension-specific reasoning traces for summary evaluation.

3.5.1.1 Stage 1: Intermediate Multi-Task Learning

In the first stage, the model was trained on a diverse multi-task dataset spanning QA, entailment, paraphrasing, span extraction, and coherence tasks (see Table 4). The use of FLAN-style instructions allowed the model to practice reasoning over different text types and question formats in a consistent, interpretable way.

The training configuration was selected to balance efficiency, stability, and generalization. FLAN-T5-Large was fine-tuned for 30 epochs using a cosine learning rate schedule with a warmup ratio of 3%. A conservative learning rate of $5e-6$ and weight decay of 0.01 were chosen

to promote smooth convergence. Input sequences were capped at 512 tokens to fit within the model’s context length, and early stopping was employed to avoid overfitting. To accommodate hardware limitations and stabilize training, a batch size of 4 was combined with gradient accumulation over 4 steps, yielding an effective batch size of 16. Mixed-precision training (BF16) helped reduce memory consumption and training time. Gradient clipping with a max norm of 1.0 was used to prevent exploding gradients.

This stage was inspired by prior work such as FLAN and T0, which demonstrated that broad multi-task instruction tuning helps models generalize to unseen formats. Unlike UniEval, it was decided not to employ continual learning with dimension-wise replay. While continual fine-tuning can mitigate negative transfer when tasks are highly specialized [64] (e.g., coherence vs. consistency), the intermediate dataset focused on general QA reasoning rather than early exposure to dimension-specific evaluative patterns. For this reason, a single-stage multi-task strategy was opted for rather than sequential dimension introduction.

Hyperparameter	Value
Batch Size	4
Gradient Accumulation	4 steps (effective batch size = 16)
Learning Rate	5e-6
Weight Decay	0.01
Learning Rate Scheduler	Cosine
Warmup Ratio	0.03
Epochs	30
Max Gradient Norm	1.0
Precision	BF16

Table 7: Key hyperparameters used during the intermediate multi-task learning stage of QAG module training.

3.5.1.2 Stage 2: Dimension-Specific Fine-Tuning

Once the model acquired general instruction reasoning capabilities, it was fine-tuned on a dataset of synthetic QA pairs specifically aligned with the four evaluation dimensions: consistency, relevance, fluency, and coherence (see Table 5). These QA pairs were generated using GPT-4.1 with handcrafted instruction prompts designed to elicit dimension-aware questions and answers based on summaries and their corresponding source texts. Each summary was expanded into four samples, one per dimension, and formatted with FLAN-style instructions to maintain alignment with the previous training stage.

This stage used a higher learning rate (1e-4) and a larger batch size (16) compared to Stage 1, combined with a shorter training duration (5 epochs) and 500 warmup steps. These choices allowed the model to rapidly adapt to dimension-specific patterns without overwriting the general-purpose capabilities acquired during intermediate training. Cosine scheduling, weight decay (0.01), and BF16 precision were again used to maintain stability and efficiency. Gradient clipping (1.0) ensured training robustness, and the absence of gradient accumulation (effective batch size = 16) accelerated fine-tuning on available hardware.

The decision to treat dimension alignment as a separate fine-tuning stage, rather than integrating it into Stage 1, was motivated by the desire to isolate general reasoning from evaluative judgment. This separation allowed for more controlled learning of quality-specific

behavior, avoiding early anchoring on any single evaluative dimension. Unlike UniEval, which explores continual training to incrementally introduce dimensions, this approach separates general reasoning from evaluative specialization by first training on broad instruction tasks and then refining the model through focused fine-tuning. This two-stage strategy enables the QAG module to perform both general and evaluative reasoning in a structured and interpretable manner.

Hyperparameter	Value
Batch Size	16
Gradient Accumulation	1 (effective batch size = 16)
Learning Rate	1e-4
Weight Decay	0.01
Learning Rate Scheduler	Cosine
Warmup Steps	500
Epochs	5
Max Gradient Norm	1.0
Precision	BF16

Table 8: Hyperparameters used during the dimension-specific QA fine-tuning stage of the QAG module.

3.5.2 Scoring Evaluator Training

The final stage of the QAG-Eval pipeline focuses on training the scoring evaluator, a regression model designed to assign real-valued quality scores (1.0–5.0) to summaries based on structured QA reasoning traces produced by the QAG module. Unlike earlier stages that focus on generation, this stage emphasizes learning alignment with human judgments.

Supervision was provided by two annotated benchmarks: SummEval and DialSummEval (see Table 6). Each dataset contains multiple model-generated summaries per source, rated across consistency, relevance, fluency, and coherence. The QAG module was used to generate one QA pair per dimension, forming structured input traces. The average of multiple human scores per dimension served as the regression target.

As described in Section 3.4.3, the number of unique annotated summaries was limited. To expand coverage and improve generalization, the dataset was augmented through controlled paraphrasing: multiple QA traces were generated per summary–dimension pair using varied GPT-based prompts. This resulted in a final dataset of 140,000 examples and encouraged **score-invariant reasoning**, where different formulations of the same evaluative logic receive the same score.

The evaluator was implemented using DeBERTa-v3-Large, selected for its strong semantic encoding and regression performance. A single regression head was shared across dimensions, leveraging structured inputs that combine the generated question, answer, and original summary. The model was trained using MSE loss, with early stopping based on Spearman correlation to prioritize rank consistency. To address score imbalance, class weighting was applied during training.

To explore how training strategy impacts performance, two variants of the scoring evaluator were developed: one trained via multi-task learning across all dimensions simultaneously, and one trained using a sequential continual learning setup with replay from previous dimensions.

3.5.2.1 Training Strategies: Multi-Task vs. Continual Learning

While the scoring evaluator in QAG-Eval is designed to operate over a unified regression dataset with consistent supervision (1–5 Likert scores), this framework also investigates two alternative training strategies to assess whether training dynamics affect alignment with human judgment. Both strategies use the same core data (i.e., structured (Q, A, summary) triplets paired with human scores), but organize the data differently and apply distinct optimization approaches.

Multi-Task Learning Setup

The primary training strategy described earlier employs a multi-task setup, where the model is jointly trained on samples from all four evaluation dimensions: fluency, coherence, consistency, and relevance. Each sample includes a question-answer pair specifically targeted at one dimension, but the training objective is shared across all dimensions using a single regression head.

This strategy allows the model to learn general representations that span different quality dimensions simultaneously. It treats dimension-specific variation as part of the overall input distribution, relying on shared capacity to capture both universal and specialized evaluative signals. The multi-task setup is trained using a stratified version of the scoring dataset that preserves balance across dimensions and score bins (see Section 3.4.3).

The selected hyperparameters reflect the scoring task’s sensitivity to subtle textual signals and limited supervision scale. A smaller learning rate (8e-6) and higher batch size (32) were used to stabilize training, while the cosine learning rate scheduler with a warmup ratio of 0.2 supported smooth convergence. Gradient clipping and weight decay was applied to help control overfitting. Unlike the generative stages, FP16 precision was used to accelerate training with minimal loss in accuracy. The selected hyperparameters can be seen summarized in Table 9.

Hyperparameter	Value
Batch Size	32
Gradient Accumulation	1 (effective batch size = 32)
Learning Rate	8e-6
Weight Decay	0.02
Learning Rate Scheduler	Cosine
Warmup Ratio	0.2
Epochs	5
Max Gradient Norm	0.5
Precision	FP16

Table 9: Hyperparameters used during multi-task scoring evaluator training. The model was trained using MSE loss and Spearman correlation as the early stopping metric.

This final training stage grounds the QAG-Eval pipeline in real-world evaluation standards, enabling the model to generalize from structured QA traces to scalar predictions that mirror human scoring behavior.

Continual Learning Setup with Replay

In parallel, a continual learning strategy is explored, inspired by the approach proposed by UniEval. In this setup, the model is trained sequentially on each evaluation dimension, following a fixed ordering:

coherence → fluency → consistency → relevance

After completing training on one dimension, the model continues with the next, while retaining a 20% replay buffer from all previously seen dimensions. This ensures that past dimensions are revisited during training, mitigating the risk of catastrophic forgetting.

To support this setup, a separate set of training and validation datasets is constructed. Each training stage consists of:

- 100% of the current dimension’s data,
- +20% randomly sampled replay from each previously seen dimension.

The validation set follows the same structure, ensuring that evaluation reflects both retention and adaptation. For example, when training on consistency, the dataset consists of:

- 100% consistency samples,
- +20% coherence samples,
- +20% fluency samples.

As for the selected hyperparameters, new values were chosen compared to the multi-task learning setup. A higher learning rate is chosen (1e-5), a similar batch size (32), and a lower warmup ratio of 0.1 was used to allow the model to adapt more quickly at the beginning of each stage, since the amount of new data per stage was comparatively small. Gradient clipping and weight decay to help control overfitting. The selected hyperparameters can be seen in Table 10.

Hyperparameter	Value
Batch Size	32
Gradient Accumulation	1 (effective batch size = 32)
Learning Rate	1e-5
Weight Decay	0.01
Learning Rate Scheduler	Cosine
Warmup Ratio	0.1
Epochs	5
Max Gradient Norm	1.0
Precision	FP16
Class Weighting	Applied

Table 10: Hyperparameters used during continual scoring evaluator training. The model was trained using MSE loss and Spearman correlation as the early stopping metric.

Training Objectives and Supervision

Both training strategies use identical supervision, mean human scores on a continuous scale from 1.0 to 5.0, and are trained using the same model architecture (DeBERTa-v3-Large) and regression objective (MSE loss). Input formatting, tokenizer configuration, and evaluation metrics are kept constant across strategies.

This design enables a fair comparison between training strategies by keeping conditions consistent, so that any differences in performance can be attributed to how the data is organized and how the training is conducted, rather than to factors like model size or label quality.

4 Experimentation

This chapter presents the experimental evaluation of the proposed QAG-Eval framework, guided by the two research questions outlined in Section 2.6. The primary goal is to assess whether the framework meets its design objectives: providing interpretable evaluations (RQ1) and aligning well with human quality judgments (RQ2).

To this end, the chapter introduces the experimental setup, including datasets, baseline systems, and evaluation metrics. Each of the following sections then addresses one of the research questions through targeted experiments. Quantitative analyses, such as score correlation and distribution comparisons, are complemented by qualitative examples that illustrate how the framework behaves in different evaluation scenarios.

4.1 Evaluation Overview

This section outlines the experimental design used to evaluate the QAG-Eval framework across interpretability and scoring accuracy. It complements the dataset creation and formatting processes detailed in Section 3 by describing how the datasets were grouped for training and evaluation, which comparative baselines and metrics were used, and the technical environment in which experiments were conducted.

4.1.1 Datasets

This section introduces the evaluation dataset used across all experiments in this chapter, including the test set for QAG-Eval and the retraining dataset for UniEval. While the full corpus was divided into training, validation, and test splits, only the held-out test set is discussed here. Score distributions for the other splits are available in Section B.1..

4.1.1.1 Evaluation Dataset

The evaluation dataset consists of a held-out test set constructed from the combined SummEval and DialSummEval datasets. These datasets were merged and stratified into training, validation, and test splits using a sampling strategy based on dataset source and mean human-annotated quality scores. This ensured representative coverage across the full range of summary qualities.

The final test set contains 289 unique samples, balanced across the two sources. This same test set is reused consistently across all experiments in this chapter to support fair comparison between models.

Figure 14 shows the distribution of human-annotated scores across the four quality dimensions in the test set. Most scores fall within the 3–4 range, highlighting the importance of accurately capturing mid-quality variation in summary evaluation.

Section B.1.

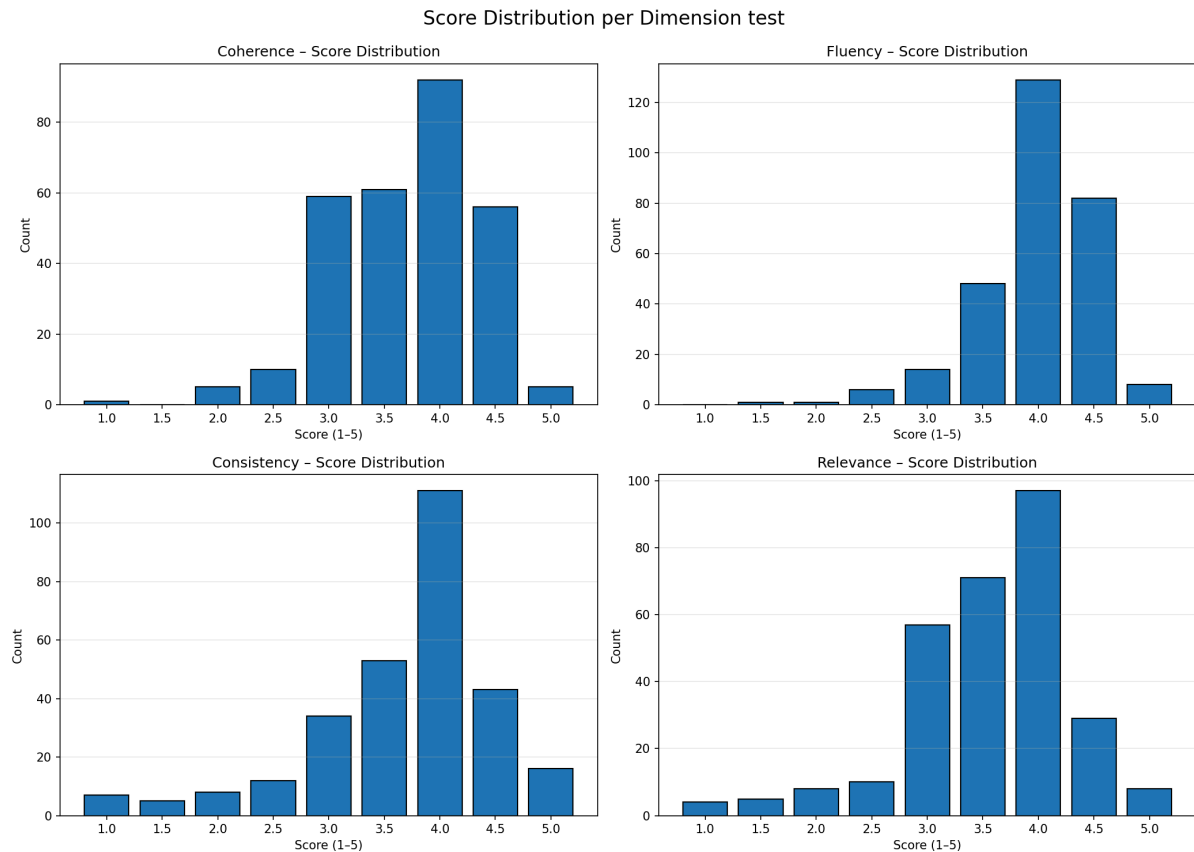


Figure 14: Each histogram displays the frequency of the human-annotated scores (1-5) for each of the quality dimensions.

4.1.1.2 Retraining Dataset for UniEval

To support a fair and direct comparison, the UniEval baseline is retrained on the same source data used for QAG-Eval. However, due to architectural and training constraints, UniEval operates in a Boolean QA format, where supervision is framed as a binary classification task (i.e., “Yes” or “No”).

Unlike QAG-Eval, which uses scalar 1–5 scores as direct supervision, UniEval is trained on synthetically generated binary-labeled data, following the original methodology outlined by the UniEval authors. Specifically:

- Positive samples (“Yes”) are constructed from the original (summary, source) pairs. These examples are assumed to represent high-quality outputs and are paired with dimension-specific yes/no questions, resulting in a “Yes” label.
- Negative samples (“No”) are created by applying dimension-specific pseudo-data transformations, adapted from UniEval’s official implementation. These include:
 - Disfluency transformations for fluency (e.g., introducing broken syntax),
 - Incoherence transformations for coherence and consistency (e.g., sentence swapping or retrieval-based perturbation),
 - Irrelevance transformations for relevance (e.g., injecting unrelated content).

Each example is framed using the original task-specific prompt templates provided in the UniEval codebase. The final dataset contains one QA pair per dimension per sample, with a corresponding binary label.

At inference time, UniEval produces softmax-normalized probabilities for “Yes” and “No” answers to each quality question. The final score is derived as the normalized confidence of a “Yes” response:

$$s_i = \frac{P(\text{Yes})}{P(\text{Yes}) + P(\text{No})}$$

This results in a continuous value in the 0-1 range, which can be interpreted as a probabilistic judgment of quality. No explicit mapping to the 1–5 Likert scale is performed in the default UniEval setup.

To confirm supervision balance, a visualization will be made for:

- The number of Boolean QA samples per dimension,
- The distribution of “Yes” vs. “No” labels across dimensions.

These plots ensure that the retraining process is based on a structurally similar dataset, though architecturally adapted, enabling a controlled and meaningful performance comparison with QAG-Eval.

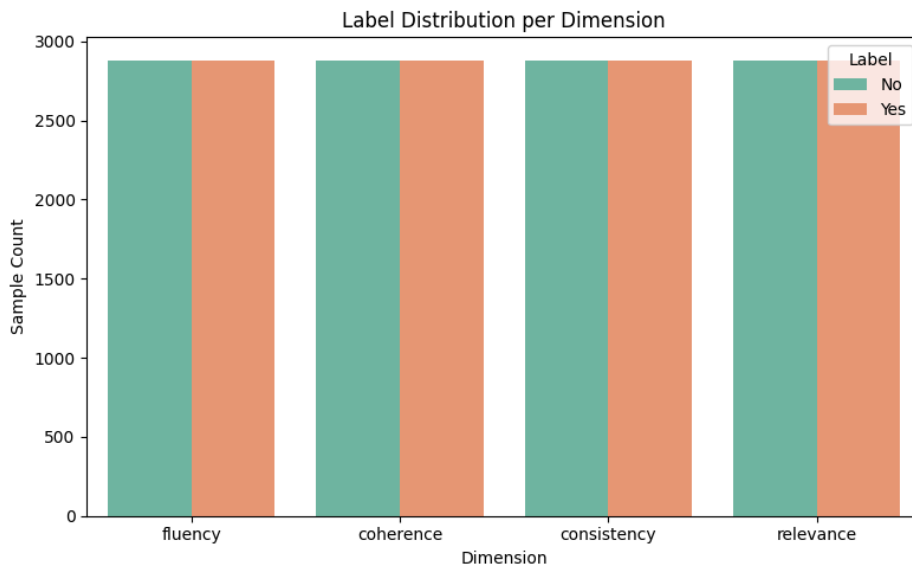


Figure 15: Balanced label distribution across evaluation dimensions in the UniEval dataset. Each dimension contains an equal number of “Yes” and “No” QA pairs, ensuring binary supervision is uniformly distributed.

The plots shown in Figure 15 indicate that the retrained UniEval dataset achieves a comparable structure to that of QAG-Eval, with balanced supervision across all quality dimensions and an even distribution of binary “Yes” and “No” labels. This setup helps ensure that any observed performance differences are primarily due to differences in evaluation architecture or methodology, rather than disparities in data distribution.

4.1.2 Baselines and Comparative models

To assess the performance of the proposed QAG-Eval framework, a set of representative evaluation models and metrics was selected. These fall into three broad categories: similarity-based metrics, learned evaluators, and multi-dimensional evaluators. It is important to note that not all baselines are included in every experiment. Instead, evaluations are separated by research question, with baselines chosen to match the nature of the comparison and training assumptions.

- **Similarity-based Metrics:**

The following commonly used metrics are included as standard general-purpose baselines. These methods require access to human-written reference summaries and primarily rely on surface-level or embedding-based similarity:

- **ROUGE-1/-2/-L** (Lin, 2004 ROUGE): is a family of lexical overlap metrics that compare the n-grams in a candidate summary against a reference summary. ROUGE-1 measures unigram (single word) overlap, providing a coarse but often informative signal of content presence. ROUGE-2 measures bigram (two-word) overlap, capturing short phrasal matches and offering a more fine-grained signal of fluency and coherence. ROUGE-L computes the length of the Longest Common Subsequence (LCS), capturing matches that preserve word order while allowing for gaps. This variant is less rigid than strict n-gram matching and is often seen as a proxy for fluency.
- **UniEval** (Zhong et al., 2022) is a unified model that formulates multi-dimensional evaluation as a Boolean QA task. UniEval generates binary answers (Yes/No) to dimension-specific questions and produces a confidence score, which is normalized to obtain a probabilistic estimate of quality. This score serves as its continuous output for comparison against scalar human ratings. For this thesis, UniEval is retrained on the same combined dataset as QAG-Eval to ensure a fair and consistent comparison. Both the multi-task and continual learning training variants are included.
- **QAG-Eval** (Proposed evaluation framework): The proposed framework is evaluated in two variants, similar to UniEval: one model trained using continual learning across dimensions and the other trained using multi-task learning. Both use the same data and training protocol as the retrained UniEval models.

Motivation for Baseline Selection

The selected baselines represent a broad range of automatic evaluation strategies, including lexical and semantic methods. ROUGE serve as standard benchmarks for surface and embedding-level similarity. UniEval, as a task-specific, learned evaluator, provides the most direct point of comparison with QAG-Eval.

In particular, the inclusion of UniEval as a retrained baseline is intended to enable a direct and fair comparison between evaluation frameworks. In the original UniEval paper, the model was benchmarked against these same similarity-based metrics using a separate training setup. In this work, UniEval is retrained on the exact same dataset used to train QAG-Eval, allowing for a controlled “apples-to-apples” comparison. This setup ensures that any observed performance differences between the two models can be attributed to differences in architectural design and reasoning methodology, rather than disparities in data exposure or training supervision.

4.1.3 Evaluation Metrics

Model predictions are evaluated against human-annotated quality ratings across the four quality dimensions: consistency, relevance, fluency, and coherence, using both rank- and regression-based metrics. All evaluations are performed across three levels of granularity:

- Sample-level: Correlation on individual summary predictions
- Summary-level: Correlation across system summaries for a single source
- System-level: Correlation on overall means for each summarization model

For measuring alignment with human judgment, the following correlation metrics are used:

- Spearman rank correlation (ρ): Evaluates monotonic agreement, reflecting whether the model ranks outputs in the same order as human annotators.
- Kendall's tau (τ): Assesses pairwise ranking similarity, providing a stricter alternative to Spearman.

A detailed explanation of each metric and its purpose is provided in Chapter 3, Section 3.2.

4.2 Framework Interpretability (RQ1)

This section describes the evaluation strategy for assessing whether the proposed QAG-Eval framework offers more interpretable and transparent quality assessments than black-box evaluation models such as UniEval. The evaluation addresses RQ1, formulated as:

RQ1: To what extent does the proposed evaluation framework provide interpretable and transparent quality judgments compared to black-box models such as UniEval?

4.2.1 Motivation

Interpretability is a critical property for evaluation systems, especially in applications where model outputs may impact decision-making, user trust, or downstream learning processes. While UniEval produces soft binary scores based on internal model confidence, it does not expose any intermediate reasoning or provide insight into *why* a summary is judged to be either consistent-inconsistent, relevant-irrelevant, fluent-nonfluent, or coherent-incoherent.

In contrast, QAG-Eval is explicitly designed for interpretability. It first generates dimension-specific questions about a summary, then answers them in a natural language using a reasoning-capable QA module. These structured QA pairs represent explicit evaluative judgments that are passed to a separate scoring model. This modular pipeline not only enables traceability but also surfaces the rationale behind each score in a human-readable form. Figure 16, shows an example of a question and answer generated by the QAG module.

Summary: paul merson has restarted his row with andros townsend . the tottenham midfielder was brought on with only seven minutes remaining in his team 's 0-0 draw with burnley . townsend was brought on in the 83rd minute for tottenham as they drew 0-0 with burnley .

Question: Are there any grammatical or stylistic errors in the summary that affect its overall fluency and readability?

Answer: Yes, the summary contains several grammatical and stylistic errors that affect its fluency and readability. The sentences are fragmented and lack proper connectors, such as missing articles (*"paul merson has restarted his row"* should be *"Paul Merson has restarted his row"*). Additionally, the phrase *"with only seven minutes remaining in his team's 0-0 draw with burnley"* is awkward and unclear, disrupting the flow. Overall, the summary would benefit from more varied sentence structures and clearer phrasing to improve readability..

Figure 16: Example of a fluency QA pair generated by the QAG module. The summary is analyzed using a dimension-specific question about grammatical and stylistic quality, followed by a reasoning-based answer that identifies specific issues affecting fluency and readability.

4.2.2 Evaluation Objectives

Since direct human assessment of interpretability is outside the scope of this thesis, interpretability is evaluated using structural and content-based analysis of the model’s outputs. The evaluation focuses on the following criteria:

1. Trace Completeness
 - Confirm that each dimension score is accompanied by a corresponding question and natural language answer (QA trace).
2. Explanation Depth
 - Compute the average token length of generated answers per dimension.
 - Identify short (< 13 token) answers that may reflect minimal or binary reasoning.
3. Justification Content
 - Analyze the presence of factual references, named entities, and linguistic markers in each answer
4. Qualitative Output Examples
 - Present a small set of representative examples showing reasoning traces for summaries of varying quality.
5. Modular Transparency
 - Discuss how QAG-Eval’s pipeline enables human-readable reasoning steps compared to black-box scoring systems.

4.2.3 Evaluation Procedure

- Sample selection: A stratified subset of summaries from the held-out test set was used to demonstrate qualitative differences across varying quality levels (low, mid, high).
- Traceability Analysis: Each summary was checked for the presence of four complete QA traces, one per evaluation dimension.
- Quantitative Analysis: Token lengths and justification cues were computed for all generated answers. Dimensions were compared based on average reasoning length and justification frequency.
- Qualitative Showcase: Three representative examples were selected to illustrate the nature of dimension-specific reasoning in QAG-Eval’s outputs.

4.3 Scoring Accuracy and Granularity (RQ2)

This section describes the evaluation strategy for assessing how accurately and finely the QAG-Eval framework reflects human quality judgments compared to existing evaluation methods, specifically UniEval and standard similarity-based metrics. The evaluation addresses RQ2, formulated as:

RQ2: How well do the scores predicted by the proposed framework correlate with human annotations, and how effectively does it distinguish between intermediate levels of summary quality compared to UniEval and standard metrics?

4.3.1 Motivation

While UniEval converts binary “Yes”/“No” answers into soft probabilistic scores, as previously discussed in Chapter 1. Section 2.4.2.2, this binary framing may oversimplify nuanced or ambiguous quality differences in human summaries, particularly in borderline cases. Consider Figure 17, an example dialogue and summary about a company’s quarterly financial results:

CEO: Our revenue increased by 15% this quarter. That’s mostly due to streamlining operations and cutting non-essential spending.

CFO: Yes, the saving from operational adjustments were significant. We’ve also seen a small improvement in margins, though sales figures remained flat.

(a) Dialogue excerpt illustrating a key financial statement and its context.

The CEO announced that revenue increased significantly, and the company plans to expand into new markets. The CFO also mentioned an improvement in operating margins.

(b) Summary example that omits a critical causal detail while maintaining surface fluency.

Figure 17: Examples of instruction-based training prompts from the intermediate multi-task learning phase. Each prompt is constructed from the same QReCC sample and used independently for question or answer generation.

The summary is arguably quite fluent and accurate in tone, but it omits a key detail: the source of the revenue increase (cost-cutting rather than sales growth), which is a critical piece of context. Human annotators might rate the summary as 3 or 4, acknowledging its surface coherence and fluency while penalizing the omission under the consistency or relevance dimension. A binary QA evaluator like UniEval must answer a general question such as: *“Is this summary consistent with the document?”* and is therefore likely to produce a confident “Yes” answer. This can be reflected as a score of 4–5, effectively overlooking the partial inconsistency. In contrast, QAG-Eval, with its dimension-specific prompting and reasoning outputs, can generate a “Yes”/“No” answer along with an explanation that explicitly accounts for what is missing. The reasoning module might conclude that the summary is only partially consistent with the source document. As such, the dimension-specific score can more faithfully reflect the partial omission through a value in the range of 3.0–4.0.

4.3.2 Evaluation Objectives

In order to test this hypothesis, the following objectives will be pursued:

1. Score Correlation With Human Annotations

- Compute Spearman and Kendall correlations between predicted scores and human labels across each evaluation dimension.
- Evaluate separately for each quality dimension (consistency, relevance, fluency, coherence).
- Compare QAG-Eval’s correlation against:
 - A retrained UniEval model (continual and multi-task variant).
 - Standard metrics: ROUGE-1, ROUGE-2, ROUGE-L.

2. Score Distribution Analysis

- Analyze the spread and shape of predicted scores across the full human rating scale (1.0–5.0), focusing on how models distribute scores rather than just rank them.
- Compare model and human distributions to assess how closely each model reflects the range, density, and central tendencies of human scoring.
- Visualize model vs. human distributions using histograms across all four evaluation dimensions.

4.3.3 Evaluation Procedure

- **Dataset:** The models will be trained on a combined SummEval and DialSummEval dataset, which includes 1-5 scale human quality ratings from multiple annotators. The models will be tested on a separate held-out test dataset from SummEval and DialSummEval.
- **Dimensions:** Evaluation is performed across all four quality dimensions using both QAG-Eval and UniEval variants (continual and multi-task).

4.4 Experiment Results

This chapter presents the empirical results of the experiments described in Section 4. The findings are grouped by research question and reported using objective criteria, including direct outputs, correlation scores, distribution patterns, and quantitative statistics. Interpretive or subjective analysis is deferred to Section 5.

4.4.1 RQ1: Interpretability Results

This section presents objective results regarding the interpretability of outputs generated by the QAG-Eval framework. The analysis includes quantitative statistics on QA trace coverage, explanation lengths, and justification content, followed by representative examples of reasoning outputs. All findings are reported without subjective interpretation or evaluation.

To assess the traceability of QAG-Eval outputs, the percentage of summaries in the evaluation set that include a complete set of four QA traces, one for each evaluation dimension, was computed. Full QA trace coverage is considered a prerequisite for transparent, modular evaluation.

Analysis confirms that 100% of system-generated summaries in the test set are accompanied by a complete set of dimension-specific QA traces. This outcome is not incidental, but a direct result of the framework’s design, which enforces modular question generation and answering as part of the scoring pipeline. As such, we can confirm that QA trace coverage is complete:

100% of system-generated summaries include a full set of four QA traces

The presence of a QA trace for each evaluation dimension ensures that every predicted score is explicitly backed by a structured, interpretable reasoning step. Unlike black-box scoring models that produce numerical outputs without rationale, QAG-Eval’s full traceability provides transparency into how and why each dimension was evaluated in a particular way. This completeness also facilitates downstream analysis, such as explanation quality filtering or error diagnostics, and supports integration in high-stakes contexts where reasoning transparency is critical.

4.4.1.1 Explanation Lengths

The token length of each answer generated by the QAG module was computed to serve as a proxy indicator of explanation depth. While not a direct measure of quality, longer responses are more likely to reflect greater elaboration, detail, and reasoning, whereas shorter answers often indicate limited justification or surface-level assessments. Short answers tend to reflect categorical judgments (e.g., “Yes.” or “No.”), whereas longer responses are more likely to contain supporting arguments, references to the source content, or explicit descriptions of identified issues.

Based on empirical observations, it is assumed that answers in the range of 3–12 tokens generally lack sufficient explanatory depth and may not qualify as fully reasoned traces. This assumption is supported by a controlled tokenization test using the FLAN-T5 tokenizer, which showed that even extended “Yes”/“No” responses rarely exceed 12 tokens. Therefore, token length serves

not only as an approximation of explanation complexity, but also as a useful diagnostic for identifying minimal or incomplete reasoning outputs. Table 11 reports average token lengths per dimension, while Table 12 provides illustrative tokenization examples.

Dimension	Avg. Token Length
Consistency	87.03
Relevance	101.71
Coherence	88.43
Fluency	81.22

Table 11: Average token length of QAG-Eval answers grouped by evaluation dimension. Longer responses are more likely to reflect deeper justification and content traceability.

Example answers	Token Length
“Yes, I agree with the statement.”	10
“No, I do not think that is correct.”	12
“Yes, it is a valid point.”	10
“No, I disagree with that conclusion.”	10
“Yes”	3
“No”	3

Table 12: Token length of typical categorical “Yes”/“No” answers as processed by the FLAN-T5 tokenizer. These examples support the threshold assumption that answers under 13 tokens likely lack sufficient explanation depth.

These results show a clear difference in average explanation length across evaluation dimensions. Relevance answers have the highest average token count (101.71), followed by coherence (88.43), consistency (87.03), and fluency (81.22). This variation reflects the different types of information involved in each evaluation task. Relevance assessments frequently involve references to multiple content points and their inclusion or omission, which may result in longer textual reasoning. Coherence and consistency involve structural and factual alignment, which also require multi-sentence reasoning. In contrast, fluency assessments are typically confined to surface-level grammar and phrasing, and are often expressed in shorter spans. In all cases, the average token length exceeds the 3–12 token threshold associated with minimal or categorical responses, indicating that QAG-Eval consistently generates answers containing more than basic binary confirmation.

4.4.1.2 Justification Presence/Content

To assess the types of justification present in QAG-Eval’s generated answers, each response was automatically analyzed for three distinct categories of interpretability cues:

1. Entity Mentions

Named entities, such as people, places, organizations, dates, and other proper nouns, were identified using spaCy [65], an open-source Python library for advanced NLP. Specifically, the `en_core_web_trf` model was used for named entity recognition (NER). Any answer containing at least one entity with a recognized label (e.g., PERSON, ORG, GPE, DATE) was counted as containing an entity reference. spaCy is a widely known library recognized for its comprehensive set of advanced NLP tools and techniques, and used in various industry use-cases

2. Factual Content Indicators

Factual grounding was detected through the presence of lexical indicators suggesting inclusion or omission of source content. The keyword set included terms such as *mention*, *includes*, *omits*, *missing*, *covers*, *details*, and *relevant*. These were matched in a case-insensitive manner using simple string pattern checks. Answers containing one or more of these keywords were considered factually grounded.

3. Linguistic Cues

Linguistic or fluency-related justifications were detected through the presence of terms referring to grammar, phrasing, sentence structure, or readability. The keyword set included *awkward*, *smooth*, *grammar*, *natural*, *readable*, and *phrasing*. These cues typically reflect stylistic judgments and are characteristic of the fluency dimension. As with factual cues, keyword presence was determined via substring matching in lowercase text.

The proportion of answers containing each justification type was computed for each evaluation dimension and is summarized in Table 13.

Dimension	Entity Mention (%)	Factual Content (%)	Linguistic Cue (%)
Coherence	91.3	23.8	49.7
Consistency	95.5	67.1	0.7
Fluency	22.7	2.8	100.0
Relevance	95.1	99.0	0.0

Table 13: Proportion of QAG-Eval answers containing named entities, factual indicators, or linguistic cues. The distribution reflects each dimension’s focus: fluency traces contain fluency markers, while relevance and consistency contain source-grounded references.

These results reveal clear patterns across dimensions. Fluency-related answers almost universally include linguistic cues (100.0%) but rarely contain factual references or named entities. In contrast, relevance and consistency answers exhibit high rates of entity mentions (95.1% and 95.5%, respectively) and factual indicators (99.0% and 67.1%). Coherence answers often refer to entities but contain a more mixed presence of stylistic terms and lower factual density, reflecting the dimension’s more abstract evaluative focus.

4.4.1.3 Qualitative Reasoning Trace Showcase

To illustrate the nature and diversity of QAG-Eval’s generated reasoning traces, three representative examples were selected from the evaluation set. These examples correspond to a relatively lower-scoring, a mid-range, and a higher-scoring summary, based on the model’s predicted overall quality scores. Each case includes the source context (either document or dialogue), the system-generated summary, and the corresponding QA pairs produced by QAG-Eval for each evaluation dimension: consistency, relevance, coherence, and fluency.

This showcase is intended to demonstrate how QAG-Eval produces interpretable outputs tailored to each dimension. The selected examples exhibit a range of justification styles, including factual references, structural explanations, and surface-level linguistic feedback. Each showcase can be found in Appendix B.

These examples illustrate how QAG-Eval produces dimension-specific reasoning traces that vary in specificity and elaboration based on summary quality. All evaluated examples contain structured QA outputs with natural language justifications, supporting the framework’s design objective of interpretability and transparency.

4.4.2 RQ2: Scoring Accuracy & Granularity Results

This section presents the quantitative evaluation results for QAG-Eval and baseline methods in terms of scoring accuracy and granularity, addressing RQ2.

To assess these qualities Spearman and Kendall correlations are reported across the four quality dimensions. Results are reported at three levels of correlation: sample-level, summary-level, and system-level, previously described in Section 3.2. Each QAG-Eval variant (multi-task and continual learning) is compared against retrained UniEval baselines (multi-task and continual learning) and a suite of reference-based similarity metrics.

All models are evaluated on a held-out test set derived from DialSummEval and SummEval. The subsections below report correlation performance and score distribution behavior, followed by a discussion of trends and comparative observations.

4.4.2.1 Correlation with Human Judgments

This section presents the alignment between model-predicted quality scores and human annotations, evaluated at three correlation granularities: sample-level, summary-level, and system-level. Results are reported separately for each of the four evaluation dimensions: consistency, relevance, fluency, and coherence, and Spearman (ρ) and Kendall’s τ correlations.

Correlation tables are reported in full in Appendix C. Below, key observations are highlighted from the results.

Sample-Level Correlation

As previously mentioned in Section 3.2.3, UniEval introduced sample-level correlation as a measure for determining the correlation between a model’s predicted score and the corresponding mean human rating for each individual summary. This correlation level offers a fine-grained, instance-specific perspective on how closely the model approximates human judgments at the per-summary level. A higher correlation at this level indicates the metric is well-calibrated for detecting subtle quality differences across summaries, regardless of their source system.

The results for the sample-level correlation for all metric types are presented in Appendix C, Table 14. Each model’s predicted quality scores are compared to the mean human annotation on a per-summary basis across the four quality dimensions, with both Spearman (ρ) and Kendall’s tau (τ).

Among the similarity-based metrics, ROUGE-1 achieves the highest average correlation ($\rho = 0.355, \tau = 0.243$), followed by ROUGE-2 and ROUGE-L. Across dimensions, all ROUGE variants perform best on Coherence and Fluency, while showing consistently low correlation with Consistency and Relevance. Specifically, ROUGE-L obtains the highest correlation for Coherence ($\rho = 0.385, \tau = 0.274$) and Fluency ($\rho = 0.325, \tau = 0.226$), while ROUGE-2 leads in Consistency ($\rho = 0.190, \tau = 0.130$) and ROUGE-1 slightly outperforms the others in Relevance ($\rho = 0.182, \tau = 0.125$). These findings suggest that ROUGE’s surface-level lexical overlap is better suited for capturing surface-level properties, such as grammaticality and sentence-level fluency, but is less effective for deeper semantic properties like factual consistency or content relevance.

The multi-dimensional evaluators demonstrate more varied performance. Both the continual and multi-task variants of the retrained UniEval model perform comparably, though their correlations are generally low across all dimensions. Coherence and Consistency exhibit the weakest alignment, where both UniEval variants fall below even ROUGE-1. The multi-task variant achieves slightly higher average correlation ($\rho = 0.364, \tau = 0.247$), while the continual

variant performs marginally better in Coherence ($\rho = 0.184, \tau = 0.127$) and Relevance ($\rho = 0.245, \tau = 0.166$), though most of these scores do not exceed the best-performing similarity-based baselines. The only dimension where both variants exceed the baselines is in Relevance, with correlations well above 0.200.

In contrast, the proposed QAG-Eval framework shows strong and consistent performance across all quality dimensions. Both the continual and multi-task variants substantially outperform all baselines, including UniEval. While the continual variant achieves the best results on Fluency ($\rho = 0.533, \tau = 0.387$), the multi-task variant achieves the highest average sample-level correlation ($\rho = 0.598, \tau = 0.435$), and outperforms all other models in Coherence ($\rho = 0.664, \tau = 0.497$), Consistency ($\rho = 0.606, \tau = 0.449$), and Relevance ($\rho = 0.612, \tau = 0.460$). These results indicate that QAG-Eval is highly effective at approximating human-perceived summary quality and is particularly well-aligned with human judgment on semantically complex dimensions such as consistency and relevance.

Summary-Level Correlation

Summary-level correlation captures the correlation between human and predicted rankings of multiple system outputs for the same input document. This measure reflects the metric’s ability to preserve human-preferred ordering when comparing summaries generated from identical source content. A high summary-level correlation suggests that the metric reliably differentiates between better and worse outputs on a per-document basis, aligning with human ranking patterns.

The results for the summary-level correlation for all metric types are presented in Appendix C, Table 15. These results assess the ranking between predicted and human scores across all system outputs for each individual source document, offering insight into whether models preserve document-specific ranking patterns. Correlations are calculated per document and averaged across the dataset.

Among the similarity-based metrics, ROUGE-L achieves the highest average correlation ($\rho = 0.260, \tau = 0.255$), particularly on Coherence ($\rho = 0.511, \tau = 0.493$) and Fluency ($\rho = 0.288, \tau = 0.288$). In contrast, correlations with Consistency and Relevance remain relatively low across all ROUGE variants. ROUGE-1/-2 performs moderately on Coherence ($\rho = 0.390, \tau = 0.376$) and ($\rho = 0.311, \tau = 0.307$), respectively. These results suggest that ROUGE—L aligns more with surface-level fluency than with semantic or factual correctness, and may better capture coherence within document-specific rankings.

Similarly to the sample-level correlation, the multi-dimensional evaluators demonstrate more varied performance. Both the continual and multi-task variants of the retrained UniEval model achieve similar performance. Coherence and Fluency exhibit the weakest alignment, where both UniEval variants fall below all baseline metrics. The multi-task variant slightly exceeds the continual variant on average correlation ($\rho = 0.213, \tau = 0.206$), while the continual variant performs moderately better in Coherence ($\rho = 0.201, \tau = 0.180$) and Relevance ($\rho = 0.238, \tau = 0.233$), but falls short on Fluency and Consistency. Most notably, the variants both exceed the baselines significantly in Consistency and Relevance.

In contrast, QAG-Eval generally shows strong and consistent performance across all quality dimensions. The continual variant moderately outperforms ROUGE-L, the best performing baseline, on Coherence and Fluency ($\rho = 0.540, \tau = 0.522$), and significantly outperforms it in Consistency ($\rho = 0.283, \tau = 0.260$) and Relevance ($\rho = 0.419, \tau = 0.412$). The multi-task variant performs slightly worse than ROUGE-L in Coherence ($\rho = 0.469, \tau = 0.451$) and Fluency ($\rho =$

0.286, $\tau = 0.270$), while outperforming it moderately in Consistency ($\rho = 0.220, \tau = 0.205$) and significantly in Relevance ($\rho = 0.420, \tau = 0.402$). Both variants generally outperform the UniEval variants by a large margin. These results suggest that QAG-Eval more effectively captures per-document level human preferences across all quality dimensions, particularly in ranking system outputs for semantic properties such as Consistency and Relevance.

System-Level Correlation

System-level correlation evaluates the correlation between model-predicted and human-assigned mean scores aggregated over all outputs from each system. This level of correlation captures how well a metric reflects overall system performance across the entire test set. A high system-level correlation suggests that the metric is effective at ranking systems in a manner consistent with human evaluators, providing the most stable and interpretable signal for leaderboard-style comparisons. However, it may mask local inconsistencies at the summary or document level, as it averages over many outputs.

The results for the system-level correlation for all metric types are presented in Appendix C, Table 16. This level of evaluation aggregates predicted and human scores across all summaries produced by each system, assessing whether the overall ranking of systems according to the metric aligns with human judgments.

Among the similarity-based metrics, ROUGE-2 achieves the highest average correlation ($\rho = 0.411, \tau = 0.301$). ROUGE-L slightly outperforms ROUGE-1 and ROUGE-2 in Coherence ($\rho = 0.660, \tau = 0.514$) and Fluency ($\rho = 0.764, \tau = 0.557$), although differences are marginal. ROUGE-2 achieves the highest Consistency ($\rho = 0.054, \tau = 0.044$) and Relevance ($\rho = 0.132, \tau = 0.092$) scores among the ROUGE metrics. Notably, ROUGE-1 and ROUGE-L produce negative correlations on Consistency ($\rho = -0.099, \tau = -0.085$) and ($\rho = -0.116, \tau = -0.071$), respectively, and ROUGE-L also exhibits negative correlation on Relevance ($\rho = -0.016, \tau = -0.014$). This reversal suggests that these metrics may reward lexical overlap patterns that do not align with human preferences for consistency or relevance.

The multi-dimensional UniEval models show mixed performance. Both variants underperform on Coherence and Fluency but perform better in Consistency and Relevance relative to the similarity-based metrics. The multi-task variant achieves the highest average correlation ($\rho = 0.451, \tau = 0.310$), but underperforms on Consistency ($\rho = -0.049, \tau = -0.018$). A similar trend is observed in Relevance, where both UniEval variants outperform ROUGE but remain well below the QAG-Eval variants.

In contrast, QAG-Eval demonstrates superior system-level alignment across all dimensions. The multi-task variant achieves the highest overall performance ($\rho = 0.733, \tau = 0.572$), with especially strong results in Consistency ($\rho = 0.840, \tau = 0.669$) and Relevance ($\rho = 0.801, \tau = 0.621$). The continual variant performs comparably well, even exceeding the multi-task variant on Relevance ($\rho = 0.825, \tau = 0.635$). These results confirm that QAG-Eval consistently captures human preferences when ranking system performance, offering significantly more reliable and semantically grounded estimates than both ROUGE and UniEval.

4.4.2.2 Score Distribution

While correlation measures assess the ranking alignment between model predictions and human judgments, they do not reveal how evaluators behave across the full scoring spectrum. This section investigates whether QAG-Eval produces fine-grained and human-aligned quality scores, with particular attention to the mid-range region (approximately 3.0–4.0), where most human scores tend to concentrate.

This range is especially challenging because summaries often exhibit a mix of strengths and weaknesses, making it difficult even for human annotators to assign a definitive quality label. Unlike clear-cut examples that deserve a 1 or a 5, mid-range cases reflect partial coverage, vague phrasing, or minor factual errors that resist binary classification. The goal is to examine whether model predictions reflect this nuanced distribution rather than collapsing toward extremes.

The analysis is structured in two parts:

- **Histogram analysis and Intermediate Scoring:**
 - Visual comparison of predicted vs. human score distributions across quality dimensions.
 - Evaluation of how models assign scores in the mid-range, where most human ratings tend to cluster
 - Assessment of how each model utilizes the 1.0–5.0 scale and whether scores are narrowly or broadly distributed
- **Raw UniEval scores:**
 - Visualization of UniEval’s output confidence scores to illustrate the impact of its binary classification framing

Histogram Analysis and Intermediate Scoring

To assess the distributions of predicted vs. human scores, grouped histograms are plotted across the four quality dimensions. A comparison of the frequency of scores (on the 1.0–5.0 scale) is made between each histogram. These visualizations provide a direct comparison between human-annotated scores and the outputs produced by the two QAG-Eval variants (UniEval is omitted from the plot because its raw scores sit on a different scale).

As shown in Figure 18, human scores (grey bars) exhibit a strong central tendency toward the mid-range (3.0–4.0), with a clear peak in the 3.5–4.0 range, reflecting the subjective and often ambiguous nature of summary quality. Most summaries are not entirely correct or incorrect but instead exhibit partial correctness, which human raters can express through intermediate scores. A well-calibrated model should reflect this tendency by avoiding collapsed predictions at the extremes or specific scores.

Both QAG-Eval variants capture this mid-range behavior effectively. Rather than clustering around a few discrete bins, both models make active use of nearly every 0.5-point interval across the full 1.0–5.0 scale. The distributions follow the general shape of human annotations, especially in the fluency and coherence dimensions.

Both models tend to overpopulate the 3.5 bin, particularly in the consistency and relevance dimensions, suggesting a mild prediction bias toward the central scores. While both variants show broadly similar patterns, the continual learning model appears to more closely mirror the overall human distribution, especially around the peaks. This may be a reflection of the benefits of the continual learning with replay training strategy. Overall, these trends suggest that QAG-Eval not only avoids collapsing to extreme or median scores, but also learns to represent intermediate levels of quality.

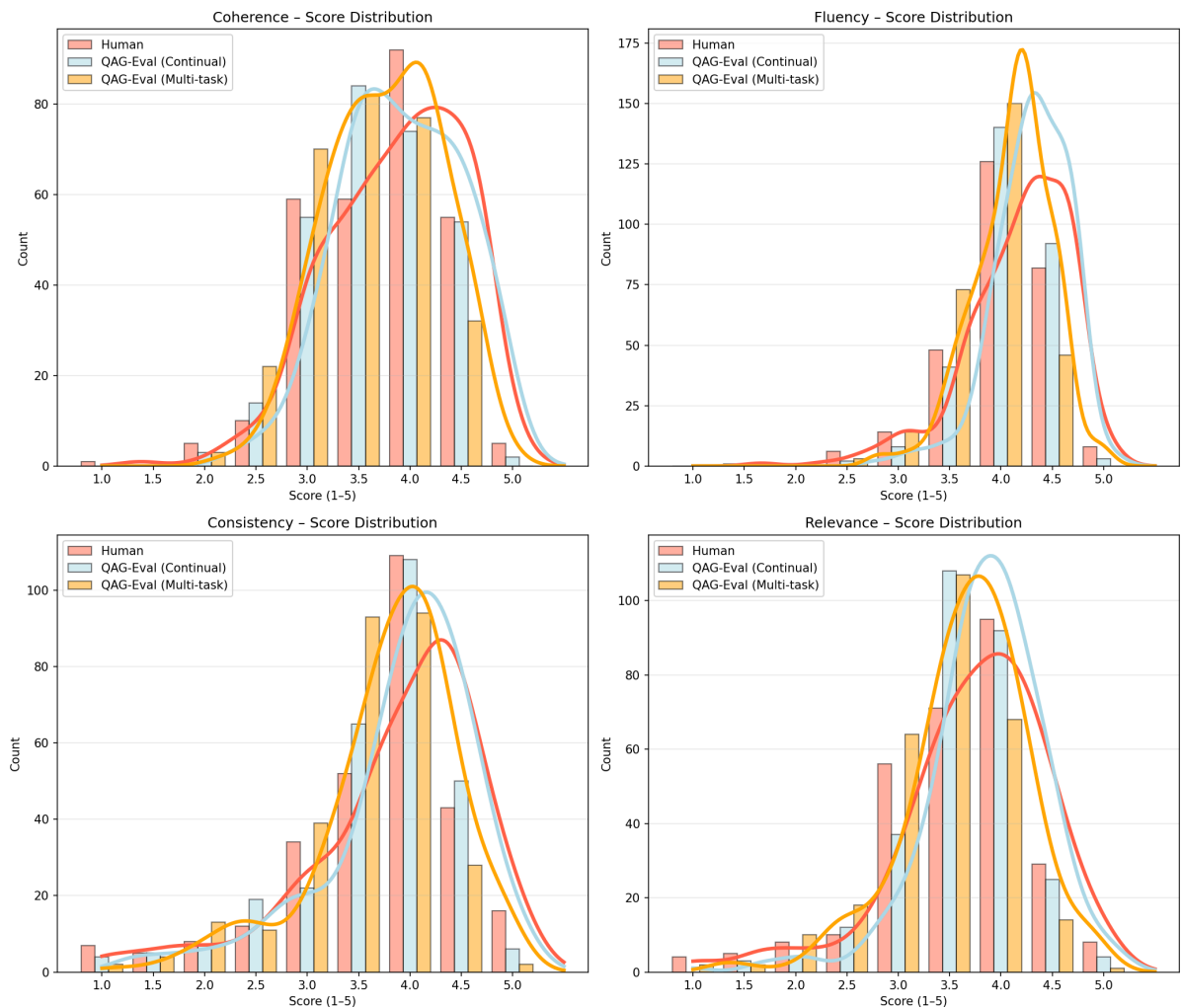


Figure 18: Grouped histograms showing predicted and human score distributions across the four evaluation dimensions. Each bar represents the number of predictions at a given score level (1.0–5.0).

Raw UniEval Scores

To better understand the limitations of binary-oriented evaluation models, this subsection analyzes the raw softmax confidence scores output by UniEval. These scores reflect the model’s internal certainty when assigning a “Yes” label for each quality dimension.

As shown in Figure 19, the raw confidence distributions for coherence, fluency, consistency, and relevance are all heavily skewed toward maximum values (≥ 0.97), with most scores concentrated in the 0.99–1.0 range. This uniform pattern is consistent across dimensions and is clearly visible, highlighting the lack of variance.

This extreme concentration reflects a key limitation of UniEval’s binary QA-based formulation. Trained to produce categorical yes/no decisions, the model tends to assign high-confidence outputs even in borderline or partially correct cases. As a result, it struggles to express degrees of quality, particularly in the ambiguous mid-range (e.g., 3.0–4.0 on a 1–5 scale) where human annotations are most densely distributed.

In contrast, QAG-Eval generates dimension-specific reasoning traces and converts them into continuous scores using a context-aware scoring mechanism. This enables it to reflect a broader and more human-like spectrum of quality, including subtle distinctions between moderately

good and clearly flawed summaries. These findings reinforce that UniEval’s scoring formulation imposes structural limitations on granularity. Its overconfident predictions and compressed output range reduce its suitability for tasks that require fine-grained quality assessment.

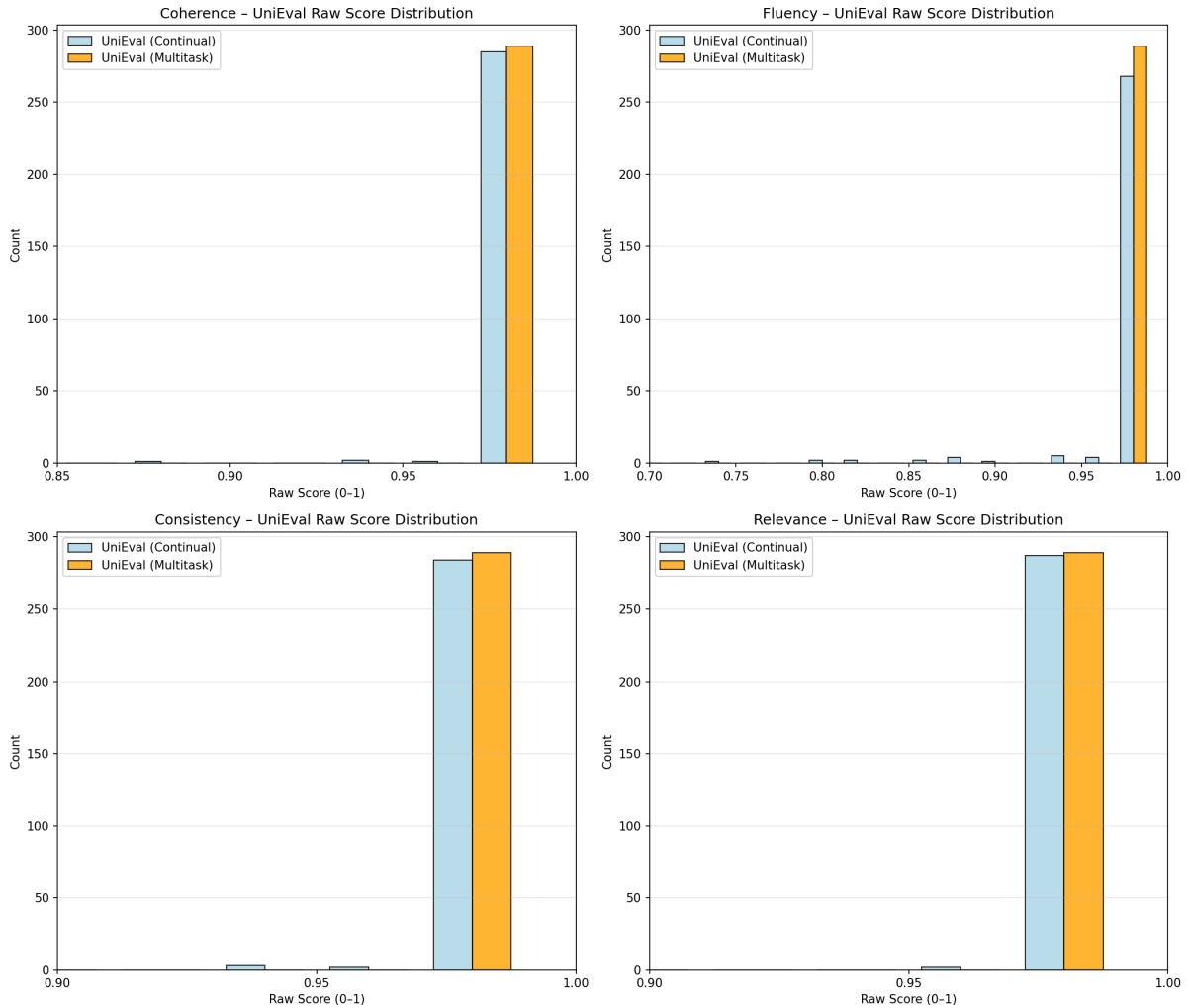


Figure 19: Each plot shows the raw softmax confidence assigned to the “Yes” label (scaled 0–1) for each quality dimension on the held-out test set. All distributions are highly concentrated near 1.0 (≥ 0.98), with minimal variation, indicating UniEval’s inability to represent partial or uncertain quality. This behavior is consistent across all four dimensions.

5 Discussion

This chapter discusses the experimental results from Chapter 4, focusing on their relevance with respect to the research questions and evaluation goals of the thesis. It provides a deeper interpretation of QAG-Eval’s empirical performance, highlights its strengths and limitations, and compares it to existing evaluation methods

5.1 Framework Interpretability

This section serves to address RQ1: To what extent does the proposed evaluation framework provide interpretable and transparent quality judgments compared to black-box models such as UniEval?

The motivation behind this research question stems from the need for evaluation frameworks to offer transparency into *why* specific quality scores are assigned, a particularly important requirement in high-stakes or diagnostic use-cases. This issue is not unique to NLP tasks such as summarization but is broadly relevant across domains. Notably, the UniEval authors (the main comparison reference) explicitly acknowledge this as a core limitation in their own design, as well as in most existing evaluators.

QAG-Eval was designed to promote interpretability through reasoning traces. Unlike black-box models such as UniEval, which output numerical confidence scores without justification. QAG-Eval decomposes each quality assessment into a dimension-specific QA pair. Each summary is evaluated across four dimensions through the generation of dimension-specific human-readable questions and explanatory answers.

5.1.1 Evaluation Outcomes

QA trace coverage

Empirical results confirm that QAG-Eval achieves 100% QA trace coverage, with every system-generated summary accompanied by a complete set of four reasoning traces. It is clear from the results shown in Appendix B that the design guarantees modular transparency and ensures that each quality score is grounded in a structured explanation. However, during implementation, it was observed that QA pairs with answers explicitly beginning with “Yes, the summary...” or “No, the summary...” introduced ambiguity. The scoring evaluator often overfit to these binary patterns, which led to degraded performance. So, while the framework design seems to imply that including additional contextual data (reasoning traces) helps the model learn the intricacies of human judgment, this needs to be further explored.

Justification Depth and Features

As mentioned previously in Section 4.2, answer token length is used as a proxy indicator for explanation depth. While not a direct measure of quality, longer responses are more likely to reflect greater elaboration, detail, and reasoning. This not only suggests deeper explanation content but also reinforces QAG-Eval’s interpretability objectives.

Results showed that average response lengths exceeded 80 tokens, significantly above typical short response lengths. Interestingly, answers in regards to the relevance dimension averaged 101.71 tokens. This suggests a high degree of justification and reasoning, with particularly extensive elaboration in semantically complex dimensions such as consistency and relevance. These dimensions require the model not only to interpret what the summary conveys but also to understand the underlying meaning of the text and verify its alignment with the source, including checks for factual entailment, contradiction, or hallucination.

These findings are further supported by the justification content analysis, which examined whether model responses included key evidential features aligned with each evaluation dimension. Specifically, the analysis focused on the presence of named entities, factual references, and linguistic markers, elements that reflect dimension-specific reasoning. For instance, relevance and consistency evaluations depend heavily on factual grounding and entity-level alignment with the source text. Without these elements, the justification may lack sufficient evidence to justify high scores for relevance or consistency. Results showed that the relevance and consistency dimension answers contained entity mentions and factual content in over 95% and 67% of cases, respectively, indicating the answers are strongly grounded in the source material. In contrast, fluency explanations exhibited 100% presence of linguistic cues related to grammar, phrasing, sentence structure, and readability, which highlights the dimension-specific nature of reasoning. Coherence answers contained factual content in approximately 23% of cases and linguistic cues in 49%, highlighting that this dimension is less concerned with factual grounding and more with how well ideas are connected and organized.

5.2 Scoring Accuracy and Granularity

This section serves to address RQ2: How well do the scores predicted by the proposed framework correlate with human annotations, and how effectively does it distinguish between intermediate levels of summary quality compared to UniEval and standard metrics?

The primary analysis was conducted across three levels of correlation: sample-level, summary-level, and system-level, each providing insight into different aspects of scoring accuracy and granularity. Sample-level correlation evaluates how closely a metric’s predicted scores align with human scores on an individual summary basis. This level is the most fine-grained and reflects the model’s ability to assign accurate scores to specific input-summary pairs. Summary-level correlation aggregates scores per summary and assesses whether the relative ranking of those summaries aligns with human rankings. This level is particularly informative in multi-system comparisons, as it captures whether a metric can consistently rank better and worse summaries across systems. System-level correlation assesses whether the metric can correctly rank entire summarization systems in alignment with human preferences. While often used for leaderboard evaluations, system-level correlation also offers insight into how robustly a metric captures overall performance trends across models.

Together, these levels provide a comprehensive view of a metric’s effectiveness, from individual scoring behavior to global system rankings. Strong performance across all three levels indicates that the metric not only mimics human judgments but also preserves score granularity across varying quality levels.

Additionally, an analysis of score distributions was conducted to gain deeper insight into the model’s predictive behavior and assess whether it captures the nuanced and subjective nature of human judgments, particularly within mid-range scores (3.0–4.0), without collapsing to dominant or extreme values.

5.2.1 Correlation Findings

Sample-Level Correlation

At sample-level, both QAG-Eval variants significantly outperform all baselines (ROUGE-1/-2/-L) and retrained UniEval variants across all dimensions. The multi-task QAG-Eval model achieves the highest average sample-level Spearman and Kendall ($\rho = 0.598, \tau = 0.435$) correlations, particularly excelling in coherence ($\rho = 0.664, \tau = 0.497$), consistency ($\rho = 0.606, \tau =$

0.449), and relevance ($\rho = 0.612, \tau = 0.460$), only beaten slightly by the continual variant in Fluency ($\rho = 0.533, \tau = 0.387$). Overall, these results demonstrate that QAG-Eval is highly effective at predicting scores that align closely with human judgments.

Summary-Level Correlation

At summary-level, the continual QAG-Eval variant only slightly outperforms the baseline ROUGE-L metric in coherence ($\rho = 0.540, \tau = 0.522$) and fluency ($\rho = 0.305, \tau = 0.292$), while ROUGE-L achieves ($\rho = 0.511, \tau = 0.493$) and ($\rho = 0.288, \tau = 0.288$) respectively. However, it greatly outperforms it in consistency ($\rho = 0.283, \tau = 0.260$) and relevance ($\rho = 0.419, \tau = 0.412$), showcasing QAG-Eval’s superior ability to capture and understand complex semantics. The continual variant also achieves the highest average summary-level Spearman and Kendall ($\rho = 0.318, \tau = 0.307$). Overall, these results show QAG-Eval can correctly capture and preserve the relative ranking of summaries, which aligns with human rankings.

System-Level Correlation

The strongest results appear at system-level, where the multi-task QAG-Eval variant achieves the highest average Spearman and Kendall ($\rho = 0.733, \tau = 0.572$). Generally, across all dimensions, both variants approach near-perfect alignment in Spearman ranking (above $\rho = 0.800$), demonstrating QAG-Eval can reliably rank systems in agreement with human evaluation. Remarkably, ROUGE-L also achieves high correlation in coherence ($\rho = 0.660, \tau = 0.512$) and fluency ($\rho = 0.764, \tau = 0.557$). Overall, these results show that QAG-Eval is effective at ranking systems that align with human assessment.

5.2.2 Score Distribution Insights

Beyond correlation metrics, the analysis of score distributions reveals that QAG-Eval effectively captures mid-range scoring behavior. Both QAG-Eval variants utilize the full 1–5 scale and show particularly strong alignment with the 3.5–4.0 range, where the majority of human scores are concentrated. This result is likely influenced by the distribution of the training data, which contains a higher density of mid-range scores (3.0–4.5) and relatively fewer examples at the extreme low (1.0–3.0) and high (4.5–5.0) ends. As a result, the model has more training signals in this range, making it naturally better at scoring moderate-quality summaries.

Despite this imbalance, the model still demonstrates the ability to assign scores across the full scale, including the extremes. This suggests that QAG-Eval is capable of generalizing beyond the most frequent score ranges, supporting the reliability of its scoring behavior. Incorporating reasoning traces likely contributes to this robustness by providing richer contextual signals for learning accurate input–summary score mappings.

While the multi-task variant achieves a stronger correlation with human annotations at the sample level, the continual learning variant produces a score distribution that more closely mirrors the shape of the human score distribution. Specifically, it demonstrates broader coverage across the full 1–5 scale, including the underrepresented extremes. This suggests that while the multi-task model excels in per-instance scoring accuracy, the continual variant more effectively captures the broader distributional patterns present in human judgments.

UniEval’s raw confidence distributions show they are heavily skewed toward maximum values (≥ 0.97), which could be a likely consequence of its binary scoring. This results in compressed score distributions and poor granularity. In contrast, QAG-Eval’s continuous scoring, derived from free-form reasoning, allows for more human-like scoring patterns. However, these results may also be due to incorrect reproduction of the UniEval models and incorrect data processing.

Early signs during implementation and training showcased signs of partial overfitting to the task, which may be attributed to the inherent simplicity of the task. While training losses and validation losses were steadily decreasing, even validating at quarter epochs, the models showcased perfect scoring across a variety of metrics. As such, the results shown should be taken with a grain of salt and are due for further exploration.

5.3 Comparative analysis

This section aim to provide a fair comparative analysis of the findings across baselines, UniEval, and QAG-Eval. The goal is not only to highlight QAG-Eval’s strengths, but also to critically examine why existing approaches may fall short for nuanced quality evaluation.

5.3.1 UniEval Architectural Limitations

While UniEval was retrained on the same datasets as QAG-Eval for an apples-to-apples comparison, it exhibited notable limitations in both score granularity and interpretability. Most notably, UniEval’s raw confidence distributions were heavily skewed toward high values (≥ 0.97) across all quality dimensions. This compression likely stems from its binary QA design, which frames each dimension as “Yes”/“No” questions. Such a setup inherently restricts the model’s ability to express ambiguity or partial correctness, especially in the mid-range scoring where most human judgments cluster.

However, it is important to note that these outcomes may not solely reflect architectural design. As previously mentioned, the retraining process, which relied on synthetic binary supervision, may have contributed to early overfitting. During training, both training and validation losses slowly decreased, and evaluation metrics appeared to be saturated early on after only a quarter epoch. This could suggest that the task formulation, involving perturbed positive/negative pairs, may have led to premature convergence and overconfident predictions. As such, the lack of score diversity and sensitivity may be as much a reflection of training data assumptions as of model design.

Therefore, while UniEval’s binary framing likely limits its ability to reflect nuanced or uncertain quality judgments, the models’ observed behavior under the current setup and conditions should be interpreted with caution.

5.3.2 Interpretability Gaps

Another area where QAG-Eval diverges meaningfully is in interpretability. By design, UniEval offers no explanation for its judgments, making it difficult to examine decisions, assess confidence, or identify reasoning failures. In contrast, QAG-Eval decomposes each score into structured QA pairs with natural language justifications, enabling downstream error analysis and model accountability. While UniEval’s simplicity may make it more efficient, scalable, and generalizable in some contexts, the absence of intermediate reasoning significantly limits its utility in scenarios requiring these insights.

5.4 Strengths and Limitations

This section reflects on the practical and theoretical strengths of the QAG-Eval Framework, as well as its current limitations. While the results across RQ1 and RQ2 demonstrate QAG-Eval’s capacity to generate interpretable and accurate evaluations, several caveats remain that should be addressed in future iterations.

Strengths

1. Dimension-Specific Reasoning

The core strength and contribution of QAG-Eval is its ability to reason separately about consistency, relevance, coherence, and fluency through QA reasoning traces. This modular structure enables dimension-specific evaluation, allowing the model to focus on grammar for fluency, discourse flow for coherence, factual alignment for consistency, and content coverage for relevance.

2. Fine-Grained and Human-Like Scoring Behavior

Unlike evaluators such as UniEval, QAG-Eval produces scalar outputs along a continuous 1–5 scale. These scores exhibit meaningful variation across mid-range values, which closely mirrors the subjective nature of human judgments. The scoring granularity is especially valuable in real-world use cases involving partially correct or ambiguous summaries. Moreover, QAG-Eval avoids prediction collapse by actively utilizing the full scale, with a particular strength in replicating the 3.0–4.0 range where most human annotations are shown to cluster.

3. Transparency and Interpretable Outputs

By generating explanatory answers alongside each score, QAG-Eval offers clear insight into why a summary received a particular evaluation. This interpretability supports downstream error analysis, refinement, and integration into high-stakes environments.

4. Modular and Generalizable Architecture

While the final scoring evaluator was trained primarily on dialogue and news-centered datasets (SummEval and DialSummEval), the core QAG module, which is responsible for generating QA reasoning traces, was trained on a diverse set of tasks and domains. This multi-domain pre-training equips QAG-Eval with broad reasoning capabilities, making the overall architecture modular and extensible. With additional labeled data or calibration strategies, the full framework could be adapted for other generation tasks such as translation evaluation, captioning, or more diverse question answering.

Limitations

1. Length-Sensitive Generation

The natural language justifications produced by the QA module can be verbose and vary significantly in length. While longer responses may reflect deeper reasoning, they also increase computational overhead and risk diluting the signal with unnecessary elaboration.

2. Prediction Bias

Despite the use of scalar scoring, both QAG-Eval variants exhibit a mild clustering around 3.5–4.0, especially in relevance and consistency dimensions. This is highly likely to stem from the skewed score distribution in the training data and suggests a potential need for score calibration, regularization to avoid central bias, or simply increasing the underrepresented sample scores.

3. Data and Training Demands

The framework’s modularity and traceability come at the cost of increased data and training requirements. QAG-Eval relies on large-scale instruction-style QA/QG datasets, as well as fine-tuning on domain-aligned evaluation data. This raises the barrier for deployment in low-resource or rapid prototyping settings.

4. Continual Learning Complexity

The continual learning variant, while promising in terms of score distribution generalization, introduces non-trivial training complexity. Managing replay buffers and avoiding catastrophic forgetting requires additional infrastructure, which may not be trivial to implement at scale.

5. Sensitivity to Answer Phrasing

As noted in the interpretability discussion, early experiments revealed that QA answers starting with fixed binary phrases (“Yes, the summary...” or “No, the summary...”) led to overfitting in the scoring model. This suggests that even with modularity, the phrasing and structure of generated answers significantly impact score quality, an area requiring further control or normalization.

6. Context Length Constraints

The underlying model architecture (FLAN-T5) imposes a fixed input length limit of 512 tokens, which restricts the amount of source content and summary that can be encoded together. In multi-turn dialogues or long documents, this often leads to input truncation, potentially excluding critical context required for accurate QA generation or score prediction.

7. Reliance on Synthetic QA-generation Supervision

The QAG module’s dimension-specific question and answer generation was fine-tuned on synthetic data generated by GPT-based models rather than human-created evaluative traces. While this approach enabled scalable training across a wide range of tasks and dimensions, it introduces potential artifacts. Some questions or answers may exhibit unnatural phrasing, verbosity, or model-internal reasoning patterns that differ from human evaluators. This could affect both the quality of reasoning traces and the scoring evaluator’s alignment with truly human judgment.

5.5 Future Works

While QAG-Eval demonstrates promising results in both interpretability and scoring accuracy, several areas remain for further development. These directions focus on improving calibration, generalization, and integration, as well as addressing the architectural and training limitations discussed previously.

1. Mitigating Mid-Range Score Bias

While QAG-Eval makes full use of the 1–5 scale, both variants show a mild prediction bias toward central scores, particularly in the 3.0–4.0 range. While this tendency may reflect the distribution of training data, it also highlights an opportunity to explore different calibration strategies to help the model express a fuller range of quality distinctions.

2. Unified QAG and Scoring Pipeline

The current framework operates as a two-stage pipeline, with QA generation feeding into a separate scoring model. Future work could explore a unified architecture capable of performing both reasoning and score prediction in a single pass. Such an approach could reduce inference latency, improve consistency between generated explanations and scores, and simplify deployment in hardware-constrained environments.

3. Human Supervision for QA Traces

Since the current QAG module is trained on synthetic QA data generated by GPT models, future work could incorporate human-created QA traces for fine-tuning or validation. This

would help ensure natural phrasing, reduce stylistic artifacts, and further improve alignment with human judgment.

4. Training on Multi-Domain Evaluation Data

The scoring evaluator is currently limited to two datasets, SummEval and DialSummEval, which constrains its exposure to diverse summarization styles. Incorporating additional human-annotated scoring datasets from domains such as legal, scientific, or open-domain summarization would enable better generalization and more robust performance.

5. Longer Context Support

To overcome the 512-token input limit of FLAN-T5, future iterations could adopt long-context transformer variants. This would improve the model's ability to handle multi-turn conversations or documents with extended context, reducing reliance on input truncation or aggressive summarization.

6. Generalizability and Robustness

While this thesis primarily focused on in-domain evaluation, an important open question is how well QAG-Eval generalizes to new domains or noisy inputs. Future work should assess its robustness under out-of-domain, paraphrasing, or hallucinated content.

7. Model Evaluation Comparison

This thesis focused on a targeted set of evaluative baselines, namely, the ROUGE family, UniEval, and QAG-Eval. While this allowed for controlled comparisons, future work could expand the analysis to include a broader range of reference-free metrics, learned metrics, or current state-of-the-art evaluators. This would offer a more comprehensive understanding of where QAG-Eval stands in the current evaluation landscape.

6 Conclusion

This thesis presented QAG-Eval, a modular evaluation framework designed to assess the quality of abstractive summaries through interpretable reasoning and fine-grained scoring. Motivated by the limitations of existing black-box evaluators, such as UniEval and reference-based metrics like ROUGE, QAG-Eval aims to bridge the gap between accuracy and transparency in automatic evaluation. While the QAG module itself was trained on a diverse set of domains and tasks, the full framework’s generalization to unseen domains remains an open area for future work.

The framework is built on a two-stage architecture. The first stage generates dimension-specific questions and answers that reflect human-like reasoning across four evaluation dimensions: consistency, relevance, coherence, and fluency. The second stage maps these reasoning traces to scalar quality scores along a 1–5 scale using a separate scoring model. This modular design allows for both human interpretability and machine-learned precision.

To evaluate the framework, a series of empirical experiments were conducted corresponding to two primary research questions. The first, RQ1, asked whether QAG-Eval could produce traceable and interpretable quality judgments. The second, RQ2, focused on the accuracy and granularity of QAG-Eval’s predicted scores in relation to human annotations.

Findings for RQ1 showed that QAG-Eval successfully produces a complete set of reasoning traces for every evaluated summary, with 100% QA trace coverage across all test samples. The generated answers demonstrated strong justification patterns, including high rates of factual grounding, entity mentions, and linguistic cues aligned with each dimension. The average token lengths of the answers, typically exceeding 80 tokens, further indicate that the framework generates elaborate, human-readable explanations rather than binary or overly simplistic judgments. These findings confirm that QAG-Eval fulfills its goal of providing interpretable outputs with rich, structured reasoning.

Results for RQ2 revealed that QAG-Eval achieves strong correlations with human quality judgments at the sample-level, summary-level, and system-level across all evaluation dimensions. The multi-task variant achieved the highest average sample-level Spearman and Kendall correlation ($\rho = 0.598, \tau = 435$) and performed especially well on dimensions like consistency and relevance. Additionally, the framework exhibited human-like behavior in its score distribution, effectively replicating the mid-range density (3.0–4.0) commonly observed in human ratings. In contrast, baselines such as UniEval and ROUGE either lacked granularity or failed to capture semantically grounded differences in quality.

Beyond empirical results, this thesis reflected on the framework’s key strengths and limitations. QAG-Eval’s core contributions lie in its interpretability, dimension-specific reasoning, and human-like scoring behavior. At the same time, its reliance on synthetic supervision, sensitivity to phrasing, input length constraints, and mid-range score bias highlight areas for improvement.

Several directions were identified for future work, including unified architectures for QA and scoring, improved calibration techniques, broader human-annotated training data, and long-context model support. Additionally, evaluating the framework’s robustness across domains and input conditions remains an important area for further research.

In conclusion, QAG-Eval offers a promising alternative to existing evaluation metrics by combining question-answering with scalar score prediction in a transparent and modular design. While refinements remain, this work takes a meaningful step toward building evaluation methods that not only approximate human judgment but also make it interpretable.

Bibliography

- [1] Luhn HP. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development. 1958;2(2):159–165.. doi:10.1147/rd.22.0159
- [2] Erkan G, Radev DR. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. CoRR. 2011 . <http://arxiv.org/abs/1109.2128>
- [3] Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia: Association for Computing Machinery; 1998. pp. 335–336. (SIGIR '98). <https://doi.org/10.1145/290941.291025>. doi:10.1145/290941.291025
- [4] Mihalcea R, Tarau P. TextRank: Bringing Order into Text. In: Lin D, Wu D, editors. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain: Association for Computational Linguistics; 2004. pp. 404–411. <https://aclanthology.org/W04-3252/>
- [5] Cheng J, Lapata M. Neural Summarization by Extracting Sentences and Words. In: Erk K, Smith NA, editors. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics; 2016. pp. 484–494. <https://aclanthology.org/P16-1046/>. doi:10.18653/v1/P16-1046
- [6] Nallapati R, Zhai F, Zhou B. SummaRuNNer: a recurrent neural network based sequence model for extractive summarization of documents. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, California, USA: AAAI Press; 2017. pp. 3075–3081. (AAAI'17).
- [7] Narayan S, Cohen SB, Lapata M. Ranking Sentences for Extractive Summarization with Reinforcement Learning. In: Walker M, Ji H, Stent A, editors. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics; 2018. pp. 1747–1759. <https://aclanthology.org/N18-1158/>. doi:10.18653/v1/N18-1158
- [8] See A, Liu PJ, Manning CD. Get To The Point: Summarization with Pointer-Generator Networks. In: Barzilay R, Kan M-Y, editors. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics; 2017. pp. 1073–1083. <https://aclanthology.org/P17-1099/>. doi:10.18653/v1/P17-1099
- [9] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR. 2018 . <http://arxiv.org/abs/1810.04805>
- [10] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research. 2020;21(140):1–67. <http://jmlr.org/papers/v21/20-074.html>
- [11] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al. Language Models are Few-Shot Learners. CoRR. 2020 . <https://arxiv.org/abs/2005.14165>

-
- [12] OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, et al. GPT-4 Technical Report. 2024 . <https://arxiv.org/abs/2303.08774>
 - [13] Zhang T, Ladhak F, Durmus E, Liang P, McKeown K, Hashimoto TB. Benchmarking Large Language Models for News Summarization. 2023 . <https://arxiv.org/abs/2301.13848>
 - [14] Lin C-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics; 2004. pp. 74–81. <https://aclanthology.org/W04-1013/>
 - [15] Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, Pennsylvania: Association for Computational Linguistics; 2002. pp. 311–318. (ACL '02). <https://doi.org/10.3115/1073083.1073135>. doi:10.3115/1073083.1073135
 - [16] Kirstein F, Wahle JP, Ruas T, Gipp B. What's under the hood: Investigating Automatic Metrics on Meeting Summarization. 2024 . <https://arxiv.org/abs/2404.11124>
 - [17] Rennard V, Shang G, Hunter J, Vazirgiannis M. Abstractive Meeting Summarization: A Survey. 2023 . <https://arxiv.org/abs/2208.04163>
 - [18] Kirstein F, Wahle JP, Gipp B, Ruas T. CADs: A Systematic Literature Review on the Challenges of Abstractive Dialogue Summarization. 2024 . <https://arxiv.org/abs/2406.07494>
 - [19] Rau LF, Jacobs PS, Zernik U. Information extraction and text summarization using linguistic knowledge acquisition. *Information Processing & Management*. 1989;25(4):419–428. <https://www.sciencedirect.com/science/article/pii/0306457389900691>. doi:[https://doi.org/10.1016/0306-4573\(89\)90069-1](https://doi.org/10.1016/0306-4573(89)90069-1)
 - [20] Jing H, McKeown KR. Cut and Paste Based Text Summarization. In: 1st Meeting of the North American Chapter of the Association for Computational Linguistics. 2000. <https://aclanthology.org/A00-2024/>
 - [21] Mani I, Maybury M. Advances in Automatic Text Summarization. *Computational Linguistics*. 1999;26:280–281. <https://api.semanticscholar.org/CorpusID:195600496>
 - [22] Rush AM, Chopra S, Weston J. A Neural Attention Model for Abstractive Sentence Summarization. In: Màrquez L, Callison-Burch C, Su J, editors. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics; 2015. pp. 379–389. <https://aclanthology.org/D15-1044/>. doi:10.18653/v1/D15-1044
 - [23] Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. 2019 . <https://arxiv.org/abs/1910.13461>
 - [24] Maynez J, Narayan S, Bohnet B, McDonald R. On Faithfulness and Factuality in Abstractive Summarization. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. pp. 1906–1919. <https://aclanthology.org/2020.acl-main.173/>. doi:10.18653/v1/2020.acl-main.173
-

-
- [25] Kryściński W, Keskar NS, McCann B, Xiong C, Socher R. Neural Text Summarization: A Critical Evaluation. 2019 . <https://arxiv.org/abs/1908.08960>
- [26] Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating Text Generation with BERT. 2020 . <https://arxiv.org/abs/1904.09675>
- [27] Zhao W, Peyrard M, Liu F, Gao Y, Meyer CM, Eger S. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. 2019 . <https://arxiv.org/abs/1909.02622>
- [28] Yuan W, Neubig G, Liu P. BARTScore: Evaluating Generated Text as Text Generation. 2021 . <https://arxiv.org/abs/2106.11520>
- [29] Durmus E, He H, Diab M. FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2020. <http://dx.doi.org/10.18653/v1/2020.acl-main.454>. doi:10.18653/v1/2020.acl-main.454
- [30] Scialom T, Dray P-A, Gallinari P, Lamprier S, Piwowski B, Staiano J, Wang A. QuestEval: Summarization Asks for Fact-based Evaluation. 2021 . <https://arxiv.org/abs/2103.12693>
- [31] Scialom T, Lamprier S, Piwowski B, Staiano J. Answers Unite! Unsupervised Metrics for Reinforced Summarization Models. 2019 . <https://arxiv.org/abs/1909.01610>
- [32] Zhong M, Liu Y, Yin D, Mao Y, Jiao Y, Liu P, Zhu C, Ji H, Han J. Towards a Unified Multi-Dimensional Evaluator for Text Generation. 2022 . <https://arxiv.org/abs/2210.07197>
- [33] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. CoRR. 2019 . <http://arxiv.org/abs/1910.10683>
- [34] Likert R. A Technique for the Measurement of Attitudes. Columbia university; 1932. (A Technique for the Measurement of Attitudes). <https://books.google.dk/books?id=9rotAAAAYAAJ>
- [35] Fabbri AR, Kryscinski W, McCann B, Xiong C, Socher R, Radev DR. SummEval: Re-evaluating Summarization Evaluation. CoRR. 2020 . <https://arxiv.org/abs/2007.12626>
- [36] Gao M, Wan X. DialSummEval: Revisiting Summarization Evaluation for Dialogues. In: Carpuat M, Marneffe M-C de, Meza Ruiz IV, editors. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States: Association for Computational Linguistics; 2022. pp. 5693–5709. <https://aclanthology.org/2022.naacl-main.418/>. doi:10.18653/v1/2022.naacl-main.418
- [37] 2005. <https://onlinelibrary.wiley.com/doi/abs/10.1002/0470011815.b2a15150>
- [38] Kendall MG. A New Measure of Rank Correlation. Biometrika. 1938 [accessed 2025 May 27];30(1/2):81–93. <http://www.jstor.org/stable/2332226>
- [39] Bhandari M, Gour PN, Ashfaq A, Liu P, Neubig G. Re-evaluating Evaluation in Text Summarization. CoRR. 2020 . <https://arxiv.org/abs/2010.07100>
-

-
- [40] Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, Li Y, Wang X, Dehghani M, Brahma S, et al. Scaling Instruction-Finetuned Language Models. 2022 . <https://arxiv.org/abs/2210.11416>
 - [41] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. CoRR. 2019 . <http://arxiv.org/abs/1910.10683>
 - [42] Kim S, Hong J-H, Kang I, Kwak N. Semantic Sentence Matching with Densely-connected Recurrent and Co-attentive Information. CoRR. 2018 . <http://arxiv.org/abs/1805.11360>
 - [43] Bowman SR, Angeli G, Potts C, Manning CD. A large annotated corpus for learning natural language inference. In: Màrquez L, Callison-Burch C, Su J, editors. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics; 2015. pp. 632–642. <https://aclanthology.org/D15-1075/>. doi:10.18653/v1/D15-1075
 - [44] Anantha R, Vakulenko S, Tu Z, Longpre S, Pulman S, Chappidi S. Open-Domain Question Answering Goes Conversational via Question Rewriting. CoRR. 2020 . <https://arxiv.org/abs/2010.04898>
 - [45] Narayan S, Cohen SB, Lapata M. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J, editors. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics; 2018. pp. 1797–1807. <https://aclanthology.org/D18-1206/>. doi:10.18653/v1/D18-1206
 - [46] Gliwa B, Mochol I, Biesek M, Wawer A. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. In: Wang L, Cheung JCK, Carenini G, Liu F, editors. Proceedings of the 2nd Workshop on New Frontiers in Summarization. Hong Kong, China: Association for Computational Linguistics; 2019. pp. 70–79. <https://aclanthology.org/D19-5409/>. doi:10.18653/v1/D19-5409
 - [47] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR. 2018 . <http://arxiv.org/abs/1810.04805>
 - [48] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR. 2019 . <http://arxiv.org/abs/1907.11692>
 - [49] He P, Liu X, Gao J, Chen W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. CoRR. 2020 . <https://arxiv.org/abs/2006.03654>
 - [50] Hermann KM, Kociský T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P. Teaching Machines to Read and Comprehend. In: NIPS. 2015. pp. 1693–1701. <http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend>
 - [51] Reddy S, Chen D, Manning CD. CoQA: A Conversational Question Answering Challenge. CoRR. 2018 . <http://arxiv.org/abs/1808.07042>
 - [52] Yang Z, Qi P, Zhang S, Bengio Y, Cohen WW, Salakhutdinov R, Manning CD. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. CoRR. 2018 . <http://arxiv.org/abs/1809.09600>

-
- [53] Trischler A, Wang T, Yuan X, Harris J, Sordoni A, Bachman P, Suleman K. NewsQA: A Machine Comprehension Dataset. CoRR. 2016 . <http://arxiv.org/abs/1611.09830>
 - [54] Wang Z, Hamza W, Florian R. Bilateral Multi-Perspective Matching for Natural Language Sentences. CoRR. 2017 . <http://arxiv.org/abs/1702.03814>
 - [55] Dasigi P, Liu NF, Marasović A, Smith NA, Gardner M. Quoref: A Reading Comprehension Dataset with Questions Requiring Coreferential Reasoning. In: Inui K, Jiang J, Ng V, Wan X, editors. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics; 2019. pp. 5925–5932. <https://aclanthology.org/D19-1606/>. doi:10.18653/v1/D19-1606
 - [56] Rajpurkar P, Jia R, Liang P. Know What You Don't Know: Unanswerable Questions for SQuAD. CoRR. 2018 . <http://arxiv.org/abs/1806.03822>
 - [57] Koupaei M, Wang WY. WikiHow: A Large Scale Text Summarization Dataset. CoRR. 2018 . <http://arxiv.org/abs/1810.09305>
 - [58] Chen Y, Liu Y, Chen L, Zhang Y. DialogSum: A Real-Life Scenario Dialogue Summarization Dataset. In: Zong C, Xia F, Li W, Navigli R, editors. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online: Association for Computational Linguistics; 2021. pp. 5062–5074. <https://aclanthology.org/2021.findings-acl.449/>. doi:10.18653/v1/2021.findings-acl.449
 - [59] Kim B, Kim H, Kim G. Abstractive Summarization of Reddit Posts with Multi-level Memory Networks. CoRR. 2018 . <http://arxiv.org/abs/1811.00783>
 - [60] Huang L, Cao S, Parulian N, Ji H, Wang L. Efficient Attentions for Long Document Summarization. 2021.
 - [61] Kornilova A, Eidelman V. BillSum: A Corpus for Automatic Summarization of US Legislation. In: Wang L, Cheung JCK, Carenini G, Liu F, editors. Proceedings of the 2nd Workshop on New Frontiers in Summarization. Hong Kong, China: Association for Computational Linguistics; 2019. pp. 48–56. <https://aclanthology.org/D19-5406/>. doi:10.18653/v1/D19-5406
 - [62] Sen P, Namata G, Bilgic M, Getoor L, Galligher B, Eliassi-Rad T. Collective Classification in Network Data. \AI Magazine. 2008;29(3):93. <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2157>. doi:10.1609/aimag.v29i3.2157
 - [63] Clement CB, Bierbaum M, O'Keeffe KP, Alemi AA. On the Use of ArXiv as a Dataset. 2019.
 - [64] Parisi GI, Kemker R, Part JL, Kanan C, Wermter S. Continual Lifelong Learning with Neural Networks: A Review. CoRR. 2018 . <http://arxiv.org/abs/1802.07569>
 - [65] Honnibal M, Montani I, Van Landeghem S, Boyd A. spaCy: Industrial-strength Natural Language Processing in Python. 2020 . <https://zenodo.org/records/10009823>. doi:10.5281/zenodo.1212303

Appendix A

A.1. Metric Definitions

This appendix provides formal definitions and formulas for the evaluation metrics used throughout this thesis. These metrics are applied to assess the alignment between predicted scores and human annotations of the QAG-Eval models.

A.1.1. Spearman Rank Correlation (ρ)

Spearman's ρ , also known as the Spearman rank-order correlation coefficient, assesses the monotonic relationship between predicted and true scores by comparing their ranks.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Where:

- d_i : The difference between the ranks of the i -th pair of observations.
- n : The number of pairs of observations.

Measures:

- 1 : Perfect monotonic positive correlation
- 0 : No monotonic correlation
- -1 : Perfect monotonic negative correlation

A.1.2. Kendall Rank Correlation Coefficient (τ)

Kendall's τ (tau) is a non-parametric measure of rank correlation that assesses the similarity of the ordering of data when ranked by quantities.

$$\tau = \frac{N_c - N_d}{\frac{1}{2}n(n - 1)}$$

Where:

- N_c : The number of concordant pairs (pairs where the ranks for both variables agree in direction).
- N_d : The number of discordant pairs (pairs where the ranks for both variables disagree in direction).
- n : The total number of observations or pairs.

Measures:

- 1 : Perfect agreement (monotonic positive correlation)
- 0 : No agreement (independence)
- -1 : Perfect disagreement (monotonic negative correlation)

Appendix B

B.1. Score Distribution per Dimension

This appendix provides the figures showing the score distribution per dimension for the training and validation dataset splits.

B.1.1. Training Score Distribution

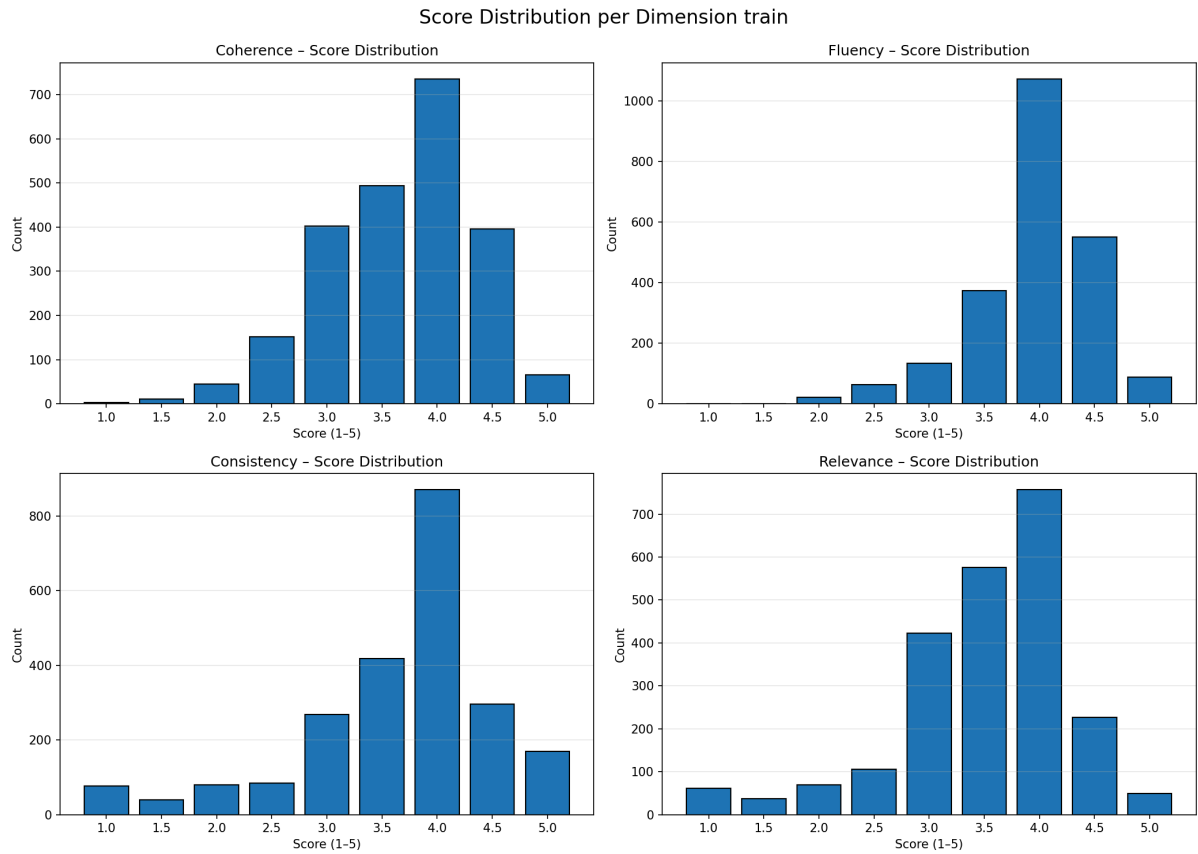


Figure 20: Each histogram displays the frequency of the human-annotated scores (1-5) for each of the quality dimensions.

B.1.2. Validation Score Distribution

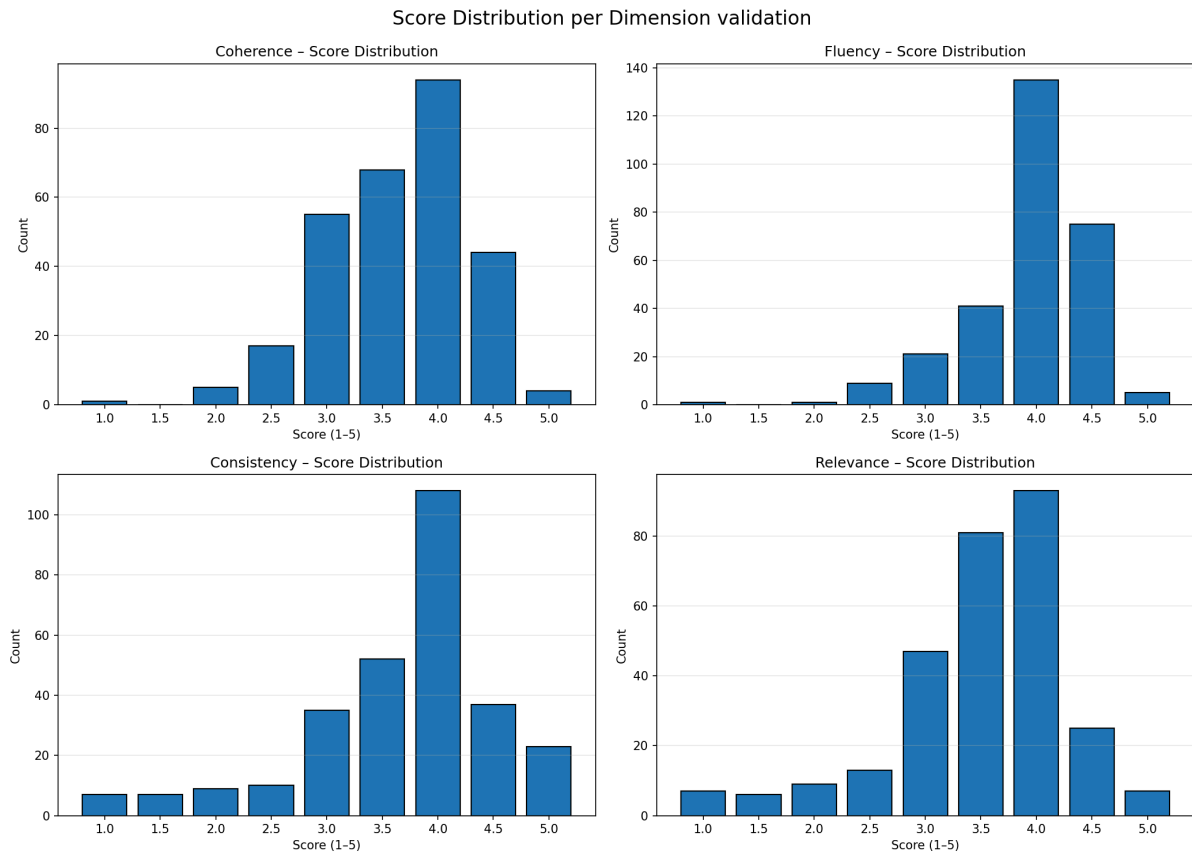


Figure 21: Each histogram displays the frequency of the human-annotated scores (1-5) for each of the quality dimensions.

B.2. Qualitative Reasoning Trace Showcase

This appendix provides examples of summaries with different quality levels. These examples showcase a low-quality, medium-quality, and high-quality summary and their corresponding *overall score*, reflecting an averaged predicted score across dimensions, the source document/context, model-generated summary, and dimension-specific QA pairs for each dimension.

B.2.1. Showcase Example 1: Low-Quality Summary

- Predicted Overall Score: 2.478
- Source:

Everett: Ralph asked me if i could give him your phone number, is that cool?
Amy: who's ralph?
Everett: my friend, i introduced him to you at the pub last week, tall, brown hair, weird laugh...
Amy: oh i remember him now, is he a psycho?
Everett: no
Amy: ok, he can have my number

- Summary:

Everett gave Ralph Amy's phone number.

Consistency:

Question-generated:

Does the summary accurately reflect that Everett gave Ralph Amy's phone number as stated in the context?

Answer-generated:

Yes, the summary accurately reflects that Everett gave Ralph Amy's phone number. The context clearly shows Everett asking Amy if he could give Ralph his phone number, and Amy agreeing, confirming that Everett gave Ralph Amy's phone number. The summary correctly captures this key detail without adding or omitting critical information.

Relevance:

Question-generated:

Does the summary include all the secondary details about Ralph's identity and the context of Everett giving Ralph Amy's phone number?

Answer-generated:

No, the summary does not include all the secondary details about Ralph's identity and the context of Everett giving Ralph Amy's phone number. While it mentions Everett giving Ralph Amy's phone number, it omits important context such as Ralph being Everett's friend, his tall, brown hair, weird laugh, and Amy's confusion about who he is and whether he is a psycho. These details are crucial for understanding the identity and context of the exchange.

Coherence:

Question-generated:

Does the summary present the sequence of events and information about Ralph's phone number in a clear and organized manner?

Answer-generated:

Yes, the summary presents the sequence of events and information about Ralph's phone number in a clear and organized manner. It begins with Everett giving Ralph Amy's phone number, which naturally follows from the conversation where Everett introduces Ralph at the pub and asks if he could give him his phone number. The summary logically follows with Amy's reaction and the agreement to give Ralph Amy's phone number, maintaining coherence throughout.

Fluency:

Question-generated:

Does the summary use clear and grammatically correct language that makes it easy to read and understand?

Answer-generated:

Yes, the summary uses clear and grammatically correct language that makes it easy to read and understand. The sentence is straightforward, concise, and free

of awkward phrasing or grammatical errors. It effectively conveys the main point without unnecessary complexity or complexity, contributing to good fluency.

B.2.2. Showcase Example 2: Mid-Quality Summary

- Predicted Overall Score: 3.638
- Source:

Former Holland forward and assistant coach Patrick Kluivert has made a winning start in World Cup qualifying as coach of Curacao. The Caribbean island team advanced to the second qualifying round in CONCACAF for the 2018 tournament late Tuesday. Curacao drew 2-2 in a second-leg match at Montserrat to win 4-3 on aggregate. Kluivert's team won 2-1 on Saturday in Willemstad. Patrick Kluivert has helped Curacao progress to the next stage of 2018 World Cup qualification. Former Holland assistant Kluivert (far right) worked with Louis van Gaal (centre) at the 2014 World Cup. The first-round encounter was headed to extra time on Tuesday until former Holland Under 21 international Charlton Vicente scored in the 87th minute, five minutes after coming on as a substitute. Curacao face Cuba in the next round, on June 8 and 16. Kluivert, whose mother is from Curacao, was a World Cup losing semi-finalist in 1998 as a player, and last year as assistant to coach Louis van Gaal.

- Summary:

Kluivert's side won 2-1 on Saturday in Willemstad

Consistency:

Question-generated:

Does the summary accurately reflect that Patrick Kluivert's side won 2-1 on Saturday in Willemstad, as stated in the context?

Answer-generated:

Yes, the summary accurately reflects that Patrick Kluivert's side won 2-1 on Saturday in Willemstad. The context confirms that Kluivert's team won 2-1 on Saturday in Willemstad, which aligns with the summary's statement. There are no hallucinated or unsupported details introduced in the summary.

Relevance:

Question-generated:

Does the summary capture the key details about Patrick Kluivert's victory and the significance of the match as described in the context?

Answer-generated:

Yes, the summary captures the key details about Patrick Kluivert's victory and the significance of the match. It mentions that Kluivert's side won 2-1 on Saturday in Willemstad, which aligns with the context mentioning the victory and the significance of the match. However, it omits important specifics such as

the match outcome (2-2 draw, 4-3 aggregate), the coach's background, and the team's progression to the second qualifying round. Including these details would provide a fuller understanding of the match's significance.

Coherence:

Question-generated:

Does the summary present the information in a clear and logically connected sequence that makes it easy to understand the sequence of events?

Answer-generated:

Yes, the summary presents the information in a clear and logically connected sequence. It begins by introducing Patrick Kluivert's winning start in World Cup qualifying, then explains the match outcome (2-1 win in Willemstad), followed by the team's progress to the second qualifying round, and concludes with the significance of the victory. This progression makes it easy to understand the sequence of events.

Fluency:

Question-generated:

Does the summary read smoothly and naturally without awkward phrasing or grammatical errors?

Answer-generated:

Yes, the summary reads smoothly and naturally without awkward phrasing or grammatical errors. The sentences are clear, concise, and well-structured, making it easy to understand. The use of punctuation and conjunctions (e.g., "*Kulivert's side won 2-1 on Saturday in Willemstad*") contributes to natural flow and readability. Overall, it is fluent and well-written.

B.2.3. Showcase Example 3: High-Quality Summary

- Predicted Overall Score: 4.591
- Source:

Peter: Did you hear what happened at school today?

Jack: No?

Peter: There's been kind of an emergency, gas leak or something.

Jack: Oh, I hope it's nothing serious. Did they called fire units?

Peter: Yep, three engines and ambulance. It turned out to be a drill.

Jack: It's better this way than the other.

- Summary:

There was an emergency at school today. Three engines and ambulance came, but it was a drill.

Consistency:**Question-generated:**

Does the summary accurately reflect that there was an emergency at school and that three engines and ambulance came, as stated in the original conversation?

Answer-generated:

Yes, the summary accurately reflects the original conversation. It correctly states that there was an emergency at school, that three engines and ambulance came, and that it was a drill. The summary aligns with the dialogue without adding or omitting key details.

Relevance:**Question-generated:**

Does the summary include all the secondary details about the emergency, ambulance response, and drill that are present in the conversation?

Answer-generated:

Yes, the summary includes all the secondary details about the emergency, ambulance response, and drill. It correctly states that three engines and ambulance came and that it was a drill, which aligns with the conversation where Peter confirms the emergency was a gas leak and fire units called fire units. The summary captures the essential points without omitting any critical information.

Coherence:**Question-generated:**

Does the summary present the events in a clear and logically connected sequence that makes it easy to understand the sequence of events?

Answer-generated:

Yes, the summary presents the events in a clear and logically connected sequence. It begins by stating the emergency at school, then explains the presence of three engines and ambulance, followed by the fact that the drill was a drill. This sequence flows naturally from the emergency to the ambulance arrival, making it easy to understand the sequence of events.

Fluency:**Question-generated:**

Does the summary use clear and grammatically correct language throughout?

Answer-generated:

Yes, the summary uses clear and grammatically correct language throughout. The sentences are straightforward and easy to understand, with proper punctuation and no awkward phrasing. The summary effectively conveys the information without grammatical errors, contributing to good readability and fluency.

Appendix C

C.1. Tables of Correlation Levels RQ2

This appendix provides the tables of sample-level, summary-level, and system-level correlation for RQ2. Each table provides an overview of the Spearman and Kendall correlation for each quality dimension as well as the average overall correlation.

C.1.1. Sample-Level Correlation

Metrics	Coherence		Fluency		Consistency		Relevance		Average	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Similarity-based Metrics										
ROUGE-1	0.379	0.269	0.300	0.209	0.166	0.108	0.182	0.125	0.355	0.243
ROUGE-2	0.278	0.198	0.259	0.183	0.190	0.130	0.177	0.126	0.334	0.230
ROUGE-L	0.385	0.274	0.325	0.226	0.100	0.069	0.096	0.067	0.305	0.208
Multi-dimensional Evaluators										
UniEval (Continual)	0.184	0.127	0.319	0.223	0.077	0.056	0.245	0.166	0.344	0.234
UniEval(Multi-Task)	0.183	0.129	0.319	0.221	0.049	0.036	0.233	0.157	0.364	0.247
QAG-Eval (Continual)	0.615	0.454	0.533	0.387	0.581	0.434	0.576	0.427	0.548	0.392
QAG-Eval (Multi-Task)	0.664	0.497	0.499	0.360	0.606	0.449	0.612	0.460	0.598	0.435

Table 14: Sample-level Spearman (ρ) and Kendall-Tau (τ) correlations of different metrics on the held-out test dataset. Bold numbers indicate the best performing metric of the its corresponding dimension column.

C.1.2. Summary-Level Correlation

Metrics	Coherence		Fluency		Consistency		Relevance		Average	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Similarity-based Metrics										
ROUGE-1	0.390	0.376	0.179	0.182	0.044	0.040	0.133	0.140	0.226	0.229
ROUGE-2	0.311	0.307	0.167	0.168	0.008	0.005	0.046	0.048	0.139	0.126
ROUGE-L	0.511	0.493	0.288	0.288	0.132	0.136	0.165	0.164	0.260	0.255
Multi-dimensional Evaluators										
UniEval (Continual)	0.201	0.180	0.148	0.133	0.211	0.216	0.238	0.233	0.160	0.165
UniEval(Multi-Task)	0.121	0.109	0.158	0.145	0.244	0.248	0.184	0.182	0.213	0.206
QAG-Eval (Continual)	0.540	0.522	0.305	0.292	0.283	0.260	0.419	0.412	0.318	0.307
QAG-Eval (Multi-Task)	0.469	0.451	0.286	0.270	0.220	0.205	0.420	0.402	0.303	0.284

Table 15: Summary-level Spearman (ρ) and Kendall-Tau (τ) correlations of different metrics on the held-out test dataset. Bold numbers indicate the best performing metric of the its corresponding dimension column.

C.1.3. System-Level Correlation

Metrics	Coherence		Fluency		Consistency		Relevance		Average	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Similarity-based Metrics										
ROUGE-1	0.647	0.500	0.752	0.552	-0.099	-0.085	0.018	0.000	0.369	0.274
ROUGE-2	0.492	0.380	0.596	0.442	0.054	0.044	0.132	0.092	0.411	0.301
ROUGE-L	0.660	0.514	0.764	0.557	-0.116	-0.071	-0.016	-0.014	0.382	0.278
Multi-dimensional Evaluators										
UniEval (Continual)	0.451	0.315	0.501	0.364	0.099	0.060	0.244	0.168	0.382	0.297
UniEval(Multi-Task)	0.405	0.274	0.611	0.480	-0.049	-0.018	0.240	0.164	0.451	0.310
QAG-Eval (Continual)	0.773	0.560	0.815	0.608	0.814	0.628	0.825	0.635	0.715	0.549
QAG-Eval (Multi-Task)	0.799	0.615	0.846	0.658	0.840	0.669	0.801	0.621	0.733	0.572

Table 16: System-level Spearman (ρ) and Kendall-Tau (τ) correlations of different metrics on the held-out test dataset. Bold numbers indicate the best performing metric of the its corresponding dimension column.