Mälardalen University
Department of Computer Science and Engineering
Västerås, Sweden

Degree of Master of Science in Computer Science - 30 credits

EVALUATING GROUNDING STRATEGIES FOR FACTUAL ACCURACY AND INFORMATION COVERAGE IN LLM-GENERATED REVIEW SUMMARIES

Abdul Wahab
awb25001@student.mdu.se

Examiner: Mobyen Uddin Ahmed
Mälardalen University, Västerås, Sweden

Supervisor: Shahina Begum
Mälardalen University, Västerås, Sweden

Company supervisor: Sasa Moilanen
Zoined Oy, Helsinki, Finland

August 24, 2026

This thesis was conducted within the Erasmus Mundus Joint Master Degree Programme on the Engineering of Data-intensive Intelligent Software Systems (EDISS).

Acknowledgments

This thesis has been a challenging and rewarding journey, and I am thankful to everyone who supported me along the way. I would first like to express my gratitude to Zoined Oy and its CEO, Sasa Moilanen, for giving me the opportunity to work on this thesis topic. I am thankful to Tor for his guidance throughout the work. His technical knowledge and expertise in AI were invaluable throughout the thesis work. I am also grateful to Karipekka Kaunisto and the rest of the team for their support and for providing the resources necessary to complete this work.

I am also grateful to my supervisor, Shahina Begum, and my examiner, Mobyen Uddin Ahmed, for their guidance, regular communication, and valuable feedback throughout the thesis process. I am also grateful to the EDISS programme for providing me with the scholarship and the opportunity to pursue my studies, allowing me to fulfill a long-standing goal of studying abroad.

Finally, I would like to thank my friends and family for their continuous support. I am especially grateful to my brother, Abdul Basit, for his invaluable support throughout my studies and for his encouragement throughout this journey.

Abdul Wahab
August 2026

Abstract

Large Language Models (LLMs) are increasingly used to summarize business data such as customer reviews. Although these models produce fluent outputs, they can generate statements that are not fully supported by the input data, a problem referred to as hallucination. In business analytics, where summaries may directly influence operational decisions, this creates a significant reliability risk. This thesis investigates how different grounding strategies affect the quality of LLM-generated restaurant review summaries. Grounding constrains generation to rely on explicitly provided external data rather than on knowledge learned during training. The research question is: How do different grounding strategies affect the quality and reliability of LLM-generated review summaries?

Four strategies are evaluated in a single-factor controlled experiment: Plain Prompting, Retrieval-Augmented Generation (RAG), Aggregated-Statistics Grounding, and Template-Constrained Generation. Each is applied to eleven evaluation scenarios built from real restaurant review data and supplemented with synthetic cases covering text-rating conflicts, ambiguous opinions, and high information density. Generated summaries are assessed using atomic-fact verification, golden-insight coverage scoring, citation-quality metrics, and operational telemetry, with an LLM-as-a-Judge pipeline as the primary evaluator.

The results reveal clear trade-offs. RAG is the most computationally efficient but underperforms in information coverage. Aggregated-Statistics Grounding and Template-Constrained Generation achieve the highest mean factual accuracy, but their advantage over the Plain Prompting long-context baseline is not statistically significant. Template-Constrained Generation shows the weakest attribution quality. No single strategy is superior across factual accuracy, coverage, attribution quality, and operational cost, so the findings provide practical guidance for selecting a strategy based on deployment requirements.

Keywords: Large Language Models, Text Summarization, Grounding Strategies, Factual Accuracy, Information Coverage, Attribution Faithfulness, Hallucination, Retrieval-Augmented Generation (RAG), Hybrid Data, LLM-as-a-Judge.

Table of Contents

1 Introduction **1**

2 Background **3**
2.1 Key Concepts and Terminology . . . 3
2.2 Automated Evaluation: LLM-as-a-Judge . . . 5
2.3 Evaluation Metrics . . . 5
2.3.1 Factual Accuracy . . . 5
2.3.2 Information Coverage . . . 5
2.3.3 Attribution Faithfulness . . . 6
2.3.4 Operational Cost . . . 7

3 Related Work **8**
3.1 Hallucinations and the Need for Grounding . . . 8
3.2 Grounding Architectures: Long-Context Processing and Retrieval . . . 9
3.3 Hybrid Data, Numeracy, and Constrained Generation . . . 9
3.4 Attribution Faithfulness and Grounded Trustworthiness . . . 10
3.5 Evaluation Frameworks for Factuality, Coverage, and Cost . . . 10
3.6 Research Gaps and Positioning of This Thesis . . . 12

4 Problem Formulation **15**
4.1 Problem Definition . . . 15
4.2 Purpose and Research Questions . . . 16
4.3 Motivation . . . 16
4.4 Goals . . . 17
4.5 Scope and Limitations . . . 18

5 Method **19**
5.1 Research Design . . . 19
5.2 Data and Scenario-Based Evaluation Design . . . 20
5.3 Selection of Grounding Strategies . . . 21
5.4 Evaluation Methodology . . . 24
5.5 Statistical Analysis . . . 25
5.6 Threats to Validity . . . 25

6 Ethical and Societal Considerations **27**

7 Experimental Setup and Implementation **28**
7.1 Dataset and Hybrid Data Format . . . 28
7.2 Base Task Instruction and What It Tests . . . 28
7.3 Scenario Design and Construction . . . 30
7.4 Golden Insight Design . . . 32
7.5 Model Selection and Configuration . . . 33
7.6 Prompt Construction Per Strategy . . . 34
7.7 Automated Evaluation Framework . . . 38
7.8 Manual Audit Protocol . . . 41

8 Results **42**
8.1 Overview of Results . . . 42
8.2 Statistical Testing Protocol . . . 43
8.3 RQ1: Factual Accuracy and Information Coverage . . . 44
8.3.1 Factual Accuracy . . . 44
8.3.2 Information Coverage . . . 46
8.4 RQ2: Attribution Faithfulness . . . 48
8.4.1 Citation Precision . . . 48
8.4.2 Citation Recall . . . 49
8.4.3 Fully Supported Claims Ratio . . . 49
8.5 RQ3: Inference Latency and Token Volume . . . 51
8.5.1 Inference Latency . . . 51
8.5.2 Token Volume . . . 53
8.6 Manual Audit of Judge Verdicts . . . 53
8.7 Summary of Main Findings . . . 55

9 Discussion **58**
9.1 Interpretation of Results by Research Question . . . 58
9.2 Relation to Previous Work . . . 59
9.3 Practical Recommendations . . . 60
9.4 Limitations and Future Work . . . 60

10 Conclusion **62**

References **64**

Appendix A Scenario Descriptive Statistics **67**

B Scenario 11 (Text-Rating Conflict): Full Input Data **69**
B.1 Aggregated Statistics . . . 69
B.2 Review Records . . . 70
B.3 Golden Insights . . . 70

C Scenario 11 (Text-Rating Conflict): Generation and Evaluation **71**
C.1 Generated Summaries (Run 1) . . . 72
C.2 Coverage Results by Golden Insight . . . 74
C.3 Attribution Results . . . 75
C.4 Summary . . . 76

D Evaluation Prompts and Supporting Schemas **76**
D.1 Judge System Prompt . . . 77
D.2 Atomic Fact Decomposition Prompt . . . 77
D.3 Atomic Fact Verification Prompt . . . 79
D.4 Information Coverage Evaluation Prompt . . . 80
D.5 Attribution Faithfulness Evaluation Prompt . . . 81
D.6 Template-Constrained Output Schema . . . 82
D.7 RAG Keyword Extraction Prompt . . . 83
D.8 Shared Generation Contract Blocks . . . 84

E Per-Scenario Result Heatmaps **84**

F Use of AI Tools **88**

1 Introduction

Large Language Models (LLMs) have become widely used for automating text-based tasks such as summarization, question answering, and report generation. In business analytics, especially in the restaurant domain, these models are increasingly used to summarize customer reviews, enabling businesses to extract insights from hybrid data consisting of unstructured review text and structured numerical signals such as ratings. Such summaries can support decision-making by highlighting common customer opinions, strengths, and weaknesses.

Despite their strong language generation capabilities, LLMs may produce outputs that are not fully supported by the input data. This issue, often referred to as hallucination, can lead to misleading or incorrect summaries. In business contexts, where summaries can influence operational or strategic decisions, this creates a significant risk. Errors such as misreporting ratings or customer sentiment can lead to incorrect conclusions, while the lack of traceability makes it difficult to verify generated claims. Ensuring that summaries are both factually correct and supported by provided evidence is therefore essential.

One approach to mitigating hallucinations is "grounding": guiding generation with explicitly provided external evidence [1, 2] rather than relying solely on the model's implicit parametric knowledge from training [3]. This thesis evaluates four grounding strategies that intervene at different points of the pipeline: full-context input (Plain Prompting), filtered input (RAG), augmented input (Aggregated-Statistics Grounding), and constrained output (Template-Constrained Generation). Sections 5 and 7 describe each strategy in detail.

However, there are two major gaps. First, prominent evaluation frameworks heavily emphasize precision and attribution, assessing whether generated statements are correct and properly supported by sources [4, 5]. This often comes at the expense of information coverage. For example, foundational metrics like FActScore explicitly neglect recall, which means a model can achieve a high score by omitting important insights [4]. This is particularly problematic in review summarization, where models tend to miss relevant information as input size increases [6].

Second, prior work does not adequately address hybrid data, where unstructured text must be reconciled with structured numerical signals. LLMs have been shown to struggle with numerical reasoning, often leading to inconsistencies in quantitative information [7]. While data-to-text generation research considers structured inputs [8], there is a lack of comparative studies that evaluate grounding strategies under such hybrid conditions.

Standard evaluation settings also often rely on relatively consistent benchmark datasets, such as Wikipedia, with well-defined reference material [5, 4], whereas customer review data is inherently subjective, heterogeneous, and often contradictory. This creates additional challenges for factual evaluation, where multiple conflicting perspectives must be preserved without introducing unsupported conclusions.

This thesis investigates different ways of grounding LLMs using external business data, with a focus on customer review summarization. The main objective is to evaluate how these grounding approaches influence key quality dimensions, including factual accuracy, information coverage, attribution faithfulness, and operational cost. The central research question is: *How do different grounding strategies affect the quality and reliability of LLM-generated review summaries?*

This problem is particularly relevant in real-world business analytics, where inaccurate or incomplete summaries can directly impact decision-making and reduce trust in automated insights. To address this, the work here adopts a comparative experimental approach, evaluating four grounding strategies across eleven evaluation scenarios derived from real restaurant review data and supplemented with synthetically constructed cases that cover conditions such as text-rating conflicts, ambiguous opinions, and high information density. A controlled experiment is appropriate because it isolates the effect of the grounding strategy from other variables, such as model choice or data characteristics. Each strategy is evaluated in the same set of scenarios using the same generation model, task instruction, and evaluation pipeline. This makes the results directly comparable across strategies and allows observed differences to be attributed to the grounding strategy rather than to other parts of the setup.

The results show that grounding strategies lead to clear trade-offs. RAG is the most computationally efficient but substantially underperforms in information coverage. Aggregated-Statistics Grounding and Template-Constrained Generation reach the highest mean factual accuracy, but their advantage over the Plain Prompting long-context baseline is not statistically significant, and Template-Constrained Generation shows the weakest attribution quality. No single strategy dominates across all evaluation dimensions, and the most appropriate choice depends on the deployment requirements of the application. These findings are based on a specific restaurant-domain customer review dataset and evaluation setup, and may not transfer directly to other domains or data distributions.

The main contributions of this work are threefold. First, it provides a controlled head-to-head comparison of four grounding strategies on hybrid business review data. Second, it shows that enforcing a rigid output schema does not improve citation quality and can lead to over-citation, where the model attaches loosely relevant sources to satisfy the format rather than to provide genuine evidential support. Third, it introduces a multi-dimensional evaluation framework combining atomic-fact verification, golden-insight coverage recall, and citation-level attribution quality, applied here to subjective review data rather than to the cleaner factual benchmarks commonly used in prior work. Together, these contributions provide practical guidance for designing more reliable summarization systems and contribute to a better understanding of how input conditioning influences factual text generation.

This thesis is intended for AI practitioners, data scientists, and business analytics providers such as Zoined Oy, a business analytics platform that aggregates customer feedback for restaurant and retail clients. It is written for readers interested in deploying reliable LLM-based summarization systems in settings where factual accuracy, information coverage, and computational cost matter.

The remainder of the report is organized as follows. Sections 2–3 present the background and related work, followed by the problem formulation in Section 4 and the method in Section 5. Ethical and societal considerations are discussed in Section 6, and the experimental setup in Section 7. The results are presented in Section 8, discussed in Section 9, and concluded in Section 10.

2 Background

The task studied in this thesis is a form of multi-document summarization. The system receives a set of review records for a single restaurant, where each record consists of an optional free-text comment, an overall star rating, optional aspect-level ratings, and a timestamp, and the system must produce a single natural-language summary of that set.

A further property of this task is that review data is inherently heterogeneous. It combines unstructured text, such as free-form customer comments, with structured numerical signals, such as star ratings and aspect-level scores for food, service, and atmosphere. Generating reliable summaries from this kind of hybrid data requires more than fluent language generation. The model must also handle quantitative information correctly, preserve factual accuracy, and produce outputs that are transparent enough to be verified against the source.

This section introduces the core concepts and evaluation metrics used throughout the thesis. Readers already familiar with LLM-based summarization and grounding may use it as a terminology reference.

2.1 Key Concepts and Terminology

- **Large Language Models (LLMs).** Neural language models trained on large text corpora to generate fluent natural language. While they perform well on tasks such as summarization, they may also produce statements that are not supported by the provided input data, especially when handling long or heterogeneous inputs.
- **Hallucination.** The generation of content that is factually incorrect, unsupported, or unverifiable with respect to the given source data. In review summarization, this may include inventing trends, misrepresenting ratings, or exaggerating customer sentiment.
- **Grounding.** The process of constraining a language model to generate outputs based on explicitly provided input data rather than relying only on implicit knowledge learned during training. In this thesis, grounding is operationalized as different ways of selecting, structuring, or augmenting the input data, as well as constraining the output format used by the model.
- **Grounding Strategies.** Concrete mechanisms used to guide model generation. This thesis evaluates four strategies, each targeting a different aspect of the grounding problem:
  - **Plain Prompting (Long-Context Baseline).** The model receives all available reviews as raw text without retrieval or output constraints. It serves as the baseline, showing how far the model can go using its native long-context ability alone.
  - **Retrieval-Augmented Generation (RAG).** A subset of the most relevant reviews is selected using a retrieval step and passed to the model instead of the full set. This reduces token usage and noise, but may also omit relevant evidence that was not retrieved.

  - **Aggregated-Statistics Grounding.** Pre-computed numerical summaries, such as average ratings and rating distributions, are injected into the prompt alongside the review text. This makes quantitative signals explicit rather than leaving the model to infer them from raw text, reducing the risk of numerical errors in the output.
  - **Template-Constrained Generation.** The model is required to produce its output according to a predefined structured schema. The schema fields correspond directly to the evaluation themes in the shared task instruction, with an additional open-ended field for any insights that do not fall under those categories. This design ensures that the output structure reflects the analytical goals of the task rather than imposing an unrelated format on the model.

- **Review Summarization.** The task of generating a concise, informative summary from a collection of customer reviews.
- **Hybrid Review Data.** Input data that combines unstructured text, such as free-form customer comments, with structured numerical signals, such as overall star ratings and aspect-level scores for food, service, and atmosphere. For example, a dataset entry may contain the written comment "slow service and cold food" together with a 2-star overall rating and sub-ratings for food, service, and atmosphere.
- **Factual Accuracy.** The extent to which statements in a generated summary are supported by the input data. Measured at the level of individual atomic claims, following the FActScore paradigm [4]. Defined formally in Section 2.3.
- **Information Coverage.** The degree to which a summary captures the key insights present in the source data. A summary can be factually precise while still omitting important information, so coverage is measured separately from accuracy. Defined formally in Section 2.3.
- **Attribution Faithfulness.** The extent to which generated claims can be traced back to specific input sources, such as individual reviews or numerical values. In this thesis, attribution faithfulness is used as a practical way to evaluate explainability and is assessed through three complementary citation-based metrics defined in Section 2.3.
- **Operational Cost.** The computational resources consumed when generating a summary using a given grounding strategy. In this thesis, it is measured along two dimensions: inference latency, which is the wall-clock time required to produce a summary, and token volume, which is the total number of tokens consumed by the prompt and the generated output combined.
- **Golden Insights.** Manually written, scenario-specific statements that capture the key facts and patterns present in the source data. They serve as the reference against which information coverage is measured: a good summary should express at least the meaning of each golden insight, even if the wording differs.

2.2 Automated Evaluation: LLM-as-a-Judge

Evaluating a large number of generated summaries manually is difficult to scale consistently. This thesis therefore uses an *LLM-as-a-Judge* framework, where a strong instruction-tuned language model acts as an automated evaluator [9]. The judge is given fixed evaluation prompts and checks each summary against the source scenario data and golden insights using structured criteria. This allows multiple quality dimensions to be assessed in a systematic and repeatable way across all scenarios and strategies. The judge model, its configuration, and evaluation prompts are described in Section 7.

2.3 Evaluation Metrics

Generated summaries are evaluated across four dimensions: factual accuracy, information coverage, attribution faithfulness, and operational cost. These metrics are designed to go beyond surface fluency and measure whether a summary is correct, sufficiently complete, traceable, and practical to deploy.

2.3.1 Factual Accuracy

Each generated summary is first decomposed into a list of independent atomic claims, where each claim expresses exactly one verifiable piece of information. This follows the FActScore evaluation paradigm [4], which evaluates factual precision at the level of individual claims rather than whole paragraphs. The decomposition step also separates non-verifiable output, such as conversational filler phrases like "Here is a summary of the reviews", which are excluded from evaluation since they make no factual claim about the restaurant. Each remaining claim is then verified against the complete source scenario data by the judge model, including raw reviews and structured numerical fields. This keeps generation and evaluation separate: strategies may receive different inputs, but all outputs are judged against the same complete business-data reference.

Let A denote the set of verifiable atomic claims extracted from a summary, and let $A_{\text{supported}}$ denote the subset that is supported by the source data. Factual accuracy is defined as:

$$\text{FActScore} = \frac{|A_{\text{supported}}|}{|A|}$$

The denominator $|A|$ therefore counts only verifiable claims. A high FActScore means that most of what the model asserted about the restaurant can be traced back to the source data. A low score indicates that the summary contains unsupported or fabricated claims.

2.3.2 Information Coverage

Factual accuracy alone does not reveal whether important information was left out. A summary can achieve a high FActScore by making only a small number of safe claims while omitting critical insights entirely. Coverage captures this recall dimension.

For each scenario, a set of golden insights G is manually written. Each insight expresses one key fact or pattern that a good summary should mention. Let G_{covered}

denote the subset of golden insights whose core meaning is represented in the generated summary. Coverage is defined as:

$$\text{Coverage} = \frac{|G_{\text{covered}}|}{|G|}$$

The judge checks each golden insight independently and determines whether its meaning is present in the summary, without requiring exact wording.

2.3.3 Attribution Faithfulness

Attribution faithfulness measures whether the evidence cited for a generated claim actually supports that claim. In this thesis, most strategies cite review identifiers as evidence. The Aggregated-Statistics strategy can also cite statistical fields, because it is the only strategy that includes an explicit statistics block in the prompt.

For example, an Aggregated-Statistics summary may contain the following cited claim:

> "Customers generally rate this restaurant positively, with an average rating of 4.1/5 and a median of 4.0/5 across 50 reviews [STAT: average_rating, STAT: median_rating, STAT: total_reviews]."

Here, the sentence is the claim, and the statistical fields in brackets are the cited evidence. For the other strategies, citations normally refer to review IDs, such as:

> "Customers repeatedly praise the food, especially baked potatoes [REVIEW: 3, REVIEW: 8, REVIEW: 14, REVIEW: 15, REVIEW: 19]."

A summary can be factually correct while still citing weak or irrelevant evidence. For example, a claim may be true according to the full review set, but the specific review IDs attached to that claim may not actually support it. This makes attribution faithfulness different from factual accuracy: factual accuracy checks whether the claim is correct, while attribution faithfulness checks whether the cited evidence justifies that claim.

This distinction matters because a summary may appear transparent while still pointing to the wrong evidence. Three complementary metrics are used: Citation Precision (CP) and Citation Recall (CR), defined following the citation evaluation framework of [5], and a Fully Supported Claims Ratio (FSCR) that this thesis introduces as a stricter, claim-level summary of attribution quality.

Let R denote the full set of claims in a summary. Here, a claim is a generated statement in the summary output together with any citations attached to it. These statements are the unit of attribution evaluation and are distinct from the finer-grained atomic facts used for factual accuracy. For each claim, the model is expected to cite one or more review identifiers, or statistical fields (in the Aggregated-Statistics strategy), as supporting evidence.

Citation Recall measures the fraction of all claims that have at least one valid cited source. A claim with no citation, or whose only citations are invalid, counts as a failure:

$$\text{Citation Recall} = \frac{|\{r \in R : \text{at least one cited source for } r \text{ is valid}\}|}{|R|}$$

Citation Precision measures, across all citation acts in the summary, what fraction points to a source that genuinely supports the associated claim:

$$\text{Citation Precision} = \frac{\text{total valid cited sources}}{\text{total cited sources}}$$

Fully Supported Claims Ratio is the strictest of the three. It measures the fraction of claims for which every cited source is valid. A single invalid citation among several causes the entire claim to fail:

$$\text{Fully Supported Claims Ratio} = \frac{|\{r \in R : \substack{r \text{ has at least one cited source and} \\ \text{all cited sources for } r \text{ are valid}}\}|}{|R|}$$

Together, these three metrics separate citation quantity from citation quality. Citation Recall checks whether claims are backed at all, Citation Precision checks whether individual citations are appropriate, and the Fully Supported Claims Ratio checks whether claims are fully and correctly evidenced. In the hybrid review summarization setting studied here, these three metrics identify distinct ways a generated summary can mislead business users, each reflecting a different breakdown in citation reliability.

2.3.4 Operational Cost

Operational cost is measured using two system-level metrics logged for every run.

Inference latency is the total wall-clock time in seconds required to generate a summary for a given scenario. Token volume is the total number of tokens consumed per run, defined as:

$$\text{Tokens}_{\text{total}} = \text{Tokens}_{\text{prompt}} + \text{Tokens}_{\text{completion}}$$

These two measures together reflect the practical deployment cost of each grounding strategy. Latency affects real-time usability, while token volume is a main driver of API cost in production systems. Taken together, the concepts and metrics defined in this section form the evaluation foundation of the thesis. Factual accuracy and information coverage address the two sides of output quality: precision and recall. Attribution faithfulness operationalizes explainability. Operational cost situates each strategy within practical deployment constraints. The following sections describe how these dimensions are measured and compared across the four grounding strategies.

3 Related Work

This section positions the thesis within existing research on hallucination, grounding architectures, hybrid data handling, attribution faithfulness, and automated evaluation. The literature shows considerable progress in each of these areas individually, but reveals important gaps when they are considered together. In particular, most prior work targets pure-text benchmarks, whereas this thesis studies grounded summarization of hybrid restaurant review data that combines subjective review text with structured numerical signals such as star ratings and aspect scores.

3.1 Hallucinations and the Need for Grounding

A foundational challenge for LLM-based generation is hallucination: the production of fluent but factually unsupported or unverifiable content. The literature distinguishes two classical types [1]: *intrinsic hallucination*, where the output directly contradicts the source material, and *extrinsic hallucination*, where the output cannot be verified against the source, neither supported nor contradicted. Huang et al. [10] refine this taxonomy for the LLM era, distinguishing *factuality hallucination* (discrepancy with verifiable world facts) from *faithfulness hallucination* (divergence from the provided context or user instruction). In this thesis, the relevant failure mode is faithfulness hallucination: the evaluation asks whether generated claims are consistent with and supported by the provided reviews and statistics, not whether they are globally true as world facts. Over-reliance on parametric knowledge is identified as one important cause of this failure, as models may draw on training data rather than the input context [1, 10].

In summarization settings, faithfulness hallucination is particularly problematic because a generated summary may appear fluent and coherent while severely misrepresenting the source text [1, 10]. Importantly, unreliability in abstractive summarization is not limited to explicitly invented or contradictory content. A model can also exhibit "hallucination by omission", producing a factually safe but incomplete summary that silently drops important source signals to maintain high precision [4, 6]. Both failure modes pose severe risks in downstream business analytics, where automated summaries can directly influence operational or strategic decisions [10, 11]. As the volume of input documents scales, the recall of multi-document summarizers has been shown to decay significantly [6]. An incorrect claim about ratings or recurring complaints can mislead decision-making, but so can a summary that fails to surface a critical complaint buried among many positive reviews.

Grounding has emerged as one well-established approach to addressing this problem. Rather than relying only on knowledge encoded during pre-training, grounding constrains generation to explicitly provided external evidence [2, 3]. Song et al. [12] reinforce this point by showing that even with retrieved documents present, current LLMs often draw on parametric knowledge rather than the provided context, and that prompting-based approaches alone are insufficient to reliably suppress this tendency. This finding motivates the present thesis, which evaluates four different ways of grounding model generation, from direct context provision to structural output constraints.

3.2 Grounding Architectures: Long-Context Processing and Retrieval

A large part of the recent literature compares two dominant grounding paradigms: long-context (LC) processing and retrieval-augmented generation (RAG). Long-context approaches provide the model with the full available input, relying on expanded context windows to allow direct identification of relevant information. RAG instead selects a relevant subset and injects only that into the prompt [2].

Two complementary benchmarking studies characterize the trade-off between these approaches. Li et al. [13] conduct a systematic comparison across multiple LLMs and find that when the model is given sufficient resources, LC consistently outperforms RAG in average performance on question answering, while RAG reduces token consumption to approximately 17% of LC cost, a substantial operational advantage. A separate study by Li et al. [14] extends this analysis and shows that the advantage of LC is most pronounced for well-structured, dense contexts such as Wikipedia articles, while RAG performs better for fragmented information such as dialogue-based queries.

However, the trade-off has an important recall dimension that these studies underemphasize. Belem et al. [6] investigate hallucination specifically in multi-document summarization and find that up to 75% of generated content can be hallucinated in conversation-based settings. More strikingly, GPT-3.5-Turbo generates summaries 79.35% of the time even when no relevant topic-specific information exists in the input documents, which indicates a strong tendency to fabricate rather than acknowledge missing evidence. Their recall analysis also shows that as the number of input documents grows, recall behavior varies across models, with several showing meaningful coverage losses at higher document counts. This finding is directly relevant to this thesis: if even a long-context baseline struggles to retain coverage across many documents, a retrieval-based approach that sees only a small selected subset is likely to face an even sharper recall penalty.

3.3 Hybrid Data, Numeracy, and Constrained Generation

A distinct but related problem is the handling of hybrid data, where unstructured text must be interpreted alongside structured numerical signals. Standard Transformer-based summarization models have been shown to represent quantitative values poorly. White [7] applies probing experiments to several state-of-the-art summarization encoders and finds that most struggle to produce embeddings that adequately represent numerical inputs compared to simple baselines. One model, DistilBART-CDM, underperforms even randomly initialized representations on several quantitative probing tasks. White argues that this encoder weakness contributes directly to quantity hallucinations, where models generate numerical values that are not present in or consistent with the source text.

Related work in data-to-text generation addresses structured inputs such as tables and database records, but typically treats the structured data as the sole source and does not study how it should be combined with variable, conflicting human-written text [8]. This leaves a gap between pure-text summarization research and structured generation research when the task requires joint reasoning over both modalities simultaneously, which is exactly the setting of this thesis.

A separate line of work studies constrained generation, where the model is required to follow lexical constraints [15] or predefined output schemas [16]. Mao et al. [15]

show that lexical constraints (requiring specific source-derived tokens to appear in the summary) can improve factual consistency in abstractive summarization by reducing the model's tendency to generate content that deviates from source evidence. Geng et al. [16] evaluate constrained decoding frameworks more broadly and find that constraint enforcement can improve downstream task accuracy by up to 4% and, in some configurations, speed up generation by around 50% compared to unconstrained decoding. However, neither study examines how such constraints behave when the output must be attributed back to a large set of short, user-generated source documents, nor whether enforcing a rigid output schema helps or hurts citation quality. These questions are central to this thesis.

3.4 Attribution Faithfulness and Grounded Trustworthiness

Beyond factual correctness, an increasingly important line of research asks whether generated claims can be traced back to the source material that supports them. This matters because a summary may sound plausible while still failing to link its claims to valid evidence. Gao et al. [5] introduce the ALCE benchmark for evaluating citation quality in long-form generation. Their results are sobering: on the ELI5 dataset, around 50% of ChatGPT and GPT-4 outputs are not fully supported by the cited passages. ALCE also shows that citation recall evaluation achieves 85.1% accuracy against human judgments, validating automated citation checking as a reliable evaluation method.

Song et al. [12] build on this line of work and show that prompting approaches are insufficient to reliably ground model outputs to provided documents. Their TRUST-SCORE metric evaluates three dimensions jointly: whether the model answers correctly, whether it appropriately refuses when evidence is missing, and whether its citations are valid. They find that models frequently produce confident outputs even when the retrieved documents do not support the answer.

These findings raise a question that prior work has not fully addressed: does the way evidence is presented to the model, through retrieval, aggregated statistics, or output templates, systematically change citation behavior? Prior attribution studies have focused almost entirely on pure-text question answering settings with Wikipedia-style evidence. This thesis extends that perspective to a hybrid, multi-review summarization setting and compares attribution behavior across four architecturally different grounding pipelines.

3.5 Evaluation Frameworks for Factuality, Coverage, and Cost

Traditional overlap-based metrics such as ROUGE and BLEU are known to correlate poorly with factual correctness in open-ended generation [17, 11]. More recent work has therefore moved toward claim-level evaluation. Min et al. [4] introduce FActScore, which decomposes a generated text into independent atomic facts and measures the proportion supported by a reliable knowledge source. Their human evaluation of commercial LLMs reveals that factual precision is substantially lower than fluency would suggest: ChatGPT achieves only 58% on biography generation, and scores drop sharply for rarer entities (from around 80% to 16% as entity rarity increases). FActScore provides a much stronger basis for measuring factual precision than surface-level text overlap, and this thesis adapts its atomic-fact decomposition and source-grounded verification approach to the review summarization domain.

However, the authors of FActScore explicitly acknowledge that the framework was developed and validated on biographies using Wikipedia as the knowledge source, a domain where facts are objective and specific rather than subjective and vague [4]. They note that evaluating the generalizability of FActScore to other domains remains future work. Their proof-of-concept transfer experiment to NLP papers is encouraging, but it is explicitly described as too small-scale to support strong conclusions about generalization [4].

This limitation is directly relevant to the present thesis. Customer review data is opinion-grounded by nature: a claim such as "the service was slow" is not a world fact checkable against an encyclopedia, but a subjective experience that can only be verified against semantically equivalent statements in the review corpus. Prior work on factual consistency evaluation has shown that metrics developed on clean, objective benchmarks (e.g., news summarization) behave differently when applied to other domains such as clinical text [11], suggesting that domain shift can substantially affect evaluation behavior.

A further limitation in multi-document review data is that claims require different levels of proof. For example, a claim that "multiple customers complained about slow service" must be supported by several independent reviews. In contrast, a claim that "one customer complained about slow service" needs only one supporting review. Existing evaluation frameworks like FActScore use a binary support/non-support check against a knowledge source [4], which works well for single-fact verification against a canonical source. In a multi-document review setting, however, claims often describe broader patterns (e.g., "multiple customers complained about slow service") whose verification depends on how many independent reviews back the claim. This thesis treats that multi-source aggregation explicitly as part of the verification logic. To address this, the thesis introduces a custom verification approach in Section 7.7. The pipeline checks the actual meaning of the text rather than just matching exact words [17] and applies different evidence rules depending on whether a claim refers to one customer or to a broader pattern.

Another important limitation of FActScore, however, is that it measures precision but not recall. A model can achieve a high factual score by making a small number of safe, easily verifiable claims while omitting important information entirely. This limitation is especially serious in multi-document summarization. As Belem et al.'s results show [6], recall and coverage tend to degrade as the number of source documents grows, with some models showing meaningful drops, yet this dimension is not captured by factual precision metrics alone. This thesis therefore supplements fact-level precision with explicit information coverage evaluation against manually verified golden insights, specifically to measure what important signals were omitted from the summary.

For evaluation at scale, the LLM-as-a-Judge paradigm has become widely used. Zheng et al. [9] systematically evaluate this approach and find that GPT-4 as a judge reaches 85% agreement with expert human evaluators on the MT-bench benchmark, a figure that exceeds the measured human-human agreement of 81%. Liu et al. [18] introduce G-EVAL, a chain-of-thought evaluation framework that achieves a Spearman correlation of 0.514 with human judgments on the SummEval summarization benchmark, outperforming all prior automated metrics evaluated on that benchmark. Both studies note that judge-based evaluation must still be treated carefully because LLM judges may exhibit known biases such as position bias and

verbosity bias [9], and may prefer responses that share the generation style of the judge model itself [18].

The evaluation setup of this thesis draws on all three of these lines of work. It uses atomic-fact verification inspired by FActScore for factual accuracy, citation-based evaluation inspired by ALCE for attribution faithfulness, and a structured LLM-as-a-Judge pipeline for scalable scoring. It extends these approaches by adding explicit information coverage against golden insights and by applying the full evaluation pipeline to hybrid, subjective review data rather than the clean factual corpora used in prior work. The domain shift from Wikipedia-style passages to restaurant reviews introduces additional evaluation challenges, since claims are often about subjective experiences and recurring patterns rather than verifiable world facts.

3.6 Research Gaps and Positioning of This Thesis

The literature reveals five gaps that directly motivate this thesis.

First, existing evaluation frameworks primarily measure factual precision while giving limited attention to information coverage. A model that makes only a few safe claims can score well on precision metrics while omitting business-critical signals entirely.

Second, most studies focus on pure-text tasks and do not address the combination of unstructured review text with structured numerical indicators. The numeracy problems identified by White [7] remain largely unaddressed in multi-review summarization settings.

Third, commonly used benchmarks rely on clean, objective datasets such as Wikipedia or news corpora. These differ substantially from the subjective, heterogeneous, and often contradictory nature of customer review data, where the same restaurant can receive five-star and one-star ratings in the same week for ostensibly the same experience.

Fourth, grounding strategies are typically evaluated in isolation. Prior work studies long-context prompting, RAG, structured generation, or numerical grounding separately, rather than comparing them head-to-head under the same controlled conditions and with the same evaluation pipeline.

Fifth, attribution studies have focused almost entirely on pure-text question answering with retrieved passages, leaving open how citation behavior changes across different grounding architectures and how it behaves when evidence includes both short user-generated reviews and numerical summary statistics.

Relative to this literature, this thesis is positioned as a controlled comparative study. It brings together four grounding strategies, Plain Prompting, RAG, Aggregated-Statistics Grounding, and Template-Constrained Generation, and evaluates them within one unified pipeline on hybrid restaurant review data. It is equally important to clarify what this thesis does not attempt to contribute. The work does not introduce a new retrieval algorithm, a new LLM architecture, or a general-purpose denoising method for customer reviews. Instead, its contribution is empirical: it compares existing grounding strategies under the same dataset, prompt, model, and evaluation setup, and analyzes how they differ in factual accuracy, coverage, attribution faithfulness, and operational cost. Table 1 summarizes the most directly relevant prior studies and clarifies how this thesis relates to and extends each one.

The positioning also clarifies the scope and limits of this work. Compared with

Table 1: Representative prior studies and their relevance to this thesis.

Study	**Main Focus**	**Relation to This Thesis**
Lewis et al. [2]	Retrieval-Augmented Generation	Foundational RAG method evaluated in this thesis. Originally developed for knowledge-intensive pure-text tasks; this thesis tests it on hybrid review data and measures coverage explicitly.
Li et al. [13]	RAG vs. long-context comparison	Direct basis for comparing Plain Prompting and RAG. Prior work uses pure-text QA benchmarks; this thesis uses restaurant reviews and adds coverage and attribution as evaluation dimensions.
Li et al. [14]	Long context vs. RAG revisited	Extends the RAG/LC comparison with broader datasets. Confirms that LC is stronger for dense structured contexts and RAG for fragmented ones, which aligns with the thesis findings on coverage.
Belem et al. [6]	Hallucination in multi-document summarization	Directly motivates the information coverage metric and the scenario design. Their finding that recall drops as document count grows is tested in this thesis's high-volume Scenario 7 (Maximum Volume, 500 reviews).
White [7]	Numeracy limitations in summarization encoders	Justifies Aggregated-Statistics Grounding. Encoder-level failures in representing numerical values provide the rationale for supplying deterministic statistics directly in the prompt.
Osuji et al. [8]	Data-to-text generation survey	Relevant for structured numerical grounding but does not address joint handling of user-written review text and numerical signals. This thesis studies exactly that combination.
Mao et al. [15]; Geng et al. [16]	Constrained and structured generation	Directly relevant to Template-Constrained Generation. Prior work studies lexical constraints and schema compliance; this thesis evaluates their effect on factuality, coverage, and attribution in a hybrid summarization context.
Min et al. [4]	Atomic-fact factual precision (FActScore)	Provides the basis for the factual accuracy metric. A key difference is that this thesis supplements fact-level precision with explicit coverage evaluation against golden insights.
Gao et al. [5]	Attribution and citation evaluation (ALCE)	Provides the basis for citation precision and citation recall. This thesis extends citation evaluation to multi-review business summarization and adds the Fully Supported Claims Ratio as a third attribution metric.
Song et al. [12]	Grounded trustworthiness in RAG	Shows that prompting alone is insufficient to ground LLMs to provided documents. Motivates the comparison of structurally different grounding strategies rather than prompt variations alone.
Zheng et al. [9]; Liu et al. [18]	LLM-as-a-Judge evaluation	Justifies the automated judge-based evaluation pipeline. This thesis applies the paradigm to hybrid review summarization and validates it with a manual audit of sampled judgments.

much of the existing literature, the thesis gains realism by focusing on real-world business data and industrially relevant evaluation criteria. It also provides a broader head-to-head comparison than studies that examine only one grounding family at a time. The main limitations are domain specificity, since the data is English-language restaurant reviews from one industry partner, and a dependence on automated judge-based evaluation, even though this is supplemented by manual validation. These limitations are discussed in more detail in Section 9.

4 Problem Formulation

The preceding sections have introduced the relevant concepts and placed this work in the context of prior research. This section now defines the problem that the thesis addresses with greater precision. It describes the specific task, explains why it is difficult, states the purpose and research questions that guide the study, defines the scope of the experiment, and closes with a set of concrete, evaluable goals whose achievement corresponds directly to answering the research questions.

4.1 Problem Definition

This thesis studies the problem of *factual reliability* in AI-generated summaries of customer reviews. The work is carried out in collaboration with Zoined Oy, a business analytics platform whose customers include restaurant and retail businesses. As part of this collaboration, Zoined made Google Maps review data from its restaurant customers available for this study. The dataset is used under company data-access restrictions and is not made publicly available. The data is hybrid by nature. Each review record combines unstructured written text with structured numerical signals such as an overall star rating and aspect-level scores for food, service, and atmosphere. Although the experiments use restaurant data specifically, the grounding strategies and evaluation framework studied here are relevant to business domains where similar hybrid review data must be summarized reliably. Even though platforms such as Zoined can support different types of queries over review data, this thesis focuses only on summarization.

At first glance, this may sound like a standard summarization task. In practice, it is harder because the input is not only free text. It consists of *hybrid review data*: short and often highly variable written reviews combined with structured numerical signals such as star ratings, aspect ratings for food, service, and atmosphere, review counts, and aggregated statistics. The problem is therefore not only to generate a readable summary, but to generate one that is grounded in and consistent with all of this provided data, both textual and numerical.

This problem is difficult for several reasons. First, LLMs are known to produce text that sounds convincing but is not supported by the source input, a problem known as hallucination [1, 10]. In the terminology introduced in Section 3, the relevant failure mode here is *faithfulness hallucination*: the generated output diverges from the provided context rather than from world knowledge. In review summarization, this can appear as invented trends, exaggerated sentiment, or claims about service or food quality that are not present in the reviews.

Second, the task becomes harder when the model has to handle both text and numbers. Earlier work shows that summarization models do not always represent quantitative values reliably, which can lead to numerical errors and quantity hallucinations [7]. For example, a model may describe the overall feedback as strongly positive even when the ratings are mixed, or it may state a trend that does not match the actual rating distribution.

Third, review summarization is a multi-document problem. A single restaurant may have dozens or even hundreds of short reviews, and important information may be spread across them unevenly. Some insights appear many times, while others appear only once but still matter, such as a serious complaint about hygiene, allergy

handling, or service failure. A reliable system must surface these low-frequency but high-importance signals rather than only reflecting the majority opinion.

Fourth, verifiability is a requirement in business settings. It is not enough for a summary to sound reasonable. It must also be traceable back to the source data, so that a business user can check why a particular claim was made. Research on citation-based generation and grounded answering shows that current LLMs often fail to provide faithful attribution and may continue generating confident outputs even when the source data is incomplete or insufficient [5, 12]. In this thesis, this matters especially in scenarios with sparse text, conflicting signals, or reviews that contain ratings but little or no written explanation. In such cases, a reliable system should acknowledge the limits of the available evidence rather than overstating conclusions.

Taken together, these challenges define the main problem of this thesis: *how different grounding strategies affect the factual accuracy, information coverage, attribution faithfulness, and practical cost of LLM-generated summaries for hybrid customer review data*. The goal of this work is not to build a general-purpose summarizer, but to determine which grounding strategy is most suitable for this task within the experimental setup used in the thesis.

4.2 Purpose and Research Questions

The purpose of this thesis is to evaluate how different grounding strategies affect the reliability and practical usefulness of LLM-generated summaries for hybrid customer review data. More specifically, it studies how grounding affects four key properties of the generated summaries: factual accuracy, information coverage, attribution faithfulness, and operational cost.

To fulfill this purpose, the thesis addresses the following research questions:

- **RQ1:** How do different grounding strategies (Plain Prompting, Retrieval-Augmented Generation, Aggregated-Statistics Grounding, and Template-Constrained Generation) affect the factual accuracy and information coverage of generated summaries?
- **RQ2:** How do these grounding strategies affect explainability, specifically the model's ability to provide faithful attribution by linking generated claims back to the correct reviews or numerical inputs?
- **RQ3:** What are the operational trade-offs of these grounding strategies in terms of inference latency and token usage?

These three questions together cover the main dimensions of the purpose: output quality (RQ1), traceability (RQ2), and practical deployment cost (RQ3). When all three are answered, the purpose is fulfilled.

4.3 Motivation

This problem matters because customer review summaries can influence real decisions. In a business analytics platform such as Zoined, a summary may shape how a company understands customer satisfaction, service quality, or changes in feedback over time. If the summary is wrong, the resulting decision may also be wrong. A summary that overstates praise can hide real problems. A summary that misses recurring complaints

can delay corrective action. A summary that invents trends or misrepresents ratings leads users to trust a false picture of customer feedback.

Trust is therefore central. In an enterprise setting, a summary is not useful only because it is fluent. It must also be possible to verify. A business user should be able to ask why the system claims that customers complain about service, or why it says that food quality is inconsistent, and the system should be able to point back to the relevant reviews or statistics. Without this traceability, the summary becomes difficult to check and harder to act on. This is especially important in analytics products, where outputs may be used repeatedly by users who want concise answers but still need confidence that those answers reflect the data.

There is also a research motivation. Existing work has studied hallucination, retrieval-augmented generation, long-context prompting, citation quality, and constrained generation, but these are typically evaluated in isolation and on cleaner datasets such as Wikipedia, news articles, or question-answering benchmarks. They do not fully address the setting studied in this thesis, where user-generated restaurant reviews are combined with structured numerical business signals. This thesis contributes by evaluating four grounding strategies in a controlled, unified setup that is closer to a real restaurant review analytics use case than any existing benchmark.

Both motivations converge on the same purpose: understanding the trade-offs between grounding strategies so that the most suitable one can be selected based on the deployment requirements of the application.

4.4 Goals

To answer the research questions and fulfill the purpose of the thesis, the work is organized around five concrete, evaluable goals. Together, achieving all five goals is sufficient to answer RQ1, RQ2, and RQ3.

1. **Construct a verified evaluation dataset.** Build a manually verified evaluation set consisting of curated scenarios and golden insights that captures both common review patterns and difficult edge cases such as conflicting text-rating signals, sparse text, and mixed sentiment. This goal supports the research questions by providing the reference basis against which summary coverage is measured.

2. **Implement multiple grounding pipelines.** Develop four summarization pipelines based on Plain Prompting, Retrieval-Augmented Generation, Aggregated-Statistics Grounding, and Template-Constrained Generation, configured to run under the same experimental conditions. This goal is the prerequisite for all comparative analysis.

3. **Define and apply a systematic evaluation framework.** Evaluate generated summaries with respect to factual accuracy, information coverage, and attribution faithfulness using a consistent, automated judge-based pipeline supported by a manual audit. This goal directly addresses RQ1 and RQ2.

4. **Measure operational performance.** Record inference latency and total token usage for each grounding strategy across all scenarios. This goal directly addresses RQ3 and provides the data needed to assess practical deployment cost.

5. **Compare strategies and derive recommendations.** Analyze the results across scenarios using descriptive and inferential statistical methods to identify the strengths, weaknesses, and trade-offs of each strategy, and to determine which approach is most suitable for reliable summarization of hybrid business review data. This goal fulfills the overall purpose of the thesis.

4.5 Scope and Limitations

This subsection defines the scope of the study and highlights the main limitations of the experimental setup. Where relevant, each limitation is paired with a mitigation strategy used in the thesis.

- **Domain and language scope.** The experiments are limited to English-language Google Maps restaurant reviews provided by Zoined Oy, together with their associated numerical signals. This supports consistency and reproducibility across experimental runs, but it also means that the findings are tied to one domain, one language, and one data source. The results may not transfer directly to other domains such as medical, legal, or financial text, or to multilingual review data. The findings should therefore be interpreted as specific to restaurant reviews and closely related review domains such as hotel reviews, rather than automatically generalizable to other settings.
- **Quality of the evaluation reference.** Information coverage is measured against manually written golden insights. If an important insight is missing from the golden set, coverage may be underestimated for that scenario. To reduce this risk, the golden insights were designed to cover both common patterns and edge cases, including conflicting text-rating signals, sparse text, and missing textual evidence. Their design rules are described in Section 7.
- **Fixed model family.** Both the generator and judge come from the OpenAI GPT family, as recommended by Zoined, which currently uses OpenAI in its platform. The generator uses `gpt-5.4-mini` and the judge uses `gpt-5.4`, both fixed across all experiments. The thesis therefore compares grounding strategies within one model setup rather than across model families. The results show how the four strategies behave with these models, but not whether a different model family would yield the same rankings. Using the same family for generation and evaluation also introduces a potential same-family evaluation bias, partly mitigated by a larger judge model and a manual audit of sampled judgments.
- **Automated evaluation.** The evaluation relies heavily on an LLM-as-a-Judge pipeline. This enables large-scale evaluation and fine-grained checking of summary claims using atomic facts and citation verification [4]. However, it also introduces a dependency on the behavior of the evaluator model. To reduce this risk, the judge model is kept fixed with a low temperature and a fixed random seed, and a random sample of automated judgments is manually audited to assess agreement with human review. The audit methodology is described in Section 7.8.

5 Method

This thesis uses a controlled experiment as its primary scientific method. The research questions require objective, measurable comparison between alternative strategies along four dimensions: factual accuracy, information coverage, attribution faithfulness, and operational cost. By holding all main experimental conditions fixed and varying only the grounding strategy, the study enables a direct and reproducible comparison of each strategy's effect.

Two alternative approaches were considered and rejected. A case study could provide contextual understanding of system use but would not allow sufficient control to isolate the grounding strategy from other factors such as data drift or user behavior. A survey would capture subjective perceptions of summary quality rather than objective system behavior, and could not reliably measure factual errors, attribution quality, or computational performance.

Figure 1 gives an overview of the full experimental pipeline. The subsections that follow explain each component and justify the design choices made.

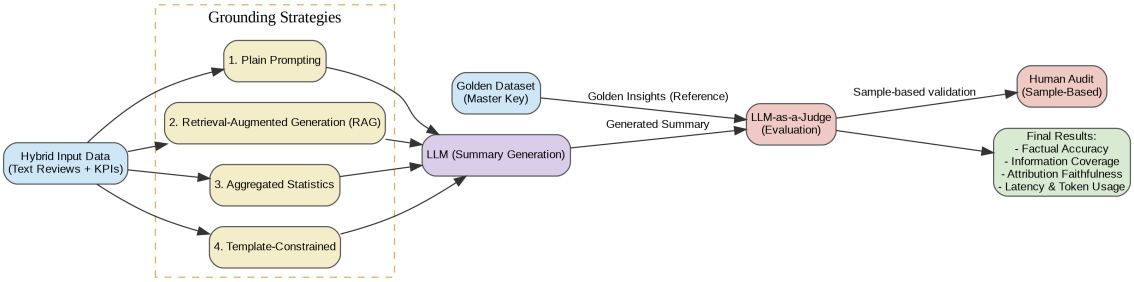

Figure 1: Overview of the experimental pipeline. Hybrid review data is passed through four grounding strategies to the LLM generator. The generated summary is evaluated by an LLM-as-a-Judge against the golden insights and source data. A human audit validates a sample of automated judgments. Final results address RQ1 (factual accuracy and coverage), RQ2 (attribution faithfulness), and RQ3 (operational cost).

5.1 Research Design

The study is designed as a single-factor controlled experiment in which the grounding strategy is the only manipulated variable. All other conditions are held fixed so that differences in results can be attributed to the grounding strategy. Table 2 summarizes the formal experimental design.

The experimental unit is one (scenario, strategy) pair. Each run applies one grounding strategy to one scenario and measures the generated output.

All model calls were configured with $T = 0$ and `seed=42`. Temperature zero reduces sampling randomness, while the fixed seed was used to make repeated calls with the same prompt and parameters as consistent as possible. However, hosted LLM APIs do not guarantee full determinism: OpenAI describes seed-based reproducibility as a best-effort mechanism and notes that outputs may still vary even when the same seed and parameters are used [19]. These settings were therefore used to minimize avoidable generation variation so that observed differences are more likely to reflect the grounding design rather than sampling noise, consistent with controlled LLM-as-a-Judge evaluation practice [9, 18]. Both the generator

Table 2: Summary of the formal experimental design.

Element	**Specification**
Independent variable	Grounding strategy
Treatments	Plain Prompting, RAG, Aggregated-Statistics Grounding, Template-Constrained Generation
Dependent variables	Factual accuracy, information coverage, attribution faithfulness, inference latency, token volume
Objects of study	Summaries generated by each strategy, evaluated against source data and golden insights
Experimental unit	One (scenario, strategy) pair
Held constant	Generator model, judge model, base task instruction, evaluation pipeline, temperature ($T = 0$), random seed (`seed=42`) for all model calls
Repetitions	Three repeated runs per unit; runs are averaged per scenario before inferential testing
Total executions	$11 \times 4 \times 3 = 132$

(`gpt-5.4-mini`) and judge (`gpt-5.4`) support an optional internal reasoning step, disabled by setting `reasoning_effort=none`. This was done for two reasons. First, it eliminates variable latency and token counts that would make operational cost comparisons unstable. Second, it reduces additional variation caused by internal model reasoning and keeps the comparison focused on the grounding strategy. This does not make the model fully deterministic, but it makes operational cost and evaluation behavior more comparable across strategies.

Three repeated runs per (scenario, strategy) pair yielded 132 executions ($11 \times 4 \times 3$). This allows run-level dispersion to be reported as a transparency measure alongside the mean. The choice of three runs was made for practical reasons, mainly to keep the API cost manageable while still allowing variation between repeated runs to be observed.

LLM-as-a-Judge evaluation carries known risks, including position bias (favoring responses placed first in the prompt) and verbosity bias (preferring longer, wordier outputs regardless of accuracy), where longer outputs may be preferred even when they are not more accurate [9]. To reduce these risks, the judge prompts used in this thesis specify explicit verdict labels, explanation fields, and domain-specific verification rules. For example, the evaluation distinguishes between claims that can be supported by a single review and broader pattern claims that require evidence from multiple reviews. The full prompts and their adaptations to the review summarization task are presented in Section 7. In addition, a random sample of automated judgments was manually audited to assess whether the judge decisions aligned with human interpretation, as discussed in Section 7.8.

5.2 Data and Scenario-Based Evaluation Design

The empirical material consists of Google Maps review data from restaurants that are customers of Zoined. The data was provided by Zoined for this study under company data-access restrictions. Before any model processing, personally identifiable information was removed from the dataset, so the LLM received only the retained review fields. These retained fields combine unstructured review text with structured

numerical signals, making the dataset hybrid by nature. Full details of the data format and data-handling procedure are provided in Section 7.

A public benchmark was not used. Common summarization benchmarks such as CNN/DailyMail or XSum [20, 21] consist of plain news text, which does not match the task here. The research questions depend on how grounding strategies handle hybrid data, where user-written comments appear alongside structured numerical signals such as star ratings and aspect scores. A domain-specific dataset was therefore necessary. This choice is consistent with prior work showing that factual evaluation methods behave differently on real-world domains than on general benchmarks [11].

Scenario-based evaluation. Rather than evaluating strategies on one large undifferentiated dataset, the data was structured into distinct evaluation scenarios, each representing a specific data condition. This design was chosen for three reasons.

First, it improves experimental control: each scenario isolates one condition so that condition-specific effects are observable explicitly.

Second, it exposes failure modes unlikely to appear in sufficient quantity in a random sample. Models have been shown to invent answers when evidence is missing [12], mishandle quantitative information [7], and lose important facts in multi-document settings [6].

Third, it enables reference-based coverage measurement. Measuring whether a summary captures important information requires a known set of expected insights per case, which a fixed scenario makes possible. Scenario construction and targets are described in Section 7.

Golden insights. For each scenario, golden insights were written manually: short, verifiable statements capturing the key facts and patterns in the data. They serve as the reference for information coverage. A summary should express the meaning of each insight even if the exact wording differs.

Manual writing was chosen over LLM generation because LLM-generated references would be stylistically similar to LLM-generated summaries, potentially biasing coverage scores toward models whose writing style resembles the reference generator. This approach was also preferred over ROUGE or BLEU, which reward lexical similarity rather than factual content and are weak indicators of factual correctness [17]. Measuring coverage at the level of meaning complements the atomic-fact approach used for factual accuracy [4]. The full insight set and design process are in Section 7.

5.3 Selection of Grounding Strategies

Four strategies were selected: Plain Prompting, RAG, Aggregated-Statistics Grounding, and Template-Constrained Generation. Each targets a different weakness in hybrid review summarization, and together they cover the main design choices for constraining LLM generation through prompt-level or output-level conditioning.

The selection was kept intentionally narrow. Other approaches were excluded for clear reasons:

- **Fine-tuning** adapts model weights on domain data and represents a fundamentally different form of grounding from prompt-level conditioning [12]. Including it would introduce a training data variable inseparable from the grounding effect.

- **Knowledge-graph grounding** anchors generation to a structured graph of entities and relations, shown to improve trustworthiness in enterprise question answering [22]. It requires a domain-specific ontology that does not exist for restaurant review data and would need substantial manual construction effort outside this study's scope.
- **Adaptive retrieval** dynamically routes queries between retrieval and long-context approaches [13], adding a routing component difficult to separate from the grounding strategy itself.

Combinations of strategies, such as adding aggregated statistics to a RAG pipeline, were not tested here but are a natural direction for future work (Section 9).

These strategies apply constraints at different stages of the pipeline, at the input level for RAG, at the output level for Template-Constrained Generation, and within the prompt context for Plain Prompting and Aggregated-Statistics Grounding. Importantly, none of these methods modify the model's weights. While early RAG systems required joint, end-to-end fine-tuning [2], all four strategies in this thesis rely on training-free grounding. This approach leverages in-context learning (ICL) during inference to anchor outputs to external evidence without changing any internal parameters [12, 13]. This has become the standard workflow for grounding modern, closed-source black-box LLMs [1].

Plain Prompting. Plain Prompting passes all reviews directly to the model with no retrieval, constraints, or pre-computed statistics. It serves as the long-context baseline, providing a fixed reference point against which the other strategies are compared. It also tests a practical question: does giving the model more context always help, or can longer input create its own problems? Prior work shows that long-context approaches increase token cost [13] and that models can miss important facts as input grows larger [6].

Retrieval-Augmented Generation (RAG). RAG is one of the most widely deployed approaches for grounding LLM outputs in external evidence [2]. A retriever selects the most relevant reviews, and only those are passed to the model. The general RAG architecture follows Lewis et al. [2]. In this thesis, it is instantiated with BM25 retrieval, chosen for transparency and reproducibility, and a fixed retrieval depth of $k = 8$, consistent with recent RAG-versus-long-context setups [13].

Retrieval in the RAG strategy used the classical BM25Okapi formulation (Robertson–Spärck Jones) via the open-source `rank_bm25` package [23], rather than dense embedding models or hierarchical retrievers such as RAPTOR. This choice was based on three considerations.

- **Lexical nature of the data.** Customer reviews are short, informal, and centered on specific entities such as dish names or complaint types, which favors exact lexical matching. Dense embedding models are known to lose accuracy on this kind of informal, out-of-domain text [10], whereas sparse retrievers such as BM25 are built specifically for literal term overlap [2, 24].
- **Sensitivity to rare signals.** Mallen et al. [3] show that sparse retrieval outperforms dense retrieval on rare, low-frequency information. This matters

here because a single mention of a serious issue, such as a hygiene complaint, is exactly the kind of rare but important signal that dense retrieval's averaged vector representations tend to smooth over and lose.

- **Absence of long-range structure.** Reviews are short, independent entries rather than a connected narrative, so there is no long-range thematic structure for a hierarchical retriever like RAPTOR to exploit [10, 14]. Such methods recursively summarize text to capture cross-document themes, adding latency and hallucination risk without any benefit on data that is already atomic.

Because reviews are short, the length-normalization corrections used in BM25L or BM25+ were unnecessary, so the standard BM25Okapi formulation was used unmodified. It was paired with a lightweight, LLM-assisted query expansion step (Section 7.6) to bridge the gap between the abstract task wording and the concrete vocabulary used in reviews, keeping the retrieval pipeline transparent and reproducible.

A retrieval depth of $k = 8$ was chosen because recent studies show that selecting the top 5 to 10 chunks typically yields the strongest performance before results begin to diminish [14]. Keeping this strict constraint fixed across all scenarios, including Scenario 7 (Maximum Volume, 500 reviews), was an intentional methodological choice. It stress-tests how a standard RAG architecture handles high data volume and exposes coverage limitations when critical signals are widely distributed. An LLM-assisted query expansion step bridges the lexical gap between abstract task wording and concrete review text [24]. Implementation details of the retrieval pipeline are in Section 7.6.

Aggregated-Statistics Grounding. This strategy augments the full review text with a numerical summary block (rating statistics, aspect-level averages, review volume counts) pre-computed from the dataset prior to runtime and injected into the prompt. These statistical attributes are derived from the current review dataset and are kept constant across all scenarios. The motivation comes from work showing that summarization models struggle to represent quantitative values reliably from raw text [7]. Supplying these aggregated values explicitly bypasses the unstructured text-processing step, which prior research identifies as the primary driver of quantity hallucinations [7]. The full statistics block format is in Section 7.6.

Template-Constrained Generation. This strategy requires the model to generate its summary within predefined JSON schema. The schema organizes the output into distinct fields derived from the base task instruction: overall sentiment, praise themes, complaint themes, and specific signals. This strategy directly investigates how structured output constraints affect citation faithfulness. Prior studies demonstrate that enforcing structured formats can successfully improve factual consistency [15] and guarantee perfect schema compliance [16]. However, these works do not investigate citation quality. When a structured template requires claims to include supporting sources, it is unclear whether this improves evidence tracking or creates "schema pressure" that encourages the model to add loosely relevant sources just to satisfy the format. By combining a predefined output template with OpenAI's JSON response format, this strategy tests whether a more disciplined output structure improves the model's evidence grounding, or conversely, whether it induces artificial citation

behaviors. The specific theme categories and full schema details are provided in Section 7.6

5.4 Evaluation Methodology

An automated LLM-as-a-Judge pipeline serves as the primary evaluation mechanism, supplemented by a manual audit. Fully manual evaluation would be difficult to scale across 132 runs and would introduce inconsistency from evaluator fatigue. ROUGE and BLEU were not used: they reward lexical similarity without assessing evidential support, and a fluent hallucinated summary can score highly on overlap [17, 11]. Using a strong LLM as a judge is supported by recent work showing frontier models can reliably perform semantic comparison rather than lexical matching [9, 18].

Figure 2 shows how each evaluation dimension is handled.

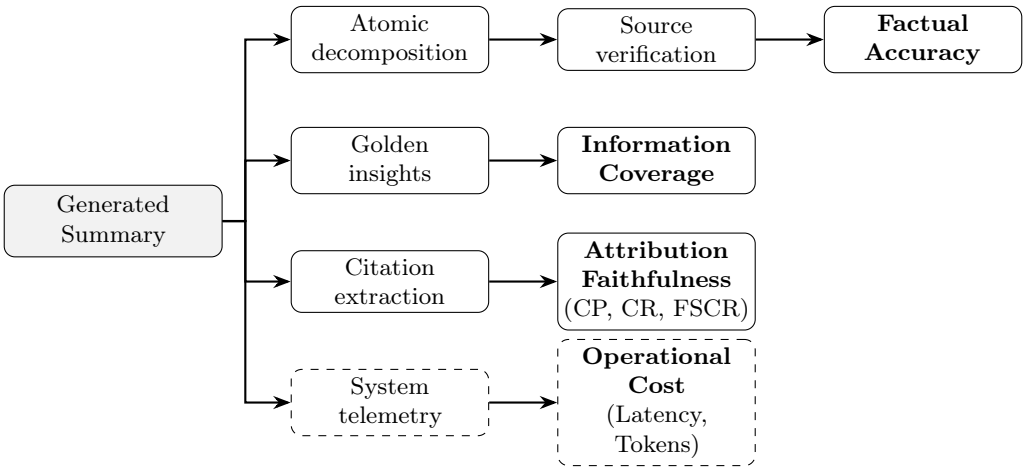

Figure 2: Evaluation pipeline for a generated summary across four dimensions: factual accuracy (atomic decomposition and source verification), coverage (golden insights), attribution faithfulness (CP = Citation Precision, CR = Citation Recall, FSCR = Fully Supported Claims Ratio), and operational cost (system telemetry, no judge involvement).

For factual accuracy, each summary is decomposed into independent atomic claims following the FActScore paradigm [4], and each claim is verified individually against the scenario's source data, as defined in Section 2.3.

For information coverage, the summary is compared against the scenario's golden insights, as defined in Section 2.3.

For attribution faithfulness, the judge verifies whether citations attached to claims actually support those claims. Three metrics are assessed: Citation Recall, Citation Precision, and the Fully Supported Claims Ratio. This builds upon the citation-focused evaluation framework of Gao et al. [5], extending their precision and recall metrics with a strict fully-supported constraint. Together, these metrics separate whether claims are backed at all, whether citations are appropriate, and whether every citation for a claim is valid.

For operational cost, inference latency and total token consumption are logged programmatically for every run, as defined in Section 2.3.

For grounded refusal, some scenarios were designed with missing or severely limited textual evidence (Section 7.4). A reliable system should acknowledge the limits of available evidence rather than generate unsupported claims. This is evaluated

explicitly because LLMs often generate confident outputs even when source data is insufficient [12].

For judge prompt structure, prompts require an explicit reasoning or explanation field before the final verdict rather than asking for a verdict alone. This improves consistency and makes judge decisions easier to audit [18, 9]. Full prompts are in Appendix D.

For the manual audit, a random sample of judge verdicts is manually reviewed to assess alignment with human judgment, serving as a construct-validity check rather than an alternative scoring mechanism.

5.5 Statistical Analysis

Differences between strategies are assessed using non-parametric tests throughout. With $n = 10$–11 paired scenarios, parametric tests such as repeated-measures ANOVA would require normality assumptions that are difficult to justify at this sample size. Several metrics are also bounded in $[0, 1]$, such as factual accuracy, information coverage, citation precision, citation recall, and fully supported claims ratio, and several scores cluster near the upper end. Non-parametric tests are therefore more appropriate for this comparison.

Before inferential testing, the three repeated runs for each scenario–strategy pair are averaged. This produces one scenario-level value per strategy and prevents repeated runs from being treated as independent observations.

For each evaluation metric, the inferential comparison is framed using the following hypotheses:

H_0**:** There is no systematic difference in performance between the four grounding strategies for the given metric.

H_1**:** At least one grounding strategy differs in performance from the others for the given metric.

The hypotheses are tested separately for factual accuracy, information coverage, citation precision, citation recall, fully supported claims ratio, inference latency, and token volume. For each metric, a Friedman test is first used to determine whether there is an overall difference among the four grounding strategies. Statistical significance is assessed at $\alpha = 0.05$. If the Friedman test is significant, Wilcoxon signed-rank tests are then used to compare the six possible pairs of strategies, with Holm–Bonferroni correction applied to account for multiple comparisons. Effect sizes are reported using Kendall's W for the overall comparison and rank-biserial correlation r for the pairwise comparisons, indicating the strength of the observed differences alongside statistical significance. The detailed results for each metric and the handling of the incomplete case of Scenario 3 (Missing Text) are presented in Section 8.

5.6 Threats to Validity

Table 3 summarizes the main validity threats and the mitigation steps used in this thesis.

Table 4 maps each research goal to the methodological choice that fulfills it.

Table 3: Threats to validity and mitigations.

Dimension	**What it asks**	**Threat and mitigation**
Internal validity	Are the observed differences really caused by the grounding strategy?	*Threat:* Differences between strategies could be caused by confounds such as model choice, prompt wording, input data, or decoding settings. *Mitigation:* All non-strategy conditions are held constant and each unit is run three times to expose run-level variation (Section 5.1).
Construct validity	Do the metrics measure the intended concepts?	*Threat:* Automated judge scores may not perfectly capture factual accuracy, coverage, or attribution faithfulness, and latency may reflect network or API load rather than the strategy itself. *Mitigation:* Each metric is explicitly defined and decomposed (atomic-fact verification, golden insights, citation-level checks; Section 5.4), a manual audit checks judge–human alignment (Section 7.8), and latency is averaged over three runs.
External validity	Can the results generalize beyond this study?	*Threat:* Results may not generalize beyond English-language restaurant reviews, Zoined's data, or the specific OpenAI models used. *Mitigation:* Scenarios cover diverse review conditions, including edge cases such as missing text and text-rating conflict (Section 5.2); generalization to other domains, languages, and model families requires further study.
Conclusion validity	Are the conclusions statistically and logically supported?	*Threat:* Statistical conclusions may be unreliable given the small number of scenarios or inappropriate test choices. *Mitigation:* Non-parametric tests with Holm–Bonferroni correction and reported effect sizes guard against this (Section 5.5).

Table 4: Mapping of research goals to methodological choices.

Research Goal	**Method**
Construct a verified evaluation dataset	Build scenario-based review datasets with manually written golden insights
Implement multiple grounding pipelines	Build four pipelines using the same generator model and same base task prompt
Define and apply a systematic evaluation framework	Use LLM-as-a-Judge evaluation with atomic-fact checking, coverage scoring, and manual audit
Measure operational performance	Log token usage and inference latency for every run
Compare strategies and derive recommendations	Compare results using descriptive statistics and statistical significance tests

6 Ethical and Societal Considerations

This thesis does not involve recruiting or directly interacting with human participants through interviews, surveys, or experiments. The evaluation data consist of customer reviews collected from Google Maps and made available by Zoined Oy for this study under company data-access restrictions. Although individual reviews may be publicly visible on Google Maps, the data relates to Zoined's commercial clients. Restaurant identities are therefore not disclosed in this thesis, and the raw dataset is not published.

To protect reviewer privacy, identifying fields such as reviewer names, profile URLs, and photo URLs were excluded from the experimental data. The retained information included the review text, overall and aspect-level ratings, publication date, and an internal review ID used for citation. Restaurant names and other business identifiers were not passed to the LLM as separate input fields.

The retained review text could nevertheless in principle be traced back to original Google Maps entries through text search. This residual privacy risk is acknowledged. To reduce the risk, the raw dataset is not published and restaurant identities are not disclosed in the thesis.

From a broader societal perspective, hallucinated or unsupported summaries can introduce misinformation and reduce trust in business analytics [1, 10]. This is particularly relevant when generated summaries are used to support business decisions. A central motivation of this thesis is therefore to evaluate whether grounding strategies can improve factual accuracy and traceability.

Operational efficiency also has economic and environmental implications. Strategies that use fewer tokens require less model processing and can reduce operational cost. Although energy consumption is not measured directly in this study, computational efficiency is relevant to economically and ecologically sustainable deployment. This remains a secondary consideration, but it supports evaluating grounding strategies not only by output quality but also by operational cost.

7 Experimental Setup and Implementation

This section describes the practical execution of the controlled experiment defined in Section 5. It covers the dataset and scenario construction, the base task instruction and its analytical goals, the golden insight framework, model configuration, the grounding-specific prompt construction for each strategy, and the automated evaluation pipeline. Technical parameters are documented in sufficient detail to allow reproduction.

7.1 Dataset and Hybrid Data Format

The evaluation dataset consists of Google restaurant reviews provided by Zoined Oy. Before any experimental processing, an anonymization step removed all reviewer identifiers, including reviewer profile URLs, reviewer names, and reviewer photo URLs. The retained fields for each review are: `text` (the written comment, which may be absent), `stars` (overall 1–5 star rating), `reviewDetailedRating` (optional aspect-level sub-ratings for food, service, and atmosphere), and `publishedAtDate` (review timestamp). Restaurant names are not disclosed to protect Zoined's customer identities.

What distinguishes this dataset from standard text summarization benchmarks [20, 21] is its hybrid structure. Each scenario contains two types of evidence: individual review records and a pre-computed aggregated statistics block. The review records provide unstructured text alongside structured star ratings, while the statistics block provides a deterministic numerical summary computed from the full review set before any experiment runs. Figure 3 illustrates this dual structure with values from Scenario 1 (Chronological Sample).

This hybrid structure creates a challenge absent from standard benchmarks: the model must reconcile written opinion with numerical ratings that may point in a different direction. It also means that some reviews carry no written text at all. A reliable system should not invent textual content when no text is present.

7.2 Base Task Instruction and What It Tests

Before describing the scenarios, it helps to introduce the base task instruction, since the scenarios were designed to stress-test specific aspects of it. The same instruction was used in all four strategies and all 11 scenarios:

> *"Based on the customer reviews, give a comprehensive summary of this restaurant. What do customers consistently praise or complain about, and are there any specific or recurring patterns worth highlighting?"*

The instruction is intentionally open-ended. It does not prescribe a format or structure, so Plain Prompting, RAG, and Aggregated-Statistics Grounding produce free prose. Template-Constrained Generation receives the same task question but adds a structured output schema aligned with the instruction. The schema design is described later in this section.

Table 5 shows this decomposition. The golden insights for each scenario were also written to include at least one insight per theme, so that coverage evaluation tests all four analytical dimensions of the instruction, not just the most obvious ones.

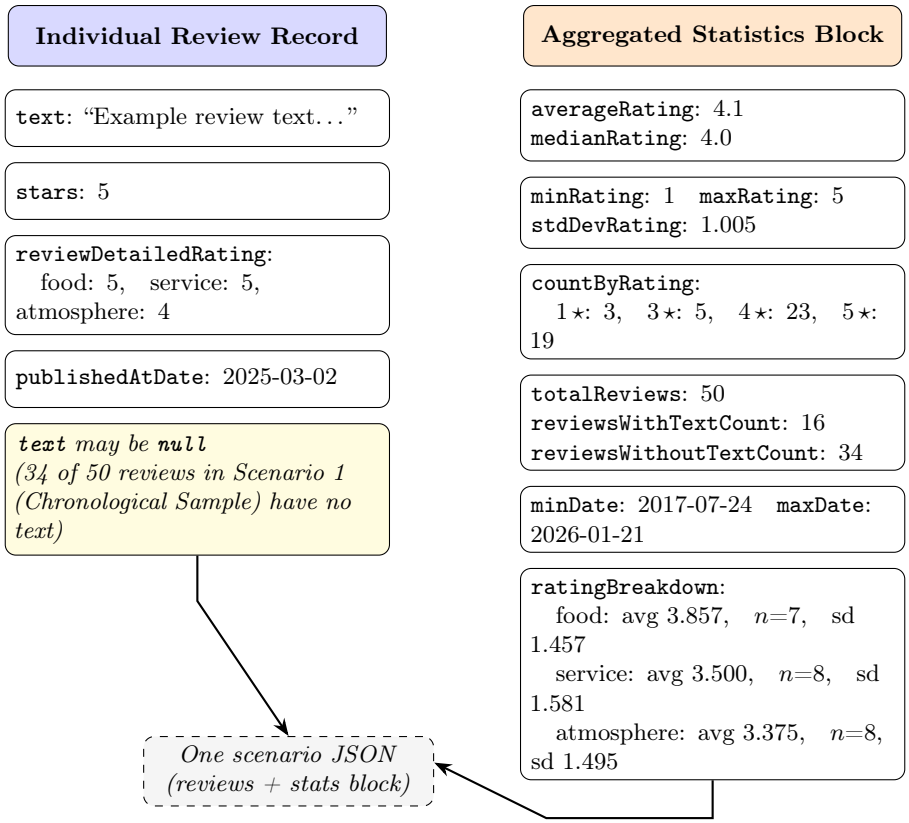

Figure 3: Hybrid data structure used in this thesis. Each scenario JSON contains individual review records (left), which may have null text, and a pre-computed aggregated statistics block (right). Values shown are from Scenario 1 (Chronological Sample, 50 reviews). The statistics block includes overall rating statistics, review volume and text availability counts, date range, and aspect-level breakdowns for food, service, and atmosphere. It is injected into the prompt only by the Aggregated-Statistics Grounding strategy; all four strategies receive the individual review records. The review text shown is a synthetic placeholder used only to illustrate the data format.

Table 5: Decomposition of the base task instruction into four evaluation themes. Each theme corresponds to a golden insight type and to a field in the Template-Constrained JSON schema.

Theme	**Instruction segment**	**What it tests**
Overall sentiment	*"give a comprehensive summary of this restaurant"*	Whether the model forms a coherent overall picture of customer sentiment rather than only listing individual details. In Scenario 1 (Chronological Sample), this corresponds to the insight *"Customer sentiment is predominantly positive, with positive ratings outnumbering negative ones."*
Praise	*"What do customers consistently praise"*	Whether the model identifies praise themes that appear across the review set. In Scenario 1 (Chronological Sample), this corresponds to *"Multiple customers specifically praise the baked potatoes"* and *"Multiple customers describe the service as friendly and fast."*
Complaint	*"or complain about"*	Whether the model surfaces recurring negative signals rather than suppressing them when positive reviews dominate. In Scenario 1 (Chronological Sample), this corresponds to *"A single customer explicitly states that customer service has become very poor."*
Specific signal	*"specific or recurring patterns worth highlighting"*	Whether the model preserves specific or recurring patterns that are worth highlighting. In Scenario 1 (Chronological Sample), one example is the insight *"A single customer was upset because their child's favorite chicken BBQ baguette was only available once despite dozens of visits."*

7.3 Scenario Design and Construction

The dataset was structured into 11 evaluation scenarios rather than one undifferentiated pool. The rationale is described in Section 5.2. This subsection describes what each scenario contains and what challenge it was designed to expose. Eight scenarios draw from real Google reviews provided by Zoined's restaurant customers. Three scenarios (Scenario 6 (High Information Density), Scenario 8 (Ambiguous Opinions), and Scenario 11 (Text-Rating Conflict)) are fully synthetic: their review texts were written specifically to satisfy conditions not present in the real provided data. Scenario 3 (Missing Text) uses the same 50 reviews as Scenario 1 (Chronological Sample) but with all text fields programmatically removed, making it a controlled modification rather than a fully synthetic case. Most scenarios contain 50 reviews, chosen as a practical balance: enough to surface recurring and minority signals, but still manageable for manual golden-insight construction and verification. Smaller and larger scenarios were added separately to test scarcity and maximum-volume conditions.

Table 6 summarizes all 11 scenarios. For each, the table gives the number of reviews, real or synthetic origin, average star rating, and the diagnostic challenge it was constructed to expose.

Table 6: Overview of the 11 evaluation scenarios. Each scenario was constructed to expose a specific challenge for grounded summarization. Average rating is on a 1–5 star scale. Detailed descriptive statistics for all scenarios are reported in Appendix A.

Scenario	**n**	**Source**	**Avg.**	**Diagnostic challenge**
Scenario 1 (Chronological Sample)	50	real	4.1	General case; 16/50 reviews have text. Tests preservation of low-frequency complaints in mixed sentiment.
Scenario 2 (Needle in a Haystack)	50	real	4.9	Overwhelmingly positive reviews with one hygiene-related needle signal. Tests whether minority business-critical signals survive [6].
Scenario 3 (Missing Text)	50	modified	4.1	Same as Scenario 1 (Chronological Sample) with all text fields nulled. Tests grounded refusal when no textual evidence for specific praise/complaints is available [12].
Scenario 4 (Polarized Opinions)	50	real	4.3	33 positive vs 17 critical reviews. Tests preservation of both sentiment poles rather than collapsing them; consistent with the multi-document coverage concerns raised in [6].
Scenario 5 (Temporal Shift)	50	real	3.2	25 highly positive (2022–23) vs 25 negative (2024) reviews. Tests detection of temporal sentiment change.
Scenario 6 (High Information Density)	50	synthetic	3.7	Each review introduces a unique signal. Tests whether details are silently dropped under coverage pressure [6, 4].
Scenario 7 (Maximum Volume)	500	real	4.2	Full review history. Tests long-context scaling and fixed top-k retrieval at scale [13].
Scenario 8 (Ambiguous Opinions)	50	synthetic	3.1	Reviews contain vague, mixed, or uncertain opinions. Tests whether the model preserves uncertainty rather than forcing a strong positive or negative conclusion; related to the tendency of LLMs to fabricate when the source signal is weak [6].
Scenario 9 (Concrete Insights)	50	real	3.5	Reviews with named dishes, exact times, specific incidents. Tests preservation of actionable detail.
Scenario 10 (Data Scarcity)	5	real	3.0	Only 5 reviews available. Tests whether the model acknowledges sample limits [12].
Scenario 11 (Text-Rating Conflict)	10	synthetic	3.0	Low-star reviews paired with positive text and high aspect ratings, high-star reviews paired with negative text and low aspect ratings. Tests whether the model can reconcile conflicting textual and numerical signals; motivated by encoder-level numeracy weaknesses documented in [7].

7.4 Golden Insight Design

For each scenario, a set of golden insights was written manually before any strategy was executed. These are short, verifiable statements that serve as the evaluation reference for information coverage: a good summary should express the meaning of each insight even if the exact wording differs.

Construction process. The content of each insight was determined by inspecting the scenario data through the lens of the base task instruction. Specifically, the same four analytical themes used to decompose the task instruction in Table 5, overall sentiment, praise, complaints, and specific signals, were used as a structured search guide: for each scenario, the data was examined to find what evidence, if any, existed for each theme. This means the golden insights are not arbitrary observations but represent the answers a good summary should provide when asked the base task question against that specific scenario. It also explains why the number of insights varies across scenarios. The number of insights therefore differs between scenarios because the amount of information in the data differs. Forcing a fixed count would have meant padding sparse scenarios with weak insights or dropping real signals from dense ones. Where a theme had no supporting evidence a refusal-style insight was added to capture the correct behavior instead of having no insight. For example, Scenario 3 (Missing Text) contains no review text at all, so no praise or complaint insights were included. Instead, the insight "There are no textual reviews available to determine specific customer praise or complaints" was added, testing whether the model acknowledges the absence of evidence rather than fabricating themes [12].

Wording and fairness. Wording was refined with LLM assistance to apply quantifier language consistently across scenarios. "Multiple customers" requires support from at least three independent reviews, "a couple" requires two, and "a single customer" requires exactly one. The factual content itself was not generated by the LLM. Two further choices kept the evaluation fair across strategies. First, insights use qualitative rather than numeric phrasing ("predominantly positive" rather than "42 of 50 reviews"), so Aggregated-Statistics Grounding does not gain an advantage by matching explicit numbers from its statistics block. Second, where data could not support the required theme, a refusal-style insight was included to test whether the model invents patterns when evidence is absent [12].

Design rules. Five rules were applied consistently to make coverage evaluation objective and reproducible.

- **Focused decomposition.** Each insight expresses one independent, verifiable claim to prevent partial credit for compound sentences.

  > *"A single customer describes a baked potato as watery and almost rotten, requiring it to be returned."*

- **Strict quantifier thresholds.** Wording is tied to counts in the full scenario data: *"a single"* means $n = 1$, *"a couple"* means $n = 2$, and *"multiple"* requires more than two independent reviews.

  > *"Multiple customers specifically praise the baked potatoes."* vs *"A single customer describes the restaurant as messy."*

- **Hybrid data integration.** Insights combine textual and numerical evidence. When written text is absent, insights check whether the model reports the absence rather than fabricating opinions from ratings.

  "None of the reviews contain written text; all feedback consists entirely of numerical star ratings."

- **Minority-signal preservation.** Some insights target evidence from only one or two reviews that carries high business relevance, ensuring coverage rewards low-frequency but important signals.

  "A single customer reports a serious food hygiene concern involving staff handling food with bare hands."

- **Grounded refusal and ambiguity preservation.** When evidence is limited or conflicting, insights test whether the model avoids unsupported conclusions and preserves uncertainty.

  "Due to the direct contradiction between written feedback and numerical ratings, it is impossible to determine a single, consistent overall customer sentiment."

In total, 114 golden insights were written across all 11 scenarios, ranging from 7 in Scenario 10 (Data Scarcity), constrained by 5 available reviews, to 15 in Scenario 6 (High Information Density), where every synthetic review introduces a distinct non-redundant signal. Table 7 shows representative examples drawn from five scenarios to illustrate how the same design rules produce qualitatively different insights depending on the data condition.

Table 7: Representative golden insight examples across five scenarios. Each example shows the design rule it primarily exercises.

Scenario	**Rule**	**Golden Insight**
Scenario 2 (Needle in a Haystack)	Minority signal	*"A single customer reports a serious food hygiene concern involving staff handling food with bare hands, which stands out as a critical negative signal within otherwise overwhelmingly positive feedback."*
Scenario 3 (Missing Text)	Hybrid data	*"None of the reviews contain written text; all feedback consists entirely of numerical star ratings, meaning there is no textual evidence available to determine specific customer praise or complaints."*
Scenario 5 (Temporal Shift)	Atomic decomposition	*"Customer sentiment shows a clear and significant shift over time; earlier reviews are overwhelmingly positive while the most recent reviews are predominantly very negative."*
Scenario 10 (Data Scarcity)	Grounded refusal	*"Due to the availability of only 5 reviews, there is a complete absence of reliable recurring patterns or consistent trends regarding the restaurant's food, service, or atmosphere."*
Scenario 11 (Text-Rating Conflict)	Ambiguity preservation	*"Due to the direct contradiction between the written feedback and the numerical ratings, it is impossible to determine a single, consistent overall customer sentiment."*

7.5 Model Selection and Configuration

With the dataset, task instruction, and evaluation reference material now defined, the following subsections turn to the generation and judging infrastructure used

to produce and score summaries. Two OpenAI GPT-family models were used: `gpt-5.4-mini` as the generator and `gpt-5.4` as the judge. Using GPT-family models was recommended by Zoined Oy, whose platform is built on the OpenAI API, and is scientifically well-grounded given that GPT-family models are the most widely evaluated in the factual accuracy and LLM-as-a-Judge literature [9, 18, 4, 5].

Keeping the roles separate reduces the risk of self-enhancement bias, where the judge assigns higher scores to outputs that match its own generation style [9]. It also allows the more capable and more expensive model to be reserved for the role where reliability matters most.

`gpt-5.4-mini` **(generator).** Selected to balance model capability with computational cost and speed. It had to run 132 times across scenarios, strategies, and repeated runs, so cost and speed mattered. Its 400,000-token context window [25] is sufficient even for Scenario 7 (Maximum Volume, 500 reviews plus statistics). Baumgärtel et al. find that medium-sized models often match or outperform larger ones on summarization tasks [26].

`gpt-5.4` **(judge).** Selected for its stronger instruction-following and structured output capabilities. Evaluation requires the judge to verify atomic claims, compare summaries against golden insights, and return structured JSON consistently across hundreds of calls. Zheng et al. find that GPT-4-class frontier judges reach 85% agreement with expert human evaluators, exceeding the human-human agreement of 81% on the same benchmark [9]. The 1,050,000-token context window [27] accommodates even Scenario 7 (Maximum Volume) evaluation calls that must hold the generated summary and full source data simultaneously.

Shared settings. Both models were configured with `temperature=0`, `seed=42`, and `reasoning_effort=none`. The rationale for these settings is given in Section 5.1.

7.6 Prompt Construction Per Strategy

The experimental pipeline uses prompts at three levels: the shared base task instruction (identical for all strategies, defined in Section 7.2), grounding-specific evidence rendering (the only component that varies), and evaluation prompts used by the judge model. Figure 4 illustrates this structure.

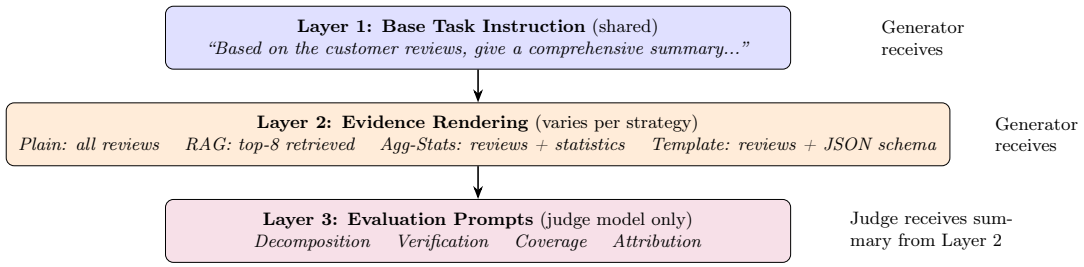

Figure 4: Three-layer prompt architecture. Layer 1 is identical for all strategies. Layer 2 is the only component that varies. Layer 3 is used only by the judge during evaluation.

Shared system-level framing

All four strategies share a system-level role definition that constrains the model to source-only generation:

> *You are an assistant summarizing customer reviews. Do not invent information. Answer only using the information provided. Do not use prior knowledge.*

Song et al. [12] demonstrate that LLMs inherently tend to fall back on parametric knowledge in retrieval-augmented settings, which motivates these strict system-level boundaries. The same shared prompt then attaches three XML contract blocks. `<output_contract>` specifies what the output must contain, `<grounding_rules>` prohibits the use of prior knowledge and forbids unsupported claims, and `<completeness_contract>` prevents selective omission of inconvenient signals. The XML tags are used only as a prompt-structuring convention for the OpenAI API; the instructions inside the tags are plain-English constraints and are therefore not specific to OpenAI models. The grounding-rules block also serves as the mechanism for grounded refusal, directly tested in Scenario 3 (Missing Text), where no review text is available. The shared generation constraints are summarized above, while their exact wording is reproduced in Appendix D which contains the evaluation prompts, the Template-Constrained output schema, the RAG keyword-extraction prompt, and the shared generation contract blocks (Appendix D.8).

Citation rule design

Every generation prompt requires each factual claim to be followed inline by its supporting evidence in brackets, using `[REVIEW: X]` for review citations or `[STAT: stat_name]` for statistics citations. Combining citations at the end of a paragraph is explicitly prohibited because it would prevent attribution evaluation from determining which source supports which specific claim. The inline format allows each claim-citation pair to be checked independently [5].

Plain Prompting. Plain Prompting uses the shared system framing, the three contract blocks, and the inline citation rule described above. The user message provides each review record structured with its ID, overall star rating, detailed aspect ratings, and review text, followed by the base task instruction. The prompt construction for this strategy is described in this section; Appendix B.2 provides a complete example of the review records for Scenario 11 (Text-Rating Conflict), and Appendix D contains the evaluation prompts used to score the generated summaries. A sample output from Scenario 10 (Data Scarcity) illustrates the citation rule in action:

> *"Customers give mixed feedback overall, with some praising the food quality and freshness while others complain about poor service, inconsistent item availability, and disappointing food quality [REVIEW: 1, REVIEW: 2, REVIEW: 3, REVIEW: 5]...."*

Every claim is followed immediately by the supporting review IDs, rather than a single citation block at the end of the paragraph.

Aggregated-Statistics Grounding. The system framing and contract blocks are identical to Plain Prompting. The key difference is a pre-computed statistics block prepended to the user message before the review records. It includes the overall

rating statistics (mean, median, standard deviation, per-star counts), aspect-level averages for food, service, and atmosphere, and review volume counts. This block is computed deterministically from the raw data before any experiment runs. Supplying it explicitly addresses the weakness identified by White, where summarization models struggle to infer aggregate numerical values reliably from raw text [7].

The citation rule is extended to support three citation types: review citations in `[REVIEW: X]` format, statistics citations in `[STAT: stat_name]` format, and hybrid claims that cite both. This allows statistical and qualitative evidence to be traced back to their respective sources independently. The prompt construction and statistics-block contents are described in this section; Appendix B provides a complete example of the aggregated statistics block for Scenario 11 (Text-Rating Conflict). A representative output extract from Scenario 10 (Data Scarcity) shows the hybrid citation system in use:

> *"Customers give this restaurant mixed but polarized feedback, with an average rating of 3.0/5 and a median of 4.0/5 [STAT: average_rating][STAT: median_rating]. The ratings span the full range from 1 to 5 [STAT: rating_range], and the distribution shows two 1-star reviews, two 4-star reviews, and one 5-star review [STAT: rating_distribution]..."*

Statistical claims cite the statistics block and qualitative claims cite specific review IDs.

RAG. RAG replaces the full review set from Plain Prompting with a retrieved subset: a retriever selects the reviews most relevant to the task instruction, and only those are passed to the model. The user message contains only the retrieved reviews, not the full set, using the same system framing and citation rules as Plain Prompting. Algorithm 1 summarizes the retrieval procedure that selects this subset.

Algorithm 1 RAG candidate retrieval

Require: task instruction; the scenario's reviews; retrieval depth $k = 8$
Ensure: the k reviews most relevant to the task instruction
1: textReviews ← the reviews that contain review text
2: **if** textReviews is empty **then**
3: **return** an empty list ▷ no review text to search; RAG inapplicable (Scenario 3 (Missing Text))
4: **end if**
5: build a BM25 search index over textReviews
6: topics ← QUERYEXPAND(task instruction) ▷ judge call; at most 8 topics, at most 6 keywords each
7: **for** each topic in topics **do**
8: score every review in textReviews against that topic's keywords using BM25
9: **end for**
10: keep, for each review, its highest score across all topics ▷ aggregate across topics
11: rank textReviews by that score, highest first
12: **return** the top k reviews
Prompt: QUERYEXPAND D.7.

The query-expansion step allows up to max 8 topics and 6 keywords per topic, providing a bounded expansion of the task instruction into concrete review-style vocabulary. This expansion helps lexical retrieval bridge the vocabulary gap between

a user query and relevant documents, a method proven to significantly improve BM25 retrieval performance [24]. For example, in Scenario 10 (Data Scarcity), a general topic such as "service" is expanded into concrete review-style keywords including "service," "waiter," "staff," and "attentive."

Retrieval uses `rank_bm25`'s BM25Okapi implementation with default parameters ($k_1 = 1.5$, $b = 0.75$), applied at the level of whole reviews. For Scenario 10 (Data Scarcity, 5 reviews total), RAG retrieved all 4 text-bearing reviews since the pool was smaller than $k = 8$, illustrating the top-k step's behavior when fewer than k text-bearing reviews are available. The summary output is structurally identical to the Plain Prompting output. Figure 5 gives a visual overview of this pipeline.

The reported RAG latency and token counts measure the summary generation step only. The query-expansion call is reported separately because it is a fixed setup step that consumed approximately 500 tokens and added around 2 seconds in this implementation. This means the operational-cost results should be interpreted as generation-stage costs rather than full end-to-end RAG pipeline costs. Including query expansion would reduce, but not remove, RAG's cost advantage.

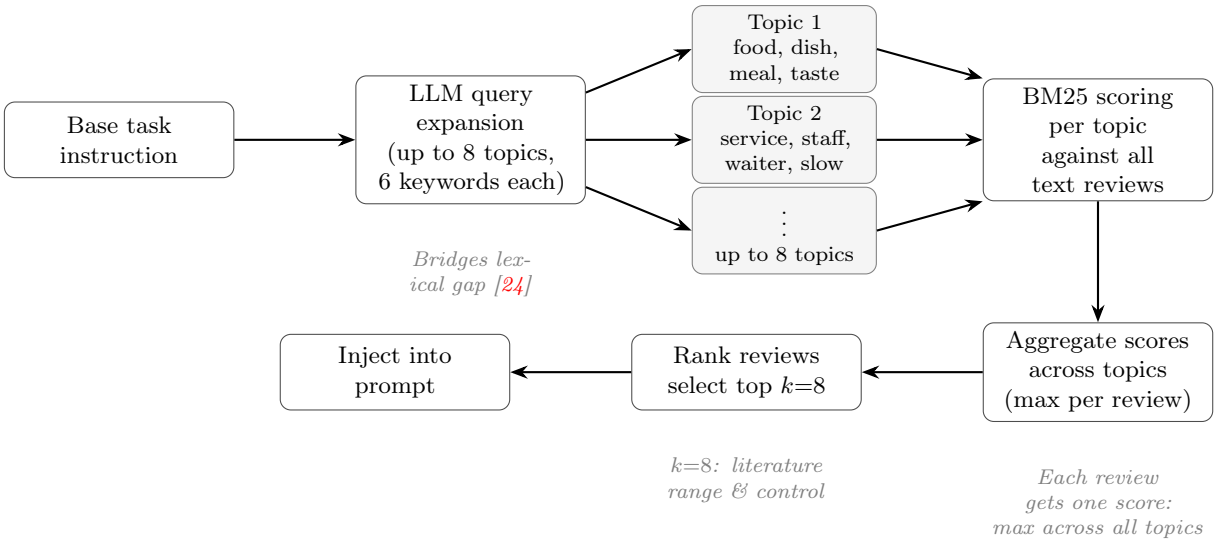

Figure 5: RAG retrieval pipeline. The base task instruction is expanded into up to 8 keyword topics. BM25 scores are computed against all reviews with text, and the top 8 reviews are injected into the prompt. Query-expansion latency is excluded from the reported RAG latency because it is constant across scenarios.

Template-Constrained Generation. Two structural additions distinguish this prompt from the others. First, the system message replaces the prose citation rule with a structured output contract that requires the model to output only valid JSON, include no explanatory text, and match the predefined schema exactly. The grounding rules remain identical to the other strategies. Second, a categorization rule instructs the model to place any valid insight that does not fit the four main schema categories into an `other_insights` field rather than discarding it. Without this rule, the model tends to force every insight into the four schema fields, losing signals that are genuinely cross-cutting or hard to classify. The strategy-specific prompt construction is described in this section, and the Template-Constrained output schema is reproduced in Appendix D.

The schema fields correspond directly to the four evaluation themes of the base task instruction. The `specific_signal` field maps to the instruction segment about "specific or recurring patterns worth highlighting."

An extract from the Scenario 10 (Data Scarcity) output illustrates this citation-padding risk:

```
{
  "overall_sentiment": {"claims": [{"statement":
    "Customer sentiment is mixed, with one very positive
     review about fresh, good food and several negative
     reviews describing poor experiences.",
    "sentiment": "Mixed", "review_ids": [1, 3, 5]}]},
  "specific_signal": {"claims": [{"statement":
    "One customer said their child only once got their
     favorite baguette despite many prior visits.",
    "review_ids": [1]}]},
  ...
}
```

The `review_ids` field in the `overall_sentiment` claim cites three reviews, but the claim about mixed sentiment is grounded across all five. The model must select a subset, and under the schema pressure to always populate the field, it may include reviews that provide only weak support.

7.7 Automated Evaluation Framework

Once a summary is generated, the judge evaluates it across three dimensions: factual accuracy, information coverage, and attribution faithfulness, as defined in Section 2.3. Operational cost is logged by system telemetry and requires no judge involvement. Algorithm 2 shows how one complete run proceeds, from prompt rendering through to scoring.

Algorithm 2 Single experiment run

Require: scenario data; grounding strategy; generator model; judge model
Ensure: generated summary with factual accuracy, coverage, and attribution scores

1: render the prompt for the chosen strategy ▷ RAG selects reviews first, Algorithm 1
2: summary ← call the generator once with that prompt
3: evalText, claims ← summary prepared for evaluation ▷ Template-Constrained reads its own JSON fields; the other three strategies parse citation tags from the raw text
4: facts, supportedFacts ← FactualAccuracy(evalText, source data) ▷ Algorithm 3
5: goldenInsights, coveredInsights ← Coverage(summary, scenario's golden insights) ▷ Algorithm 3
6: citedClaims ← Attribution(claims, source data) ▷ Algorithm 3
7: log latency and token usage
8: **return** summary, with all computed scores

Steps 4–6 above call the three evaluation procedures detailed in Algorithm 3.

All three procedures share the same judge model and are given the same source data rendered from the scenario; they differ only in what they check. Figure 2 in Section 5.4 shows how the pipeline connects. Full evaluation prompts are provided in Appendix D.

Algorithm 3 Summary evaluation pipeline

procedure FACTUALACCURACY(summary, sourceData)
    facts, discardedClaims ← DECOMPOSE(summary) ▷ judge call; discards non-verifiable claims (e.g. filler text)
    for each fact in facts **do**
        label(fact) ← VERIFY(fact, sourceData) ▷ judge call; standard (pattern vs. single-review) chosen from the fact's own wording
    end for
    supportedFacts ← facts labeled Supported
    return facts, supportedFacts ▷ used for FActScore, Section 2.3
end procedure
procedure COVERAGE(summary, goldenInsights)
    for each insight in goldenInsights **do**
        label(insight) ← JUDGECOVERAGE(summary, insight) ▷ one independent judge call per insight
    end for
    coveredInsights ← insights labeled Covered
    return goldenInsights, coveredInsights ▷ used for Coverage, Section 2.3
end procedure
procedure ATTRIBUTION(claims, sourceData) ▷ claims already extracted from the summary, see Algorithm 2
    for each claim in claims **do**
        validSources(claim) ← JUDGECITATIONS(claim, sourceData) ▷ judge checks only the sources already cited
    end for
    return claims, with the sources each one cited and which of those were valid ▷ used for CP, CR, FSCR, Section 2.3
end procedure

Prompts: DECOMPOSE D.2, VERIFY D.3, JUDGECOVERAGE D.4, JUDGECITATIONS D.5.

All evaluation prompts ask the judge to reason with defined rules before giving a verdict and return a structured JSON object. This follows the recommendation in prior LLM-as-a-Judge work that structured reasoning improves consistency and reduces arbitrary scoring [18, 9].

Some scenarios were designed with absent or severely limited textual evidence. The evaluation checks explicitly whether the model acknowledges this limit rather than generating unsupported content, a failure mode documented in prior work [12] and directly tested in Scenario 3 (Missing Text).

Factual Accuracy. The FACTUALACCURACY procedure in Algorithm 3 implements the metric defined in Section 2.3. Decomposition breaks the summary into small, independent atomic facts. A few rules govern this step:

- **Separate independent predicates.** "The food was cold and the wait was 40 minutes" becomes two facts, since each can be checked on its own.
- **Preserve relational claims.** "Recent reviews were more negative than earlier reviews" stays as one fact. Splitting it would destroy the comparison being made.
- **No duplicates.** If two potential facts say the same thing, only the more specific one is kept. This prevents the same claim from inflating the denominator.
- **No conversational filler.** Phrases like "Here is a summary of the reviews" make no factual claim and are excluded from scoring entirely.

Two design decisions shape verification, both applied within the single judge call in Algorithm 3.

First, **semantic entailment** is used rather than keyword matching. A claim that "the staff was overwhelmed" is supported by a review saying "waiters were running around in a panic and forgot our drinks", even though no words overlap [11, 17]. The verification prompt includes a calibration example of this kind to keep the judge's threshold consistent.

Second, a **dual evidentiary standard** applies based on how the claim is worded:

- **Pattern claims** use language like "multiple customers", "consistently", or "recurring". These require evidence from more than one review.
- **Standard claims** require only one supporting source. Disagreement from other reviews does not invalidate them, since reviewer disagreement is a normal feature of opinion data.

The claim's own wording determines which standard applies, not the judge's interpretation.

Information Coverage. The COVERAGE procedure in Algorithm 3 implements the metric defined in Section 2.3, checking each golden insight against the generated summary independently. Checking all insights together risks the judge confusing evidence for one insight with evidence for another, particularly when insights are thematically similar; independent calls reduce this cross-contamination risk.

Semantic entailment applies here too: the core meaning of the insight must be present, but the exact wording does not need to match. An insight is counted as covered when its main meaning is represented in the summary, even if some minor detail is omitted. To keep the judge's threshold consistent across runs, the coverage prompt includes calibration examples for covered and not-covered cases [18]. Details in Appendix D.4.

Attribution Faithfulness. The three sub-metrics, Citation Precision, Citation Recall, and Fully Supported Claims Ratio, are defined in Section 2.3 following [5].

As Algorithm 3's ATTRIBUTION procedure shows, the judge is never allowed to search for supporting sources beyond those the model already cited. The goal is to measure citation discipline: did the model cite valid evidence, not just any evidence?

A **strictness rule** governs what counts as a valid citation. A source is marked invalid if it mentions only the general topic without entailing the specific claim. For example, a review saying "the food was bad" is not a valid citation for "the soup was cold", because the specific fact is not entailed by the general complaint.

7.8 Manual Audit Protocol

The automated pipeline produces hundreds of judge verdicts across all scenarios, strategies, and runs. A sample of verdicts was manually audited as a construct-validity check [9].

Thirty verdicts were audited in total: 10 drawn from factual-verification tasks, 10 from coverage-evaluation tasks, and 10 from attribution-evaluation tasks. Within each task type, the sample was selected randomly while ensuring coverage across scenarios and grounding strategies where possible. This made the audit closer to a stratified random check than a purely random sample.

The audit was carried out by the author. For each sampled verdict, the auditor received the same inputs as the judge (the source reviews, the claim or golden insight being evaluated, and the cited review IDs for attribution tasks) and made an independent binary judgment before seeing the judge's verdict, to avoid anchoring. Cases where the auditor and judge disagreed were recorded and examined to determine whether the disagreement reflected a judge reasoning error or a prompt design limitation. The main findings from this audit are reported in Section 8.6. Because the author both wrote the golden insights and performed the audit, the audit is a construct-validity check on the judge rather than an independent replication of the evaluation. A multi-annotator study would provide stronger validation, but that was outside the scope of this experiment.

8 Results

This section reports the empirical results for the three research questions. The four grounding strategies were evaluated across 11 scenarios with three repeated runs each, giving 132 planned executions in total. RAG produced no valid output for Scenario 3 (Missing Text) because that scenario contained no review text. Standard retrieval-augmented frameworks are structurally limited to fetching unstructured natural-language passages, the system cannot index, match, or aggregate raw numerical attributes in the absence of textual narratives [2, 10]. The treatment of this missing RAG observation in the statistical analysis is described in Section 8.2.

The section is organized as follows. Section 8.1 gives a descriptive overview. Section 8.2 summarizes the statistical testing procedure, effect-size reporting, and handling of Scenario 3 (Missing Text). Sections 8.3, 8.4, and 8.5 report results for RQ1–3. Section 8.6 reports the manual audit findings, and Section 8.7 summarizes the main findings.

8.1 Overview of Results

Table 8 gives a first overview of all metrics. For each strategy and scenario, the three runs are averaged first. The table reports the mean and standard deviation across those scenario-level averages. For RAG, Scenario 3 (Missing Text) is omitted because no valid output was produced. Scenario-level patterns are examined in the figures throughout this section, and a full worked example for one scenario is provided in Appendix C.

The standard deviations in Table 8 should not be interpreted as uncertainty around a single population estimate. They describe variation across scenario-level averages, so large values indicate that a strategy behaves differently across scenario conditions. This is especially important for latency and token volume, where Scenario 7 (Maximum Volume) strongly increases the spread for long-context strategies.

Table 8: Summary statistics (mean ± standard deviation) by grounding strategy. Metrics are defined in Section 2.3. FA = Factual Accuracy; IC = Information Coverage; CP = Citation Precision; CR = Citation Recall; FSCR = Fully Supported Claims Ratio. Latency in seconds; Tokens = prompt plus completion tokens. RAG excludes Scenario 3 (Missing Text) (no valid output). RAG latency and token counts exclude the query-expansion setup call; this overhead is reported separately in Section 7.6.

Strategy	FA	IC	CP	CR	FSCR	Lat. (s)	Tokens
Plain Prompting	$0.899_{\pm 0.063}$	$0.847_{\pm 0.139}$	$0.869_{\pm 0.069}$	$0.984_{\pm 0.046}$	$0.594_{\pm 0.251}$	$6.32_{\pm 3.33}$	$8{,}246_{\pm 10{,}937}$
RAG	$0.845_{\pm 0.088}$	$0.514_{\pm 0.177}$	$0.940_{\pm 0.041}$	$1.000_{\pm 0.000}$	$0.859_{\pm 0.097}$	$3.26_{\pm 0.52}$	$1{,}464_{\pm 407}$
Template-Constrained	$0.924_{\pm 0.033}$	$0.907_{\pm 0.096}$	$0.790_{\pm 0.095}$	$0.952_{\pm 0.066}$	$0.493_{\pm 0.261}$	$8.36_{\pm 3.48}$	$8{,}987_{\pm 10{,}927}$
Aggregated-Statistics	$0.924_{\pm 0.039}$	$0.841_{\pm 0.124}$	$0.875_{\pm 0.058}$	$0.962_{\pm 0.060}$	$0.633_{\pm 0.235}$	$7.38_{\pm 4.22}$	$8{,}807_{\pm 11{,}089}$

Four patterns emerge from this table.

- First, Plain Prompting, Template-Constrained Generation, and Aggregated-Statistics Grounding cluster closely on Factual Accuracy (0.899–0.924), while RAG is lower (0.845).
- Second, Information Coverage separates the strategies more clearly: RAG reaches only 0.514 while the others are between 0.841 and 0.907.

- Third, attribution quality follows the opposite pattern: RAG has the strongest Citation Precision (0.940) and Fully Supported Claims Ratio (0.859), while Template-Constrained Generation has the weakest on both (0.790 and 0.493).
- Fourth, at the generation stage, RAG uses far fewer tokens (1,464) and completes faster (3.26 s) than the long-context strategies.

The main result is a trade-off rather than a single winner. These patterns are examined metric by metric in the sections that follow. The architectural explanations for why they arise are discussed in Section 9.

8.2 Statistical Testing Protocol

The inferential analysis follows the statistical plan defined in Section 5.5. Here, only the results-specific implementation details are summarized.

Effect sizes. Effect sizes are reported alongside p-values to indicate the magnitude of the observed differences. For the Friedman test, Kendall's W is reported as a measure of how consistently the strategies are ranked across scenarios. For pairwise Wilcoxon signed-rank tests, the rank-biserial correlation r is reported. Its absolute value indicates the strength of the pairwise difference, while its sign indicates the direction of the comparison. These values are used descriptively, and statistical significance is determined by the Holm-corrected p-values.

Handling of Scenario 3 (Missing Text). RAG produced no valid output for Scenario 3 (Missing Text) because all 50 reviews had null text fields. The problem was not that the language model could not reason about ratings, but that the RAG pipeline first had to retrieve supporting evidence. Since there was no review text, the BM25 retriever had nothing to index or select.

This reflects a broader limitation of retrieval-based grounding. RAG is effective when relevant evidence appears in retrievable text, but it can struggle when the important signal is mainly numerical or structured. Retrievers match words or semantic similarity. They do not directly model numerical order, quantities, or dataset-level aggregates in the way a statistical calculation does. In this thesis, Scenario 3 (Missing Text) therefore tests a real weakness of the implemented text-retrieval pipeline: when written review evidence is absent, RAG cannot reliably ground a summary using the numerical signals alone.

Two analyses were computed:

- **Analysis A** excludes Scenario 3 (Missing Text) and compares all four strategies on 10 scenarios. This is the primary analysis.
- **Analysis B** includes Scenario 3 (Missing Text) and imputes RAG's missing values as zero. This is used only as a robustness check.

Unless stated otherwise, all inferential claims in this section refer to Analysis A. The conclusions are the same under both analyses, and no pairwise strategy comparison reverses direction. The missing RAG output should therefore be read as a limitation of the implemented fixed BM25 RAG pipeline, not as an inherent limitation of all RAG systems. In practice, a production system could add a fallback path for ratings-only

cases, such as returning a grounded refusal or switching to an aggregated-statistics strategy.

8.3 RQ1: Factual Accuracy and Information Coverage

RQ1 asks how the four strategies differ in factual accuracy and information coverage. These two metrics are defined in Section 2.3. This section focuses on what was observed.

8.3.1 Factual Accuracy

Results. The three long-context strategies cluster closely: Aggregated-Statistics Grounding and Template-Constrained Generation both reach 0.924, while Plain Prompting reaches 0.899. RAG is lower at 0.845. Figure 6 shows the aggregate picture and Figure 7 shows the scenario-level variation.

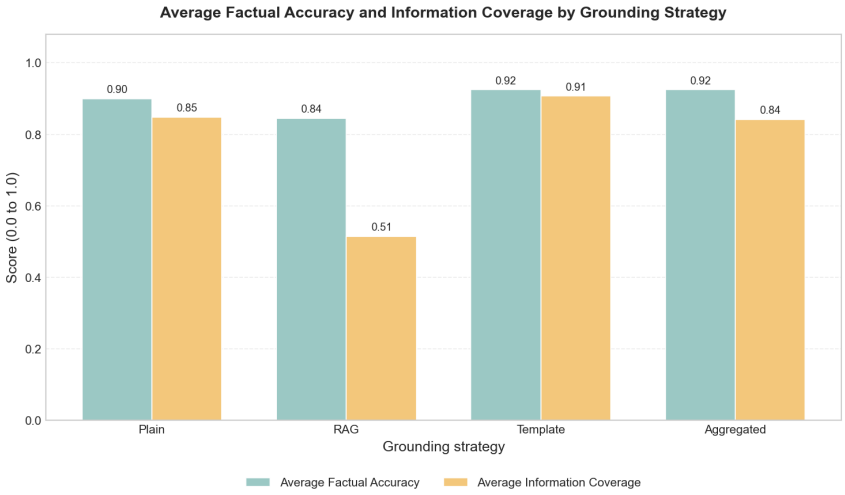

Figure 6: Average Factual Accuracy and Information Coverage by grounding strategy. The three long-context strategies cluster closely on Factual Accuracy. The largest relative gap appears in Information Coverage: RAG scores 0.51, while the next-lowest strategy scores 0.84. This means RAG's coverage is about 39% lower than the next-lowest strategy.

The Friedman test finds a statistically significant difference across strategies (Analysis A: $\chi^2(3) = 8.28$, $p = 0.041$, $W = 0.28$). Pairwise tests show that the significant differences are limited to comparisons against RAG. The full Holm-corrected results are summarized in Table 9. The factual accuracy advantage of Template-Constrained Generation and Aggregated-Statistics Grounding is therefore statistically supported only against RAG, and it is narrow.

Figure 7 shows that RAG's factual-accuracy gap is not constant across scenarios. It is closer to the other strategies in Scenario 2 (Needle in a Haystack), Scenario 8 (Ambiguous Opinions), and Scenario 10 (Data Scarcity), but drops more clearly in Scenario 1 (Chronological Sample) and Scenario 11 (Text-Rating Conflict). Figure 8 complements this scenario-level view by showing the distribution of scores across all runs, where RAG has the lowest median and the widest spread. Per-scenario values are given in Appendix E.

Figure 9 provides useful context. Template-Constrained Generation and Aggregated-Statistics Grounding produce roughly 62–66 atomic facts per run on

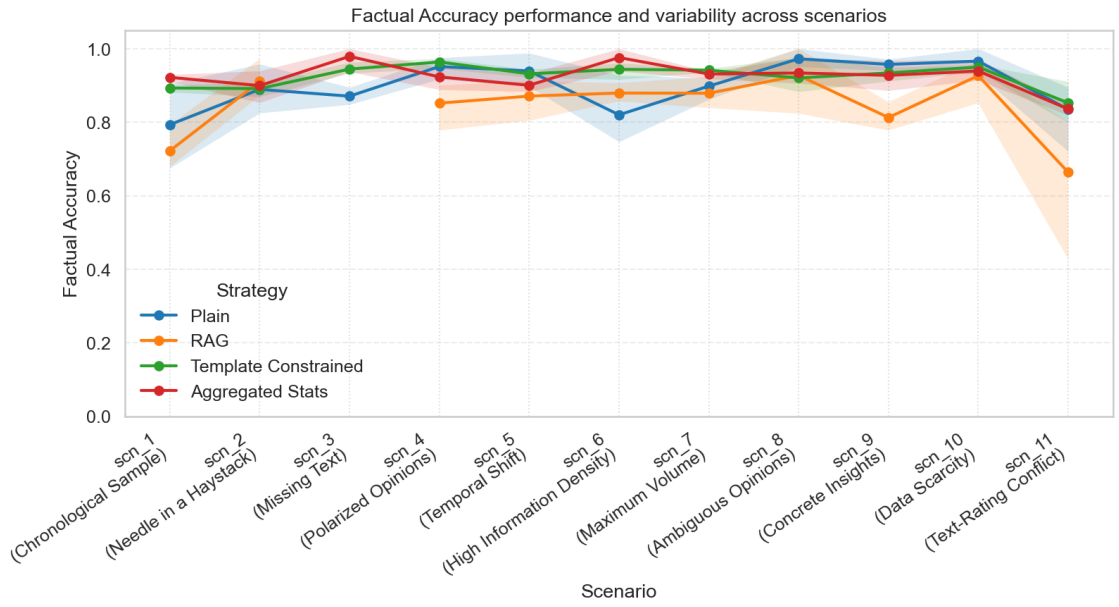

Figure 7: Factual Accuracy across scenarios by grounding strategy. Shaded bands show one standard deviation across three runs. RAG shows greater variability, with its weakest valid outputs in Scenario 1 (Chronological Sample), at 0.72, and Scenario 11 (Text-Rating Conflict), at 0.66. Scenario 3 (Missing Text) has no valid RAG output.

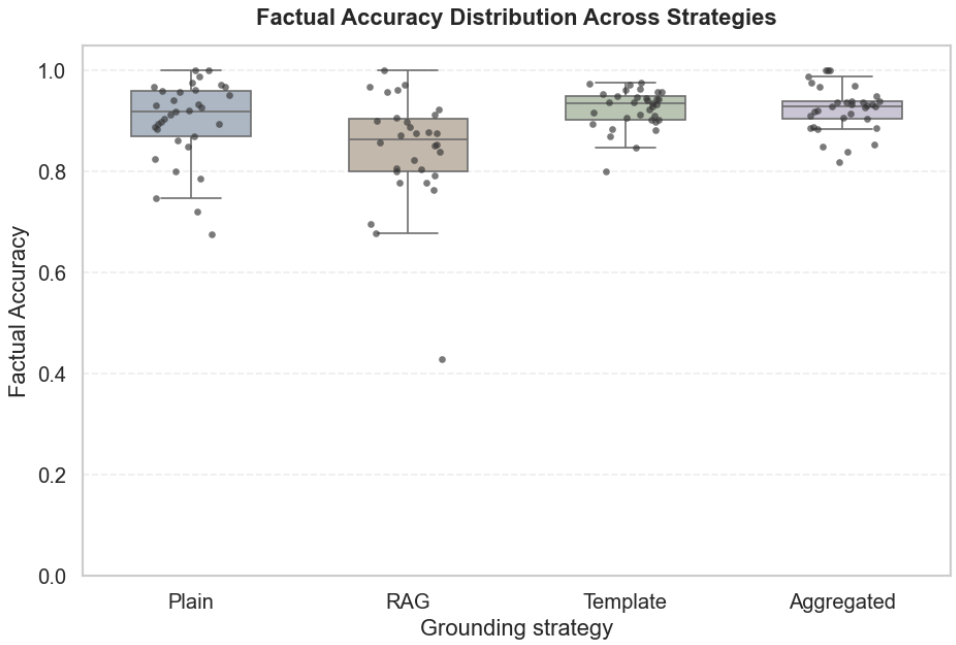

Figure 8: Distribution of Factual Accuracy scores across all scenarios and runs. RAG has the lowest median and the widest spread. Template-Constrained Generation and Aggregated-Statistics Grounding show the narrowest interquartile ranges.

average, compared with 56 for Plain Prompting and 36 for RAG. Generating more facts creates more opportunities for unsupported claims. Template-Constrained Generation and Aggregated-Statistics Grounding produce the most facts while maintaining accuracy on par with or above Plain Prompting. RAG generates the fewest facts and still achieves the lowest mean accuracy.

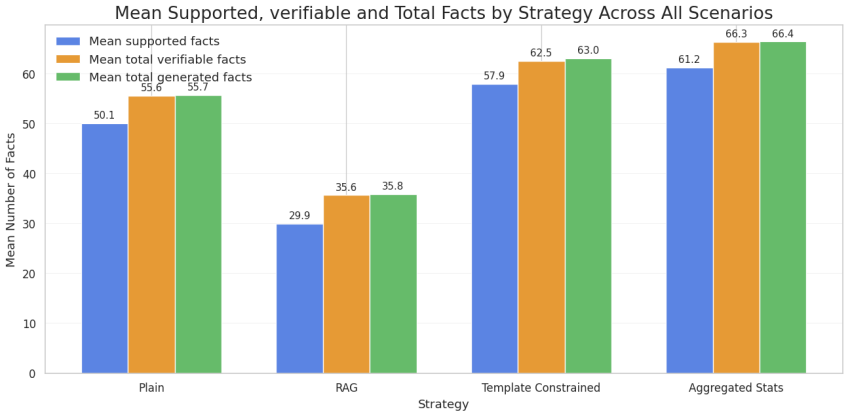

Figure 9: Mean number of atomic facts per run by strategy. Template-Constrained Generation and Aggregated-Statistics Grounding produce the most facts while maintaining similar or better accuracy than Plain Prompting. RAG generates the fewest facts and has the lowest ratio of supported to total facts.

Interpretation. The statistically significant factual accuracy gap between RAG and both Template-Constrained Generation and Aggregated-Statistics Grounding, but not between Plain Prompting and either of those strategies, is consistent with the structural difference between these approaches. Why this pattern occurs at an architectural level is discussed in Section 9.1.

8.3.2 Information Coverage

Results. Information Coverage separates the strategies more strongly than Factual Accuracy. RAG reaches 0.514 on average, while the next-lowest strategy, Aggregated-Statistics Grounding, reaches 0.841: a gap of approximately 33 percentage points. Figure 10 shows the scenario-level pattern and Figure 11 the distribution across all runs.

The Friedman test is highly significant (Analysis A: $\chi^2(3) = 17.97$, $p < 0.001$, $W = 0.60$). Pairwise tests show that all three long-context strategies significantly outperform RAG, while they do not significantly differ from each other (Table 9). Template-Constrained Generation has the highest average Information Coverage score (0.907). However, the supported conclusion is that all long-context strategies cover substantially more information than RAG.

The scenario-level view (Appendix E) shows that the coverage gap is sharpest where the number of reviews or distinct signals is large. In Scenario 6 (High Information Density, 50 reviews, each with a unique insight), RAG reaches only 0.20. In Scenario 7 (Maximum Volume, 500 reviews), RAG reaches only 0.27, while the long-context strategies stay around 0.85–0.91.

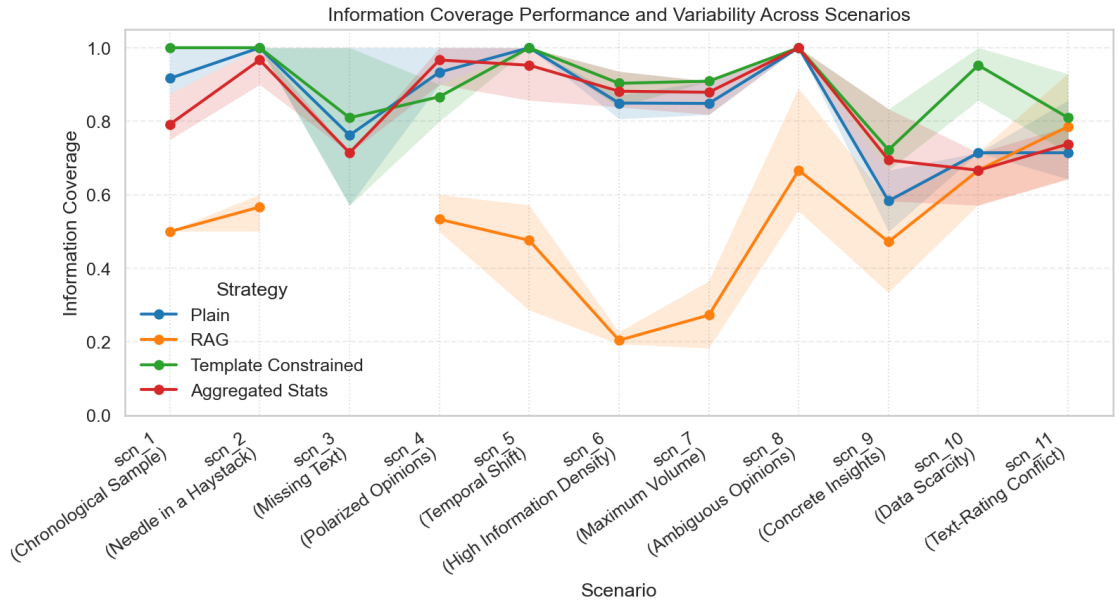

Figure 10: Information Coverage across scenarios by grounding strategy. RAG is the lowest-coverage strategy in every scenario except Scenario 11 (Text-Rating Conflict) (see Appendix C.2). The largest drops occur in Scenario 6 (High Information Density), at 0.20, and Scenario 7 (Maximum Volume), at 0.27. Scenario 3 (Missing Text) has no valid RAG output.

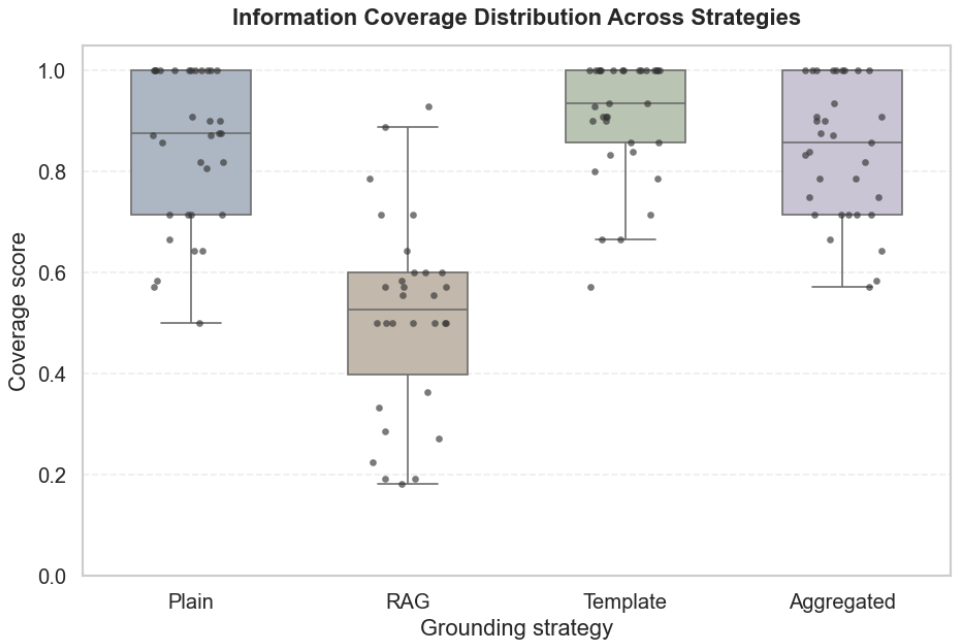

Figure 11: Distribution of Information Coverage scores. RAG's median is around 0.53 with a wide lower tail. The other strategies cluster in the 0.85–0.93 range.

The run-level plot shows the same pattern more compactly: the three long-context strategies remain high and stable, while RAG shows greater run-to-run variation. Because model calls are not fully deterministic even at temperature zero (Section 5.1), variation can occur in both the LLM-based query-expansion step, potentially changing the retrieved top-8 set, and the final generation step. When only 8 of 500 reviews are available to the generator, such variation can have a larger effect on information coverage. Figure 12 shows the run-level detail for Scenario 7 (Maximum Volume).

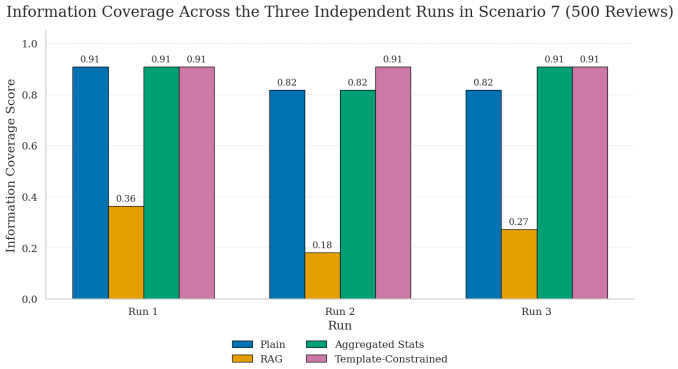

Figure 12: Information Coverage across three runs in Scenario 7 (Maximum Volume, 500 reviews). Template-Constrained Generation is stable at 0.91. RAG drops to 0.18 in run 2, showing the coverage risk of a fixed top-k retrieval setup under high review volume.

Interpretation. The coverage result answers RQ1 more clearly than the factual accuracy result. RAG is weaker mainly because it misses information, not because its stated claims are mostly wrong. The structural reason for this is that passing only 8 retrieved reviews to the generator limits what can be reported, regardless of how accurately the retrieved evidence is used. The exception is Scenario 11 (Text-Rating Conflict), where the review pool is small enough that retrieval discards little evidence (Appendix C.2). This is discussed further in Section 9.1.

8.4 RQ2: Attribution Faithfulness

RQ2 asks whether generated claims are correctly linked to their supporting evidence. The three attribution metrics (Citation Precision, Citation Recall, Fully Supported Claims Ratio) are defined in Section 2.3. The main pattern is counter-intuitive: RAG has the strongest attribution scores despite having the weakest coverage, and Template-Constrained Generation has the weakest attribution despite using the most rigid output format.

8.4.1 Citation Precision

Results. RAG achieves the highest mean Citation Precision (0.940). Template-Constrained Generation achieves the lowest (0.790). Aggregated-Statistics (0.875) and Plain Prompting (0.869) are in the middle. Figure 13 shows the scenario-level pattern, with per-scenario values in Appendix E.

The Friedman test is highly significant (Analysis A: $\chi^2(3) = 18.84$, $p < 0.001$, $W = 0.63$). Pairwise tests show that RAG has significantly higher precision than Plain Prompting and Template-Constrained Generation, while Template-Constrained

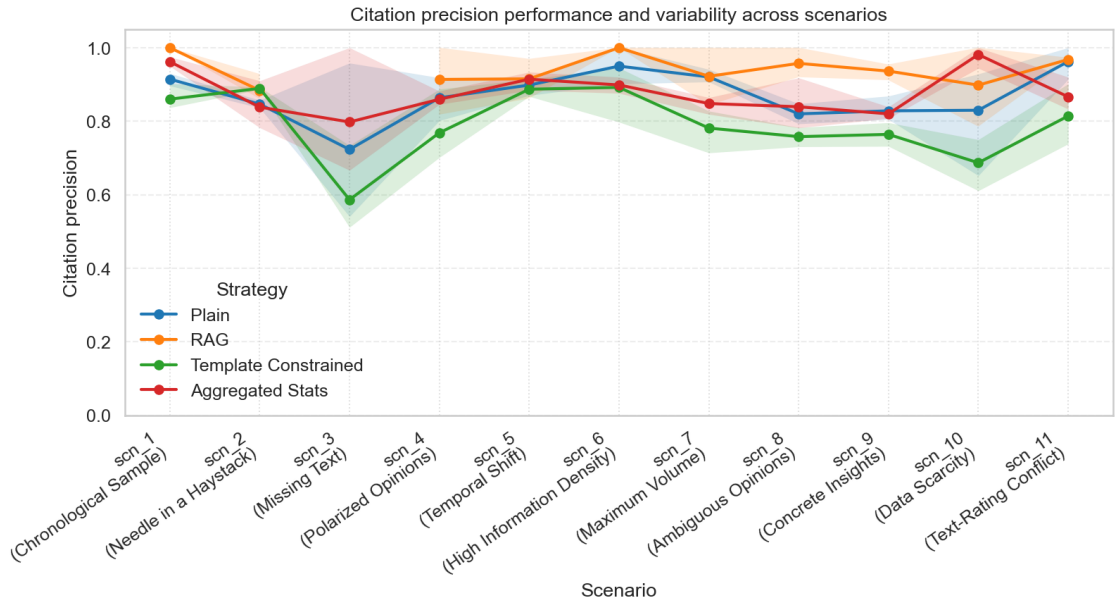

Figure 13: Citation Precision across scenarios. RAG leads in most scenarios. Template-Constrained shows the most variability, with its largest drop in Scenario 3 (Missing Text), at 0.59.

Generation is significantly lower than both Plain Prompting and Aggregated-Statistics Grounding (Table 9).

8.4.2 Citation Recall

Results. Citation Recall is uniformly high across all strategies. RAG reaches 1.000. The others are between 0.952 and 0.984. The Friedman test is significant (Analysis A: $\chi^2(3) = 13.96$, $p = 0.003$, $W = 0.47$), but no pairwise comparison survives Holm correction. This is a ceiling effect: all strategies cite at least one valid source for almost every claim, so this metric does not meaningfully discriminate between them. Figure 14 shows that values stay near the ceiling in almost every scenario; per-scenario values are in Appendix E.

Figure 15 shows why the stricter metrics matter more. Template-Constrained Generation produces the most citations on average (214.8 total, 167.7 valid per run), and the largest gap between total and valid. RAG produces the fewest (25.3 total, 23.6 valid) and the smallest gap.

8.4.3 Fully Supported Claims Ratio

Results. FSCR is the strictest attribution metric: a claim is fully supported only if every attached citation is valid. RAG achieves the highest mean FSCR (0.859), while Template-Constrained Generation has the lowest (0.493). Figure 16 shows the scenario-level breakdown, with per-scenario values in Appendix E.

FSCR produced the highest omnibus Kendall's W in the attribution group (Analysis A: $\chi^2(3) = 22.20$, $p < 0.001$, $W = 0.74$). Pairwise tests show that RAG significantly outperforms all three long-context strategies, while Template-Constrained Generation is significantly lower than Plain Prompting and Aggregated-Statistics Grounding (Table 9).

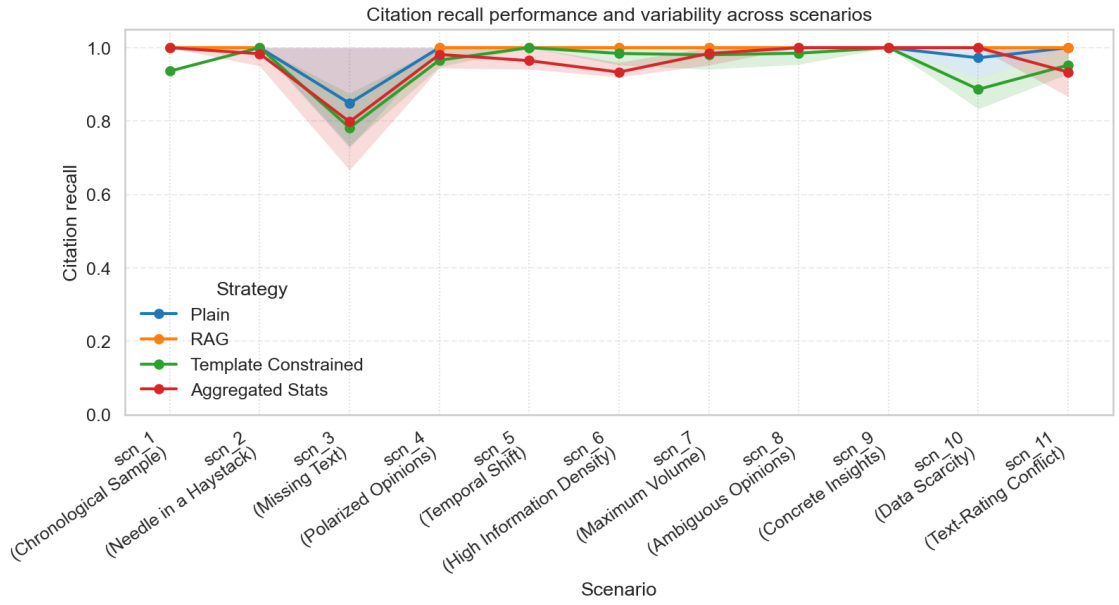

Figure 14: Citation Recall across scenarios. Most strategies stay near 1.0 in most scenarios. The main variation occurs in Scenario 3 (Missing Text), where all three strategies with valid outputs show sub-ceiling recall; RAG has no valid output for this scenario.

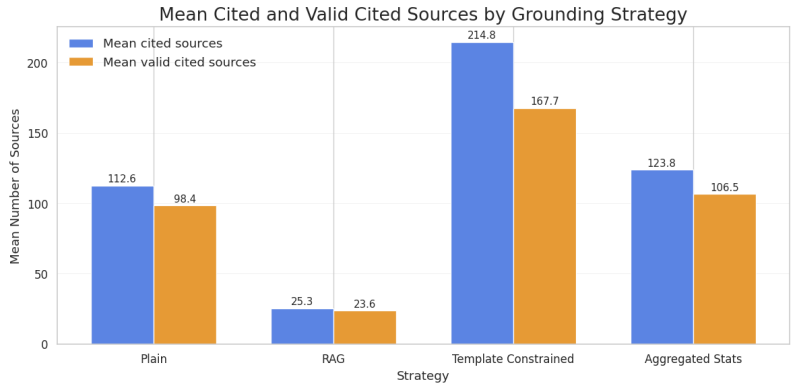

Figure 15: Mean cited and valid-cited sources per run by strategy. Template-Constrained generates the most citations and the largest gap between total and valid. RAG generates the fewest with the smallest gap.

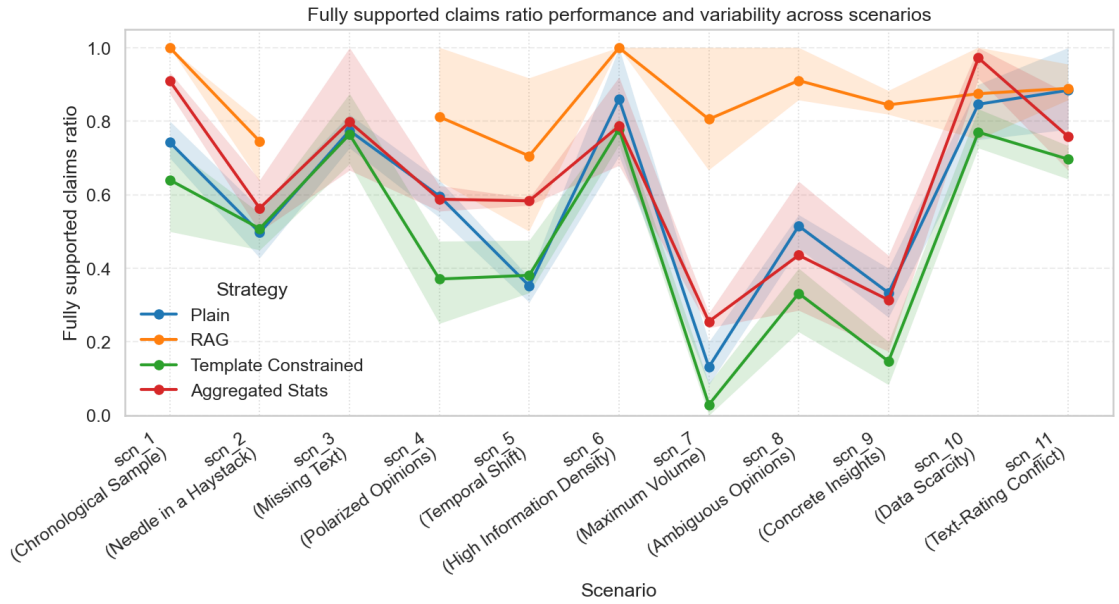

Figure 16: Fully Supported Claims Ratio across scenarios. RAG has the highest Fully Supported Claims Ratio in every scenario with a valid RAG output except Scenario 10 (Data Scarcity), where Aggregated-Statistics Grounding scores highest. RAG remains above 0.65 across all of its valid outputs. Template-Constrained drops sharply in Scenario 7 (Maximum Volume), at 0.03, and Scenario 9 (Concrete Insights), at 0.15.

Scenario 7 (Maximum Volume) shows the FSCR problem most clearly. Template-Constrained Generation produces roughly 870–980 total citations per run in this scenario, representing the range across its three runs. Its FSCR stays near zero throughout. At the same time, its factual accuracy remains high (0.943, 0.946, 0.936). The problem is not that the generated claims are mostly false. It is that many attached citations are only partially relevant.

Figure 17 places all three attribution metrics side by side.

Interpretation. The RQ2 result is that enforcing a rigid output structure does not automatically produce better attribution. The counter-intuitive pattern – where RAG leads on citation quality despite trailing on factual accuracy and information coverage – is explained in Section 9.1 in terms of evidence surface size.

8.5 RQ3: Inference Latency and Token Volume

RQ3 concerns operational cost, measured through inference latency and total token volume. The latency values are summary-generation latencies only. The excluded RAG query-expansion overhead is discussed in Section 7.6.

8.5.1 Inference Latency

Results. The omnibus test yields one of the largest Kendall's W values in the study for latency (Analysis A: $\chi^2(3) = 27.00$, $p < 0.001$, $W = 0.90$), second only to token volume. RAG is the fastest strategy (3.26 s), followed by Plain Prompting (6.32 s), Aggregated-Statistics (7.38 s), and Template-Constrained Generation (8.36 s).

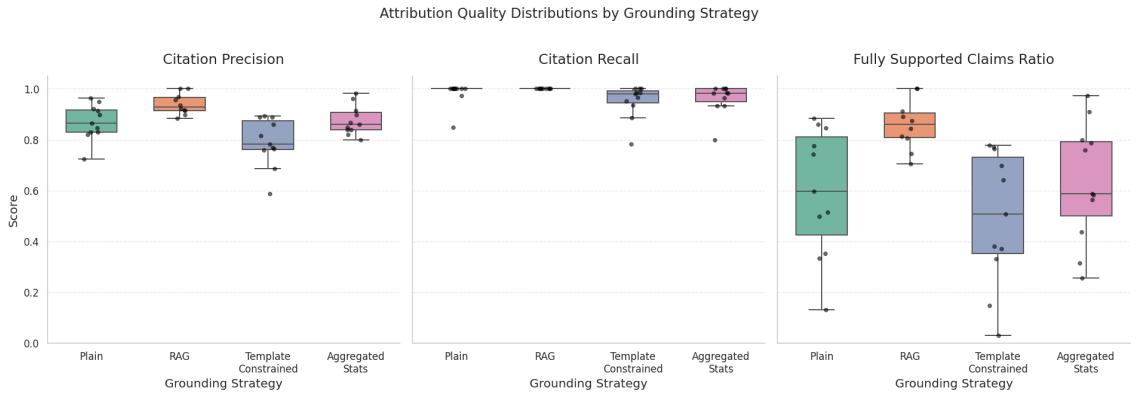

Figure 17: Attribution quality distributions by grounding strategy. Left: Citation Precision. Center: Citation Recall. Right: Fully Supported Claims Ratio. RAG leads on Precision and FSCR with the narrowest spread on both. Template-Constrained is the lowest and most variable on FSCR despite having the highest Information Coverage. Citation Recall shows a clear ceiling effect across all strategies.

Pairwise tests show that RAG is significantly faster than all other strategies, and Plain Prompting is significantly faster than both Template-Constrained Generation and Aggregated-Statistics Grounding. Aggregated-Statistics Grounding and Template-Constrained Generation do not differ significantly, although Template-Constrained has the higher average latency. The detailed pairwise results are summarized in Table 9.

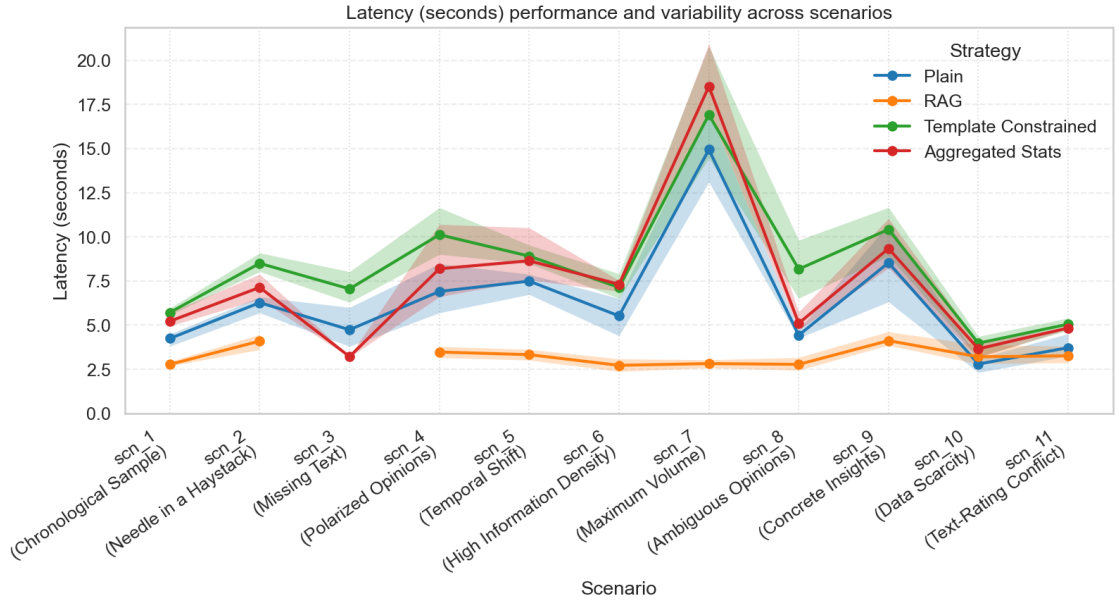

Figure 18: Inference Latency across scenarios. Scenario 7 (Maximum Volume, 500 reviews) produces the most extreme values. Template-Constrained reaches approximately 20 seconds in one run, while RAG stays around 3 seconds.

Figure 18 shows latency by scenario, with per-scenario values in Appendix E. In Scenario 7 (Maximum Volume), the long-context strategies take roughly 13–21 seconds per run while RAG completes in about 3 seconds.

8.5.2 Token Volume

Results. Token Volume produces the highest omnibus Kendall's W in the study (Analysis A: $\chi^2(3) = 28.92$, $p < 0.001$, $W = 0.96$). RAG uses far fewer tokens than every other strategy in every scenario (mean 1,464 vs 8,246–8,987 for the long-context strategies). At the generation stage, RAG uses approximately 82–84% fewer tokens than the long-context strategies. When the query-expansion setup call is included, the reduction is smaller but the overall cost advantage remains.

Pairwise tests show that RAG uses significantly fewer tokens than all other strategies. Plain Prompting also uses significantly fewer tokens than Template-Constrained Generation and Aggregated-Statistics Grounding. Aggregated-Statistics Grounding and Template-Constrained Generation are not significantly different in Analysis A. Figure 19 shows token volume across all scenarios. The detailed pairwise results are summarized in Table 9.

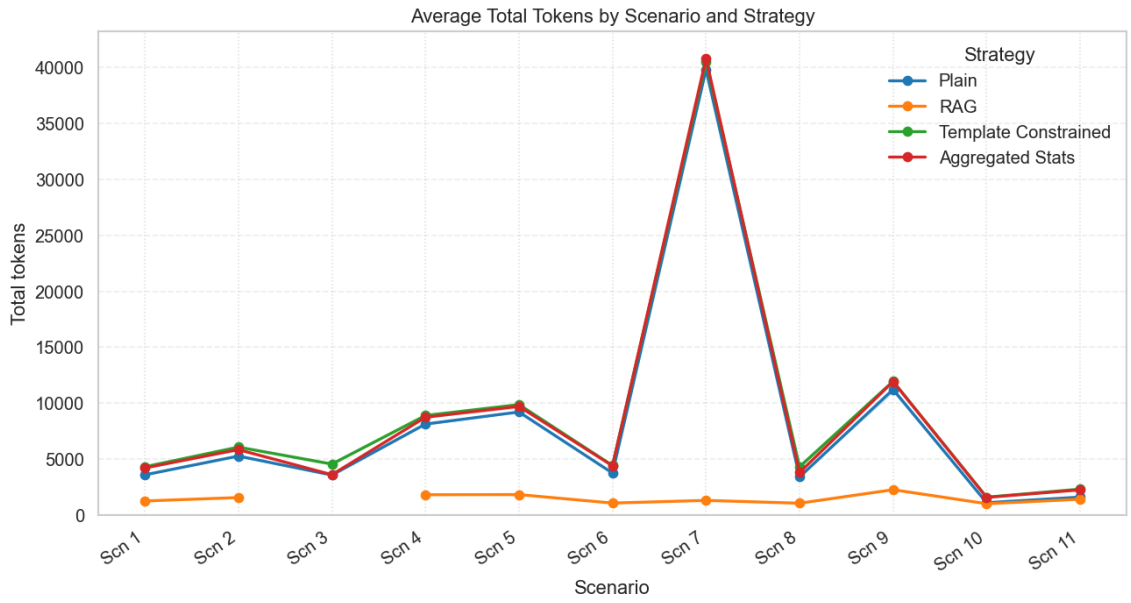

Figure 19: Total Token Volume across scenarios. The three long-context strategies track closely and spike sharply in Scenario 7 (Maximum Volume), at approximately 40,000 tokens per run. RAG remains flat and low (approximately 1,300 tokens even in Scenario 7 (Maximum Volume)).

The large standard deviations for the long-context strategies (±10,927–11,089 tokens) are driven almost entirely by Scenario 7 (Maximum Volume).

Interpretation. RAG's cost advantage is structural: its prompt stays small regardless of the total review volume. This advantage is real, but it is inseparable from the coverage limitation identified under RQ1.

8.6 Manual Audit of Judge Verdicts

Thirty judge verdicts were manually reviewed as described in Section 7.8: 10 factual-verification, 10 coverage-evaluation, and 10 attribution-evaluation cases. The cases were sampled randomly while ensuring coverage across scenarios and grounding strategies where possible. Of these, 28 agreed with the auditor's judgment, giving

an observed agreement rate of 93% on this small audit sample. Because only 30 verdicts were audited, this should be interpreted as a limited construct-validity check rather than a full validation of the judge. The result is nevertheless encouraging and broadly consistent with prior work showing high agreement between GPT-4-class judges and expert human evaluators [9].

The audit showed that the judge handled several difficult cases well. In factual-verification tasks, it correctly matched the claim "food has the highest average score" to aspect-level ratings derived from the raw review records, even when no statistics block had been given to the generator. In the 500-review scenario, it also located a staff member praised by name in only two reviews, explaining that *"[WAITER NAME] is praised by name in reviews 123 and 125."* It also recognized semantic equivalence, for example treating "not properly cooked" and "food being cold in the middle" as support for the same undercooked food claim.

In coverage tasks, the judge correctly accepted a reworded summary as covering an insight about the lack of recurring complaints. The golden insight stated that the dataset was dominated by positive feedback with no consistent negative theme, while the summary stated that *"a small number of reviews are mixed or negative, but these do not outweigh the generally favorable food feedback."* It also correctly returned Not Covered when a summary mentioned menu variety but omitted the more specific claim that frequent menu changes encourage repeat visits. In attribution tasks, the judge was usually able to identify citations that were only loosely related to a claim and explain why they did not support it.

Two disagreements were found. In one coverage case, the golden insight contained three conjunctive items. The summary covered two of them, which the prompt defines as partial coverage and therefore Covered, but the judge returned Not Covered. This suggests that the judge may handle insights containing several related details inconsistently.

The more important disagreement occurred in an attribution case in Scenario 3 (Missing Text). The Template-Constrained strategy generated the claim *"there is insufficient review text to identify detailed reasons for the ratings"*, citing all 50 reviews. The judge marked the claim as Not Supported because it evaluates each cited review individually. Each review only shows that its own text is missing, while the claim becomes supported when all 50 reviews are considered together. The manual audit therefore marked the claim as Supported. This shows a limitation of the attribution judge for claims that depend on evidence from the dataset as a whole rather than from individual reviews. To improve coverage, the sample avoided repeated runs of the same scenario-strategy pair and instead used one run per pair where possible. Within this sample, only two disagreements were found, and they do not change the main strategy rankings.

A similar case, outside this audit sample, appears in the Scenario 11 (Text-Rating Conflict) walkthrough (Appendix C.2): for Plain Prompting, Run 1, the judge marked "positive written feedback appears in multiple low-star reviews" as Covered, even though the summary only called the reviews mixed and never connected positive comments to low-star ratings. This suggests the coverage judge can be too permissive with partial semantic matches.

8.7 Summary of Main Findings

Table 9 condenses the key inferential results from Analysis A. No single strategy performs best on every dimension.

Table 9: Significant pairwise findings under Analysis A (Holm-corrected Wilcoxon, $\alpha = 0.05$, $n = 10$ scenarios). $X > Y$ means X scores significantly higher on quality; $X < Y$ means lower operational cost. Non-significant comparisons omitted.

Metric	**Friedman** p	W	**Significant pairwise results (Holm-corrected)**
Factual Accuracy	0.041	0.28	Aggregated-Stats > RAG ($\lvert r\rvert = 0.93$); Template-Constrained > RAG ($\lvert r\rvert = 0.89$)
Information Coverage	< 0.001	0.60	Plain, Aggregated-Stats, and Template-Constrained all > RAG (all $\lvert r\rvert \geq 0.93$)
Citation Precision	< 0.001	0.63	RAG > Plain ($\lvert r\rvert = 1.00$); RAG > Template-Constrained ($\lvert r\rvert = 0.96$); Plain > Template-Constrained ($\lvert r\rvert = 0.93$); Aggregated-Stats > Template-Constrained ($\lvert r\rvert = 0.89$)
Citation Recall	0.003	0.47	No pair survives Holm correction (ceiling effect).
FSCR	< 0.001	0.74	RAG > Plain ($\lvert r\rvert = 1.00$); RAG > Aggregated-Stats ($\lvert r\rvert = 0.93$); RAG > Template-Constrained ($\lvert r\rvert = 1.00$); Aggregated-Stats > Template-Constrained ($\lvert r\rvert = 1.00$); Plain > Template-Constrained ($\lvert r\rvert = 0.89$)
Inference Latency	< 0.001	0.90	RAG < Plain ($\lvert r\rvert = 0.96$); RAG < Aggregated-Stats ($\lvert r\rvert = 1.00$); RAG < Template-Constrained ($\lvert r\rvert = 1.00$); Plain < Aggregated-Stats ($\lvert r\rvert = 1.00$); Plain < Template-Constrained ($\lvert r\rvert = 1.00$)
Total Token Volume	< 0.001	0.96	RAG < Plain ($\lvert r\rvert = 1.00$); RAG < Aggregated-Stats ($\lvert r\rvert = 1.00$); RAG < Template-Constrained ($\lvert r\rvert = 1.00$); Plain < Aggregated-Stats ($\lvert r\rvert = 1.00$); Plain < Template-Constrained ($\lvert r\rvert = 1.00$)

Scenario design goals. The 11 scenarios were designed to test specific diagnostic conditions (Section 7.3). Overall, the scenario-level results show distinct strategy behavior under the intended diagnostic conditions:

- **Scenario 1 (Chronological Sample):** This scenario worked as the standard scenario. All four strategies produced valid outputs, allowing the main trade-offs between factual accuracy, coverage, attribution faithfulness, and cost to be compared. RAG already showed lower coverage than the long-context strategies, suggesting that low-frequency complaints can be missed when only a small retrieved subset is summarized.
- **Scenario 2 (Needle in a Haystack):** Plain Prompting consistently surfaced the single hygiene signal, while RAG missed it. The scenario therefore captured minority-signal retrieval failure: a business-critical but minor signal can be lost under top-k retrieval even when the summary still appears factually plausible.
- **Scenario 3 (Missing Text):** RAG produced no valid output, showing its dependence on retrievable review text even when numerical ratings and aspect scores are available. Aggregated-Statistics Grounding consistently acknowledged the absence of written review text, while Plain Prompting only partly expressed this limitation. The grounded-refusal goal was therefore met most clearly by the statistics-based strategy.

- **Scenario 4 (Polarized Opinions):** This scenario contained both positive and critical reviews. The long-context strategies preserved this mixed feedback more clearly than RAG, while RAG tended to focus on a smaller subset of evidence. This scenario tested well whether a strategy could represent both sides of customer sentiment.
- **Scenario 5 (Temporal Shift):** The temporal sentiment shift was handled most reliably by the long-context strategies, which preserved the contrast between earlier positive reviews and later negative reviews. RAG showed much lower coverage, indicating that temporal change is difficult to recover from a small retrieved subset.
- **Scenario 6 (High Information Density):** RAG's coverage dropped to 0.20, showing the pressure that fixed top-k retrieval faces when each review contains a different signal. The long-context strategies retained much broader coverage in the same high-information-density setting.
- **Scenario 7 (Maximum Volume):** The long-context strategies maintained high coverage (0.85–0.91), while RAG dropped to 0.27. The maximum-volume scenario therefore exposed the coverage-cost trade-off at scale: RAG remains efficient, but misses more distributed information.
- **Scenario 8 (Ambiguous Opinions):** The vague-review scenario largely met its goal. The long-context strategies achieved high coverage in a setting where the source evidence was deliberately uncertain and non-committal, while RAG covered less of the available evidence. This suggests that uncertainty is easier to preserve when the full evidence surface is visible, and more fragile under retrieval.
- **Scenario 9 (Concrete Insights):** The actionable-detail scenario was challenging for all strategies, with coverage lower than in many other scenarios. This supports the intended goal: concrete details such as named dishes, wait times, and specific incidents are easily compressed away during summarization.
- **Scenario 10 (Data Scarcity):** The data-scarcity scenario showed that smaller evidence sets can simplify citation behavior for some strategies, especially RAG and Aggregated-Statistics Grounding. However, attribution quality did not improve for every strategy: Template-Constrained Generation still showed weak citation precision. Data volume therefore affects citation behavior, but does not by itself guarantee faithful attribution.
- **Scenario 11 (Text-Rating Conflict):** RAG showed its weakest factual accuracy (0.66) in this scenario, showing sensitivity to the mismatch between written feedback and numerical ratings when only a small subset of reviews is retrieved. The scenario therefore tested whether strategies reconcile conflicting modalities rather than relying too strongly on one type of signal.

Per-strategy summary.

- **RAG** is the most efficient and most precisely attributed strategy. However, it covers substantially less information than the long-context strategies, misses

minority and high-diversity signals, and its failure in Scenario 3 (Missing Text) shows that it needs a fallback when no review text is available.

- **Plain Prompting** is the most balanced baseline: not the best on any single metric, but competitive on factual accuracy, coverage, and citation recall while cheaper than Template-Constrained Generation and Aggregated-Statistics Grounding.
- **Template-Constrained Generation** achieves the highest mean coverage and high factual accuracy, but has the weakest attribution quality. Enforcing a structured schema improves breadth, but does not guarantee faithful citations.
- **Aggregated-Statistics Grounding** matches Template-Constrained on factual accuracy, performs close to Plain Prompting on coverage, and has better citation discipline than Template-Constrained. It costs more than Plain Prompting but avoids the most severe citation problems of the template approach.

The next section interprets these metric-level patterns in terms of the underlying strategy mechanisms: evidence surface size, schema pressure, statistical grounding, and operational scaling.

9 Discussion

This section interprets the results from Section 8 with respect to the thesis purpose: understanding how grounding strategies affect the reliability and practical usefulness of LLM-generated summaries for hybrid customer review data. Section 9.1 answers the research questions, Section 9.2 relates the findings to prior work, Section 9.3 gives practical recommendations, and Section 9.4 discusses limitations and future work.

9.1 Interpretation of Results by Research Question

The five research goals were achieved. An 11-scenario dataset was built from real restaurant review data and manually constructed edge cases. Golden insights were written for each scenario, four grounding pipelines were implemented, outputs were evaluated with a common LLM-as-a-Judge framework, operational telemetry was logged, and a random subset of judge verdicts was manually audited as a construct-validity check (Section 8.6).

RQ1: Factual accuracy and information coverage. Factual accuracy and information coverage are affected differently by the grounding strategy. Aggregated-Statistics Grounding and Template-Constrained Generation are significantly more accurate than RAG, but not significantly more accurate than Plain Prompting. This suggests that adding a statistics block or enforcing a schema helps mainly where the model may otherwise struggle to ground broad numerical claims from raw review text.

The more important difference is coverage. RAG's lower coverage is not mainly caused by wrong claims, but by missing evidence. Because the generator receives only the top-k retrieved reviews, it can only summarize what appears in that subset. Signals distributed across many reviews are therefore invisible, especially in high-volume and high-density scenarios. In answer to RQ1, this fixed top-k RAG configuration should be treated as selected highlights rather than a comprehensive review summary.

RQ2: Attribution faithfulness. RQ2 produced the most counter-intuitive result: RAG has the strongest citation quality despite the weakest coverage, while Template-Constrained Generation has the weakest citation quality despite the highest coverage and the most rigid output format. Two mechanisms explain this.

First, evidence surface size matters. With only 8 retrieved reviews, citation selection is simpler because there are fewer sources to confuse. With 50 or 500 reviews, the model must choose among many similar sources, increasing the risk of loosely relevant citations. Second, schema pressure matters. Template-Constrained Generation requires every claim to populate the `review_ids` field. For broad pattern claims, this encourages citation padding: the model may include reviews that are related but do not individually entail the specific claim. This is visible in Scenario 7 (Maximum Volume), where Template-Constrained produces roughly 870–980 citations per run and has FSCR scores near zero, while factual accuracy remains above 0.93.

Plain Prompting and Aggregated-Statistics Grounding fall between these extremes. They avoid both RAG's narrow evidence surface and the schema pressure of Template-Constrained Generation. In answer to RQ2, citation quality is shaped more by

evidence surface size and schema pressure than by the strictness of the general grounding instruction.

RQ3: Operational trade-offs. At the generation stage, RAG uses approximately 82–84% fewer tokens than the long-context strategies and completes significantly faster. When the query-expansion setup call is included, the reduction is smaller, but the cost advantage remains. This advantage is structural: RAG's prompt remains small regardless of review-pool size, while long-context strategies scale with input size. This explains the token and latency spike in Scenario 7 (Maximum Volume). However, the same design choice that reduces cost also reduces coverage: RAG is cheaper because it processes less evidence.

Overall interpretation. The main conclusion is that grounding strategies create trade-offs rather than a single ranking. Fixed top-k RAG is efficient and well-attributed, but incomplete. Template-Constrained Generation is broad and accurate, but weak on citation discipline. Plain Prompting is a strong and cost-competitive baseline. From an engineering perspective, Aggregated-Statistics Grounding offers a defensible overall balance: it matches the highest factual accuracy, maintains high coverage, and avoids the worst citation problems of the template approach, at slightly higher cost than Plain Prompting. This is a practical multi-criteria recommendation rather than a statistically derived global optimum, since the evaluation does not assign formal weights to accuracy, coverage, attribution quality, latency, and token usage.

9.2 Relation to Previous Work

The coverage weakness of fixed top-k RAG is consistent with prior work showing that retrieval pipelines can miss information when evidence is distributed across many documents [6, 13]. As the document pool grows, a fixed top-k sample becomes less representative. Aggregated-Statistics Grounding partly supports the motivation from White [7]: explicit numerical grounding helps with hybrid numerical data. However, its factual-accuracy advantage over Plain Prompting is not statistically significant, suggesting that the generator used here can already aggregate numerical information from raw reviews. This may not hold for weaker or older models.

For attribution and constrained generation, the results add nuance to prior work. Template-Constrained Generation achieves the highest coverage, suggesting that schema fields encourage broader evidence search, but it has the weakest citation quality. Lexical constraints can improve factual consistency [15], and schema-based decoding can improve output regularity [16]. The present results show that structural enforcement alone does not guarantee faithful citations. This extends the ALCE finding that generated citations do not always support their attached claims [5] to hybrid review summarization. It also supports Song et al.'s view that trustworthy RAG requires both evidence alignment and faithful citation [12]: a narrow evidence surface improves citation faithfulness, but reduces content-level recall.

The evaluation framework follows LLM-as-a-Judge methodology [9, 18], using atomic-fact verification [4] for factual accuracy, golden-insight recall for coverage, and citation checking for attribution. The results show why one metric is not enough: factual accuracy alone hides RAG's coverage weakness, coverage alone hides

Template-Constrained Generation's citation problems, and cost alone makes RAG appear clearly superior.

9.3 Practical Recommendations

The suitable grounding strategy depends on the deployment requirements.

When comprehensive coverage matters most. If the goal is to capture many business-relevant signals, including low-frequency complaints, temporal shifts, or critical incidents, a long-context strategy is preferable to fixed top-k RAG. Aggregated-Statistics Grounding is the safer choice when both high coverage and citation traceability are needed: it matches Template-Constrained Generation on factual accuracy, performs close to Plain Prompting on coverage, and produces more reliable citations than Template-Constrained Generation. Template-Constrained Generation can be useful when structured output is needed, but its citations should be verified before being treated as evidence.

When latency and cost dominate. RAG is best when speed and API cost are primary constraints. At the generation stage, it uses approximately 82–84% fewer tokens than the long-context strategies, and the cost advantage remains even when the query-expansion setup call is included. However, its output should be presented as selected highlights from retrieved reviews, not as a complete summary of all feedback.

When citation auditability is required. If every claim must be traceable to specific source evidence, Template-Constrained Generation should not be used without secondary citation verification. Its structured format looks disciplined, but the FSCR results show that many attached review IDs do not fully support their claims. Plain Prompting and Aggregated-Statistics Grounding are safer starting points because they preserve coverage while avoiding the worst citation-padding behavior.

Broader applicability. The trade-offs observed here may also appear in product reviews, hotel feedback, employee surveys, and customer support logs, where many short texts, mixed sentiment, informal language, and structured numerical signals appear together. The exact scores will differ across domains and models, but the same tensions are likely to remain: restricting evidence lowers cost but risks missing distributed signals, while enforcing output structure improves regularity but does not guarantee faithful attribution.

9.4 Limitations and Future Work

The dataset is English-language restaurant review data from one industry partner, so replication on hotel, product, or employee review data is needed. The RAG pipeline uses BM25 retrieval with fixed $k = 8$, which is sufficient for a controlled comparison but does not cover the full RAG design space.

Evaluation framework limits. As reported in Section 8.6, the attribution prompt's per-source strictness rule cannot validate collective-absence claims, where evidence is distributed across the corpus rather than concentrated in one source. Future work could add a separate evaluation path for corpus-level inferences, following Song et al. [12]. Coverage scores also depend on manually written golden insights. Although explicit rules were used, different annotators might select slightly different key insights, especially for subjective review data. Coverage is normalized by the number of golden insights in each scenario. Since scenarios differ in their content and number of insights, coverage should be interpreted relative to each scenario's reference set. In addition, the generator and judge both come from the OpenAI GPT family, so the potential same-family evaluation bias identified in Section 4.5 also applies to these results. Repeating the evaluation with a judge from a different model family would provide a useful robustness check. The use of qualitative wording, such as "predominantly positive" instead of exact counts such as "42 of 50 reviews", made the evaluation fairer across strategies but also introduced some interpretive flexibility. This should be considered when interpreting absolute coverage values. Attribution faithfulness is treated here as the operational form of explainability, but whether business users find these explanations useful would require a user study. Such a study was outside the scope of this thesis because the research questions focus on system-level factuality, coverage, attribution, and operational cost. Run-level variation was observed even with `temperature=0` and `seed=42`; OpenAI describes the seed parameter as best-effort rather than a strict determinism guarantee [19]. Reporting means and standard deviations across three runs makes this variation visible.

Future work. The most promising direction is hybrid grounding. RAG's citation discipline and the long-context strategies' coverage are complementary, so a system could route between them based on review-pool size or text density. Long-context citation prompting could also be improved by asking the model to cite fewer but more precise sources and to distinguish between single-source and pattern-level claims. Combined strategies, such as RAG plus aggregated statistics or Template-Constrained Generation with a secondary citation-verification pass, may preserve individual strengths while reducing weaknesses. Cheaper non-LLM evaluation components, such as deterministic statistical-claim checks or embedding-based pre-filters for citation verification, are also worth testing against the full LLM-based evaluation.

10 Conclusion

This thesis investigated how four grounding strategies affect the quality and practical cost of LLM-generated summaries for hybrid restaurant review data: Plain Prompting, Retrieval-Augmented Generation, Aggregated-Statistics Grounding, and Template-Constrained Generation. The strategies were compared across 11 evaluation scenarios covering standard review samples and edge cases such as ratings-only input, text-rating conflicts, temporal shifts, and high-volume review sets. Each strategy was evaluated on factual accuracy, information coverage, attribution faithfulness, latency, and token volume.

Main findings. The central finding is that no grounding strategy is best across all dimensions. Each strategy creates a different trade-off.

Factual accuracy is similar for Plain Prompting, Template-Constrained Generation, and Aggregated-Statistics Grounding, while RAG is lower. Adding structure or numerical statistics does not significantly improve factual accuracy beyond the Plain Prompting long-context baseline in this setup. RAG is descriptively lower and significantly lower than Template-Constrained Generation and Aggregated-Statistics Grounding, although not significantly lower than Plain Prompting.

The clearest difference appears in information coverage. RAG covers substantially less information than the other three strategies, with the gap most visible in high-volume and high-density scenarios where important signals are distributed across many reviews. Fixed top-k retrieval is not sufficient when the goal is a comprehensive summary.

Attribution faithfulness shows the opposite pattern. RAG achieves the strongest Citation Precision and Fully Supported Claims Ratio, while Template-Constrained Generation performs worst on both. A narrow retrieved evidence set makes citation selection simpler, whereas a rigid schema can encourage citation-padding even when the cited evidence only partially supports the claim.

Operationally, RAG is the fastest and cheapest strategy. At the generation stage, it uses approximately 82–84% fewer tokens than the long-context strategies. When the query-expansion setup call is included, the reduction is smaller but the overall cost advantage remains. It also completes in roughly half the time of the other strategies. This efficiency advantage comes from the same design choice that causes the coverage gap: RAG processes less evidence.

Contributions. This thesis makes three contributions.

- First, it provides a controlled comparison of four grounding strategies in a realistic business review analytics setting.
- Second, it applies a multi-dimensional evaluation framework combining factual accuracy, golden-insight coverage, citation faithfulness, and operational cost.
- Third, it identifies and documents two specific failure modes: citation-padding in Template-Constrained Generation, where the model sometimes attaches weak citations because the schema expects review IDs, and a limitation of the attribution evaluator when judging claims that depend on evidence being absent across the whole review set rather than present in one specific review.

Practical takeaway. The best strategy depends on the deployment requirements. Within this experimental setup, Aggregated-Statistics Grounding offers the most defensible overall balance when both coverage and traceability matter. Template-Constrained Generation provides the strongest coverage and high factual accuracy, making it useful when broad, structured, machine-readable output is needed. However, its citations are less reliable and should be verified before being used as an audit trail. RAG is the right choice when speed and cost dominate, but its output should be framed as selected highlights rather than a comprehensive summary.

Closing remark. The main lesson is that grounding is not a single improvement that makes LLM summaries better in every way. Restricting evidence improves efficiency and citation precision but reduces coverage. Enforcing structure improves breadth but can weaken attribution. Adding deterministic statistics makes numerical patterns explicit, but it does not automatically produce a large overall improvement, since the long-context baseline can already recover many simple rating patterns. The main remaining difficulty is not only calculation, but deciding how those numerical signals should be combined with the written reviews in a faithful summary. Designing production-ready LLM summarization systems therefore requires choosing the grounding strategy that matches the intended use case.

References

[1] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, Mar. 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3571730

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds. Curran Associates, Inc., 2020, pp. 9459–9474. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[3] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, and H. Hajishirzi, "When not to trust language models: Investigating effectiveness of parametric and non-parametric memories," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 9802–9822. [Online]. Available: https://aclanthology.org/2023.acl-long.546/

[4] S. Min, K. Krishna, X. Lyu, M. Lewis, W. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "FActScore: Fine-grained atomic evaluation of factual precision in long form text generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 2023, pp. 12 076–12 100. [Online]. Available: https://aclanthology.org/2023.emnlp-main.741/

[5] T. Gao, H. Yen, J. Yu, and D. Chen, "Enabling large language models to generate text with citations," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 2023, pp. 6465–6488. [Online]. Available: https://aclanthology.org/2023.emnlp-main.398/

[6] C. G. Belém, P. Pezeshkpour, H. Iso, S. Maekawa, N. Bhutani, and E. Hruschka, "From single to multi: How LLMs hallucinate in multi-document summarization," in *Findings of the Association for Computational Linguistics: NAACL 2025*. Albuquerque, New Mexico: Association for Computational Linguistics, 2025, pp. 5291–5324. [Online]. Available: https://aclanthology.org/2025.findings-naacl.293/

[7] N. M. White, "Probing of quantitative values in abstractive summarization models," 2022. [Online]. Available: https://arxiv.org/abs/2210.00667

[8] C. C. Osuji, T. C. Ferreira, and B. Davis, "A systematic review of data-to-text NLG," 2024. [Online]. Available: https://arxiv.org/abs/2402.08496

[9] L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-Judge with MT-bench and chatbot arena," in *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and*

Benchmarks Track. Curran Associates, Inc., Dec. 2023. [Online]. Available: https://neurips.cc/virtual/2023/poster/73434

[10] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, Jan. 2025. [Online]. Available: https://dl.acm.org/doi/10.1145/3703155

[11] Z. Luo, Q. Xie, and S. Ananiadou, "Factual consistency evaluation of summarisation in the era of large language models," *Expert Systems With Applications*, vol. 254, p. 124456, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417424013228

[12] M. Song, S. H. Sim, R. Bhardwaj, H. L. Chieu, N. Majumder, and S. Poria, "Measuring and enhancing trustworthiness of LLMs in RAG through grounded attributions and learning to refuse," in *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, Singapore, Apr. 2025. [Online]. Available: https://proceedings.iclr.cc/paper_files/paper/2025/file/4c88827decab6c046b881a2c3a99c76f-Paper-Conference.pdf

[13] Z. Li, C. Li, M. Zhang, Q. Mei, and M. Bendersky, "Retrieval augmented generation or long-context LLMs? a comprehensive study and hybrid approach," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Miami, Florida, US: Association for Computational Linguistics, 2024, pp. 881–893. [Online]. Available: https://aclanthology.org/2024.emnlp-industry.66/

[14] X. Li, Y. Cao, Y. Ma, and A. Sun, "Long context vs. RAG for LLMs: An evaluation and revisits," 2024. [Online]. Available: https://arxiv.org/abs/2501.01880

[15] Y. Mao, X. Ren, H. Ji, and J. Han, "Constrained abstractive summarization: Preserving factual consistency with constrained generation," 2021. [Online]. Available: https://arxiv.org/abs/2010.12723

[16] S. Geng, H. Cooper, M. Moskal, S. Jenkins, J. Berman, N. Ranchin, R. West, E. Horvitz, and H. Nori, "Generating structured outputs from language models: Benchmark and studies," 2025. [Online]. Available: https://arxiv.org/abs/2501.10868

[17] O. Honovich, R. Aharoni, J. Herzig, H. Taitelbaum, D. Kukliansy, V. Cohen, T. Scialom, I. Szpektor, A. Hassidim, and Y. Matias, "TRUE: Re-evaluating factual consistency evaluation," in *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. 3905–3920. [Online]. Available: https://aclanthology.org/2022.naacl-main.287/

[18] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-Eval: NLG evaluation using GPT-4 with better human alignment," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore:

Association for Computational Linguistics, Dec. 2023, pp. 2511–2522. [Online]. Available: https://aclanthology.org/2023.emnlp-main.153/

[19] OpenAI, "How to make your completions outputs consistent with the seed parameter," 2023, OpenAI Cookbook. Accessed 2026-04-18. [Online]. Available: https://cookbook.openai.com/examples/reproducible_outputs_with_the_seed_parameter

[20] K. M. Hermann, T. Kočiský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Montreal, Quebec, Canada: Curran Associates, Inc., 2015, pp. 1693–1701. [Online]. Available: https://proceedings.neurips.cc/paper/2015/hash/afdec7005cc9f14302cd0474fd0f3c96-Abstract.html

[21] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 1797–1807. [Online]. Available: https://aclanthology.org/D18-1206/

[22] J. Sequeda, D. Allemang, and B. Jacob, "Knowledge graphs as a source of trust for LLM-powered enterprise question answering," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 85, p. 100858, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1570826824000441

[23] D. Brown, "Rank-BM25: A Collection of BM25 Algorithms in Python," 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4520057

[24] L. Wang, N. Yang, and F. Wei, "Query2doc: Query expansion with large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2023, pp. 9414–9423. [Online]. Available: https://aclanthology.org/2023.emnlp-main.585/

[25] OpenAI, "GPT-5.4 mini," 2026, OpenAI API documentation. Accessed 2026-04-18. [Online]. Available: https://developers.openai.com/api/docs/models/gpt-5.4-mini

[26] F. Baumgärtel, E. Bono, L. Fillinger, L. Galou, K. Keska-Izworska, S. Walter, P. Andorfer, K. Kratochwill, P. Perco, and M. Ley, "A systematic evaluation and benchmarking of text summarization methods for biomedical literature: From word-frequency methods to language models," 2026. [Online]. Available: https://www.biorxiv.org/content/10.64898/2026.01.09.697335v3

[27] OpenAI, "GPT-5.4," 2026, OpenAI API documentation. Accessed 2026-04-18. [Online]. Available: https://developers.openai.com/api/docs/models/gpt-5.4

A Scenario Descriptive Statistics

Table 10 reports the descriptive statistics for all 11 evaluation scenarios. The table is included to make the scenario construction transparent and to show the numerical properties available to the Aggregated-Statistics Grounding strategy.

Table 10: Descriptive statistics for all 11 evaluation scenarios. Aspect averages are absent for synthetic scenarios (Scenario 6 (High Information Density) and Scenario 8 (Ambiguous Opinions)) because no structured ratings were defined. Dates are omitted for Scenario 6 (High Information Density) and Scenario 8 (Ambiguous Opinions) because temporal information was not part of those synthetic diagnostic conditions.

Scenario	Avg	Med	Min	Max	SD	Rating counts	Min date	Max date	With text	No text	Total
Scenario 1 (Chronological Sample)	4.1	4.0	1	5	1.005	1:3; 3:5; 4:23; 5:19	2017-07-24	2026-01-21	16	34	50
Scenario 2 (Needle in a Haystack)	4.9	5.0	3	5	0.361	3:1; 4:3; 5:46	2023-05-09	2026-01-05	35	15	50
Scenario 3 (Missing Text)	4.1	4.0	1	5	1.005	1:3; 3:5; 4:23; 5:19	2017-07-24	2026-01-21	0	50	50
Scenario 4 (Polarized Opinions)	4.26	5.0	1	5	1.146	1:1; 2:5; 3:7; 4:4; 5:33	2021-10-04	2026-01-18	50	0	50
Scenario 5 (Temporal Shift)	3.16	3.5	1	5	1.837	1:18; 2:4; 3:3; 4:2; 5:23	2022-07-16	2024-12-30	50	0	50
Scenario 6 (High Information Density)	3.7	4.0	2	5	1.082	2:8; 3:15; 4:11; 5:16	–	–	50	0	50
Scenario 7 (Maximum Volume)	4.2	5.0	1	5	1.138	1:28; 2:22; 3:50; 4:122; 5:278	2020-02-13	2026-01-06	313	187	500
Scenario 8 (Ambiguous Opinions)	3.12	3.0	2	4	0.382	2:1; 3:42; 4:7	–	–	50	0	50
Scenario 9 (Concrete Insights)	3.46	4.0	1	5	1.431	1:6; 2:9; 3:9; 4:8; 5:18	2023-01-21	2025-12-29	50	0	50
Scenario 10 (Data Scarcity)	3.0	4.0	1	5	1.673	1:2; 4:2; 5:1	2024-04-20	2026-01-21	4	1	5
Scenario 11 (Text-Rating Conflict)	3.0	3.0	1	5	1.673	1:3; 2:2; 4:2; 5:3	2024-11-21	2026-03-21	10	0	10

Table 10: Descriptive statistics for all 11 evaluation scenarios (continued).

Scenario	**Food (avg|n|sd)**	**Service (avg|n|sd)**	**Atmosphere (avg|n|sd)**
Scenario 1 (Chronological Sample)	3.857|7|1.457	3.5|8|1.581	3.375|8|1.495
Scenario 2 (Needle in a Haystack)	4.975|40|0.156	4.658|38|0.77	4.622|37|0.672
Scenario 3 (Missing Text)	3.857|7|1.457	3.5|8|1.581	3.375|8|1.495
Scenario 4 (Polarized Opinions)	4.523|44|0.988	4.535|43|0.817	4.395|43|0.919
Scenario 5 (Temporal Shift)	3.026|38|1.724	3.375|40|1.608	4.025|40|1.369
Scenario 6 (High Information Density)	–	–	–
Scenario 7 (Maximum Volume)	4.406|229|0.969	4.458|249|1.002	4.462|247|0.813
Scenario 8 (Ambiguous Opinions)	–	–	–
Scenario 9 (Concrete Insights)	3.674|43|1.359	3.523|44|1.454	3.465|43|1.318
Scenario 10 (Data Scarcity)	3.0|3|1.633	2.75|4|1.785	2.5|4|1.5
Scenario 11 (Text-Rating Conflict)	3.2|10|1.47	3.3|10|1.676	3.9|10|0.943

B Scenario 11 (Text-Rating Conflict): Full Input Data

Scenario 11 (Text-Rating Conflict) is reproduced in full because it is synthetic and has no privacy concerns. It tests modality conflict: low-star reviews contain clearly positive written text and high aspect sub-ratings, while high-star reviews contain clearly negative text and low sub-ratings. Appendix C uses this data to trace one complete experimental run through generation and evaluation. Table 11 gives the aggregated statistics block and Table 12 lists all ten review records.

B.1 Aggregated Statistics

Table 11: Aggregated statistics for Scenario 11 (Text-Rating Conflict).

Field	**Value**
Average rating	3.0
Median rating	3.0
Std deviation	1.673
Min / Max rating	1 / 5
Rating distribution	1$\star$: 3, 2$\star$: 2, 4$\star$: 2, 5$\star$: 3
Total reviews	10
Reviews with text	10
Reviews without text	0
Date range	2024-11-21 to 2026-03-21
Food avg ($n = 10$)	3.20 (sd 1.470)
Service avg ($n = 10$)	3.30 (sd 1.676)
Atmosphere avg ($n = 10$)	3.90 (sd 0.943)

B.2 Review Records

Table 12: All 10 review records in Scenario 11 (Text-Rating Conflict). Low-star reviews tend to contain positive text and high aspect sub-ratings; high-star reviews tend to contain negative text and low sub-ratings.

ID	Stars	Review text	Food	Svc	Atm
1	1	Amazing staff and delicious pasta. We had a wonderful dinner and would happily return.	5	5	4
2	5	The soup was cold, the service was slow, and the table was dirty. Very disappointing visit.	2	1	2
3	2	Friendly service, generous portions, and excellent value for money.	4	5	4
4	4	The burger arrived raw in the middle and we waited 45 minutes. I would not recommend it.	1	1	3
5	5	The terrace view is nice, but the food was bland and the wine pairing felt overpriced.	2	3	5
6	1	Cozy room, fast service, and one of the best desserts I have had in months.	5	4	5
7	4	We were given the wrong order, the waiter never checked back, and the fish was dry.	2	1	3
8	2	Excellent cocktails, attentive staff, and beautifully presented dishes.	4	5	5
9	5	Prices were absurd for the portion size, and the pasta tasted reheated.	2	3	4
10	1	The restaurant handled our allergy request carefully, the chicken was excellent, and the meal felt thoughtful from start to finish.	5	5	4

B.3 Golden Insights

The 14 golden insights for this scenario are listed below. Each is paired with the evidence that grounds it. These insights are used as the coverage reference in Appendix C.

1. Written feedback across the reviews is mixed, with both clearly positive and clearly negative statements present.
   Evidence: positive text in reviews 1, 3, 6, 8, 10; negative text in reviews 2, 4, 5, 7, 9.

2. Positive written feedback explicitly praises the food quality.
   Evidence: pasta in review 1, desserts in review 6, chicken in review 10.

3. Multiple customers praise the friendly and attentive service.
   Evidence: reviews 1, 3, 8, 10 describe staff as amazing, friendly, attentive, or careful.

4. Specific positive mentions include good value, cocktails, and allergy handling.
   Evidence: value in review 3, cocktails in review 8, allergy handling in review 10.

5. Negative written feedback explicitly criticizes poor food quality.
   Evidence: cold soup in review 2, raw burger in review 4, bland food in review 5, dry fish in review 7, reheated pasta in review 9.

6. Multiple customers complain about slow or inattentive service.
   Evidence: slow service in review 2, 45-minute wait in review 4, waiter not checking back in review 7.

7. Specific negative mentions include dirty conditions, wrong orders, and poor value for money.
   Evidence: dirty table in review 2, wrong order in review 7, overpriced or poor value in reviews 5 and 9.

8. Ratings are distributed across the full scale from 1 to 5 stars, with no single rating level clearly dominating.
   Evidence: statistics show 1⋆: 3, 2⋆: 2, 4⋆: 2, 5⋆: 3.

9. Positive written feedback appears in multiple low-star reviews.
   Evidence: reviews 1, 3, 6, 8, 10 are all positive despite 1- or 2-star overall ratings.

10. Negative written feedback appears in multiple high-star reviews.
    Evidence: reviews 2, 4, 5, 7, 9 are all negative despite 4- or 5-star overall ratings.

11. These observations indicate a mismatch between written feedback and overall star ratings.
    Evidence: pattern consistent across all 10 reviews.

12. Detailed aspect ratings (food, service, atmosphere) generally align more closely with the written feedback than with the overall star rating.
    Evidence: positive-text reviews have high aspect scores despite low overall stars; negative-text reviews have low aspect scores despite high overall stars.

13. Specific contradictory cases exist, including positive feedback with 1-star ratings and negative feedback with high-star ratings.
    Evidence: reviews 1, 6, 10 are 1⋆ with positive text; reviews 2, 5, 9 are 5⋆ with negative text.

14. Due to the direct contradiction between the written feedback and the numerical ratings, it is impossible to determine a single, consistent overall customer sentiment.
    Evidence: written feedback is evenly split, the rating average is exactly 3.0, and ratings disagree systematically with text.

C Scenario 11 (Text-Rating Conflict): Generation and Evaluation

This appendix traces one complete experimental run (Run 1) of Scenario 11 (Text-Rating Conflict) through all four grounding strategies. It shows the actual generated summaries, the per-insight coverage verdicts, and the attribution scores. The full input data is in Appendix B. This is the same evaluation pipeline described in Section 7.7 applied to a single concrete case.

C.1 Generated Summaries (Run 1)

The four strategies received the same ten review records. They differed only in what additional evidence was provided and in the required output format.

Plain Prompting

Customers consistently praise the staff and service when it is good, describing it as amazing, friendly, attentive, fast, and careful with allergy requests [REVIEW: 1, 3, 6, 8, 10]. They also frequently compliment specific food items such as pasta, desserts, cocktails, chicken, and the presentation of dishes [REVIEW: 1, 6, 8, 10]. Several reviews mention a pleasant atmosphere or setting, including a cozy room, a nice terrace view, and a beautiful presentation [REVIEW: 5, 6, 8].

The most common complaints are about food quality and service problems [REVIEW: 2, 4, 5, 7, 9]. Customers reported cold soup, a raw burger, dry fish, bland food, reheated pasta, and an overall disappointing meal [REVIEW: 2, 4, 5, 7, 9]. Service issues included slow service, a long wait of 45 minutes, the wrong order, and a waiter who did not check back [REVIEW: 2, 4, 7]. Pricing and value were also criticised, with comments that the wine pairing felt overpriced and that prices were absurd for the portion size [REVIEW: 5, 9].

A notable pattern is that the reviews are sharply mixed: some guests describe excellent meals and thoughtful service, while others report serious problems with food preparation and attentiveness [REVIEW: 1, 2, 4, 6, 7, 10]. Another recurring theme is inconsistency, since the same restaurant is praised for dishes like pasta, desserts, and chicken, but also criticised for pasta being reheated and other dishes being poorly prepared [REVIEW: 1, 6, 9, 10].

RAG (8 reviews retrieved)

Customers consistently praise the restaurant's service, atmosphere, and presentation in several reviews, with mentions of friendly or attentive staff, fast service, a cozy room, a nice terrace view, excellent cocktails, beautifully presented dishes, and generous portions [REVIEW: 3, 8, 6, 5]. Food quality is also praised in some reviews, especially the desserts and portions, with one reviewer calling the desserts "one of the best" they had in months [REVIEW: 6, 3].

At the same time, there are repeated complaints about food quality and value, including bland food, reheated pasta, cold soup, dry fish, and a burger served raw in the middle [REVIEW: 9, 5, 2, 7, 4]. Pricing is another recurring issue, with customers describing prices as absurd for the portion size and the wine pairing as overpriced [REVIEW: 9, 5]. Service problems also appear multiple times, including slow service, a waiter who never checked back, the wrong order being delivered, and a 45-minute wait [REVIEW: 2, 7, 4].

A notable pattern is that the reviews are mixed and sometimes internally inconsistent: several reviews give very positive comments in the text while the star ratings are low, and some negative comments appear alongside high star ratings [REVIEW: 3, 8, 6, 2, 5, 4]. Overall, the most consistent themes are praise for ambiance and some aspects of service, alongside frequent criticism of food quality, pricing, and reliability of service [REVIEW: 3, 8, 6, 5, 9, 2, 7, 4].

Aggregated-Statistics Grounding

Customers give the restaurant a mixed overall impression, with an average rating of 3.0/5 and a median rating of 3.0/5 across 10 reviews [STAT: average_rating, STAT: median_rating, STAT: total_reviews]. The ratings are spread across the full range from 1 to 5, with 3 one-star reviews, 2 two-star reviews, 2 four-star reviews, and 3 five-star reviews, showing substantial variability [STAT: rating_range, STAT: rating_distribution, STAT: rating_variability].

A consistent strength is the atmosphere, which has the highest average aspect score at 3.9/5 [STAT: atmosphere_average]. Several reviewers specifically praised the setting or ambiance, including a nice terrace view, a cozy room, and a pleasant atmosphere overall [REVIEW: 5, 6]. The restaurant also receives repeated praise for staff and service in some reviews, with comments about amazing staff, friendly service, attentive staff, fast service, and careful handling of an allergy request [REVIEW: 1, 3, 6, 8, 10].

Food quality is more uneven. The food aspect averages 3.2/5, which is lower than atmosphere but still above neutral [STAT: food_average]. Positive food comments mention delicious pasta, generous portions, excellent desserts, beautifully presented dishes, excellent cocktails, and excellent chicken [REVIEW: 1, 3, 6, 8, 10]. However, several reviews complain about bland food, cold soup, a raw burger, dry fish, reheated pasta, and a generally disappointing meal [REVIEW: 2, 4, 5, 7, 9].

Service is also inconsistent, with an average of 3.3/5 and the highest variability among the aspect scores [STAT: service_average, STAT: service_stddev]. Some customers describe the service as friendly, attentive, fast, or thoughtful [REVIEW: 1, 3, 6, 8, 10]. Others report slow service, a waiter who never checked back, and poor handling of the visit overall [REVIEW: 2, 4, 7].

A recurring pattern is that several reviews pair strong atmosphere or service with weak food, suggesting the dining experience can feel uneven [REVIEW: 5, 8, 10]. Another repeated complaint is value for money: one reviewer called the wine pairing overpriced, and another said the prices were absurd for the portion size [REVIEW: 5, 9]. There are also isolated but notable operational issues, including a dirty table, a wrong order, and a long 45-minute wait [REVIEW: 2, 4, 7].

Overall, the reviews suggest a restaurant with appealing ambiance and occasional standout service, but inconsistent food quality and recurring concerns about speed, accuracy, and pricing [STAT: average_rating, STAT: atmosphere_average, STAT: food_average, STAT: service_average, REVIEW: 2, 4, 5, 7, 9].

Template-Constrained Generation

The model produced the following structured JSON output.


```
{
  "overall_sentiment": {
    "claims": [
      {
        "statement": "The reviews are mixed overall, with several customers praising the food, service, and
      atmosphere, while others report serious problems with food quality, service, and value.",
        "sentiment": "Mixed",
        "review_ids": [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
      }
    ]
  },
  "praise_theme": {
    "claims": [
      {
        "statement": "Customers frequently praise the staff and service as friendly, attentive, fast, or
      thoughtful.",
        "review_ids": [1, 3, 6, 8, 10]
      },
      {
        "statement": "Several reviews highlight specific dishes or drinks as excellent, including pasta, desserts,
       cocktails, and chicken.",
        "review_ids": [1, 6, 8, 10]
      },
      {
        "statement": "Some customers appreciate the atmosphere and setting, including a cozy room and a nice
      terrace view.",
        "review_ids": [5, 6, 8, 10]
      },
      {
        "statement": "One review praises generous portions and excellent value for money.",
        "review_ids": [3]
      }
    ]
  },
  "complaint_theme": {
    "claims": [
      {
        "statement": "Recurring complaints focus on poor food quality, including cold soup, raw burger, bland food,
       dry fish, reheated pasta, and an overall disappointing meal.",
        "review_ids": [2, 4, 5, 7, 9]
      },
      {
        "statement": "Several customers complain about slow or inattentive service, including long waits, a waiter
       not checking back, and a wrong order.",
        "review_ids": [2, 4, 7]
      },
```

```
      {
        "statement": "Price and value concerns appear in the reviews, especially with overpriced wine pairing and
      absurd prices for the portion size.",
        "review_ids": [5, 9]
      },
      {
        "statement": "One review mentions a dirty table, indicating cleanliness issues.",
        "review_ids": [2]
      }
    ]
  },
  "specific_signal": {
    "claims": [
      {
        "statement": "Portion size and pricing are a recurring value concern, with one review calling the portions
       small relative to the price and another describing prices as absurd.",
        "review_ids": [3, 9]
      },
      {
        "statement": "Food temperature and preparation quality are specific issues, including cold soup, a raw
      burger, dry fish, and reheated pasta.",
        "review_ids": [2, 4, 7, 9]
      },
      {
        "statement": "Service responsiveness is inconsistent, ranging from fast and attentive to slow,
      unresponsive, or incorrect order handling.",
        "review_ids": [2, 4, 6, 7, 8, 10]
      },
      {
        "statement": "The restaurant appears capable of accommodating special requests, as one review specifically
       praises careful handling of an allergy request.",
        "review_ids": [10]
      }
    ]
  },
  "other_insights": {
    "claims": [
      {
        "statement": "There is evidence of strong inconsistency across reviews, with some customers describing the
       same restaurant experience as excellent while others found it very disappointing.",
        "aspect": "Inconsistency",
        "sentiment": "Mixed",
        "review_ids": [1, 2, 4, 6, 8, 10]
      },
      {
        "statement": "Atmosphere-related comments are generally positive when mentioned, but they do not outweigh
      the food and service issues in negative reviews.",
        "aspect": "Atmosphere",
        "sentiment": "Mixed",
        "review_ids": [5, 6, 8, 10, 2, 4, 7]
      }
    ]
  }
}
```


C.2 Coverage Results by Golden Insight

Table 13 shows the coverage verdict for each of the 14 golden insights, for each strategy, in Run 1. A verdict of **C** (Covered) means the judge found the core meaning present in the summary; **N** (Not Covered) means it was absent. The final row shows the coverage score for each strategy.

Table 13: Coverage verdicts for each of the 14 golden insights in Scenario 11 (Text-Rating Conflict), Run 1. C = Covered, N = Not Covered.

#	Golden insight (shortened)	Plain	RAG	Agg	Tmpl
1	Written feedback is mixed (positive and negative present)	C	C	C	C
2	Positive feedback praises food quality	C	C	C	C
3	Multiple customers praise friendly/attentive service	C	C	C	C
4	Specific positive mentions: good value, cocktails, allergy handling	C	N	C	C
5	Negative feedback criticizes food quality	C	C	C	C
6	Multiple customers complain: slow or inattentive service	C	C	C	C
7	Specific negative mentions: dirty conditions, wrong orders, poor value	C	C	C	C
8	Ratings span full 1–5 star scale with no level dominating	N	C	C	C
9	Positive written feedback in multiple low-star reviews	C	C	C	C
10	Negative written feedback in multiple high-star reviews	N	C	C	C
11	Systematic mismatch between written feedback and star ratings	N	C	N	N
12	Aspect ratings align with text more than with overall star	N	C	N	N
13	Specific contradictory cases: positive text with 1-star ratings	N	C	N	N
14	Cannot determine single consistent sentiment due to contradiction	C	C	C	C
Coverage score (covered / 14)		**9/14**	**13/14**	**11/14**	**11/14**
Coverage score (fraction)		**0.643**	**0.929**	**0.786**	**0.786**

What the coverage results show. The modality-conflict insights were the hardest for the long-context strategies to express explicitly. Plain Prompting missed most of the numerical conflict structure, while Aggregated-Statistics Grounding and Template-Constrained Generation captured the general high-star/negative-text pattern but not the deeper mismatch between text, overall stars, and aspect ratings. RAG achieved the highest coverage in this run because its summary explicitly stated that positive comments appeared with low star ratings and negative comments appeared with high star ratings. This is RAG's strongest run in the scenario; averaged across all three runs, RAG (0.79) is ahead of Plain Prompting (0.71) and Aggregated-Statistics Grounding (0.74) but still trails Template-Constrained Generation (0.81). This illustrates the diagnostic value of Scenario 11 (Text-Rating Conflict): a summary can describe textual sentiment accurately while still failing to explain the structural conflict between rating modalities.

C.3 Attribution Results

Table 14 shows the attribution scores for Run 1 of Scenario 11 (Text-Rating Conflict). These are the raw pipeline outputs, not averages.

Table 14: Attribution faithfulness scores for Scenario 11 (Text-Rating Conflict), Run 1. CP = Citation Precision; CR = Citation Recall; FSCR = Fully Supported Claims Ratio.

Strategy	CP	CR	FSCR	Claims	Citations	Valid cit.
Plain Prompting	1.000	1.000	1.000	9	37	37
RAG	0.967	1.000	0.857	7	30	29
Aggregated-Statistics	0.846	0.867	0.667	15	39	33
Template-Constrained	0.902	1.000	0.733	15	61	55

What the attribution results show. The attribution pattern is easiest to read as a trade-off between breadth and citation discipline. Plain Prompting is the cleanest in this run: when it makes a claim, its citations align closely with that claim. RAG is also generally reliable, but occasional mismatches still appear. The two more structured strategies produce broader, denser summaries, which gives the reader more coverage but also creates more opportunities for citations that support only part of a claim. Aggregated-Statistics Grounding seems especially vulnerable when it combines review evidence with statistic fields, while Template-Constrained Generation shows the same general problem in a different form: the fixed JSON structure pushes the model to attach citations to many broad thematic statements. Overall, the scores suggest that saying more is not the same as citing more faithfully. The most expansive summaries are also the hardest to keep fully supported at the claim level.

C.4 Summary

Table 15 brings together the generation telemetry and evaluation scores for all four strategies in Run 1 of Scenario 11 (Text-Rating Conflict).

Table 15: Scenario 11 (Text-Rating Conflict), Run 1 summary: telemetry and evaluation scores. FA = Factual Accuracy; IC = Information Coverage.

Strategy	Tokens	Lat (s)	FA	IC	CP	CR	FSCR
Plain Prompting	1,614	3.25	0.721	0.643	1.000	1.000	1.000
RAG	1,416	3.79	0.763	0.929	0.967	1.000	0.857
Aggregated-Statistics	2,264	4.77	0.839	0.786	0.846	0.867	0.667
Template-Constrained	2,326	5.06	0.847	0.786	0.902	1.000	0.733

D Evaluation Prompts and Supporting Schemas

This appendix reproduces the prompts used in the automated evaluation pipeline and the output schema used by Template-Constrained Generation. These prompts operationalize the metrics defined in Section 2.3 and applied in Section 8. They are reproduced with minor line-wrapping and formatting changes for readability. The substantive instructions, placeholders, and JSON output contracts match the implementation, so the evaluation can be replicated with the same judge model and settings described in Section 7.5.

The evaluation pipeline runs in four stages:

1. Atomic fact decomposition (Listing 2).

2. Atomic fact verification (Listing 3).
3. Information coverage evaluation (Listing 4).
4. Attribution faithfulness evaluation (Listing 5).

All evaluation calls shared the same judge system message (Listing 1). Judge temperature was fixed at $T = 0$ and the random seed at `seed=42`.

D.1 Judge System Prompt

The message below was prepended to every evaluation call. It instructs the judge to act as a strict evaluator and return only valid JSON.

```
You are a strict evaluator. Return only valid JSON matching the
requested schema.
```

Listing 1: Shared judge system prompt, prepended to every evaluation call.

D.2 Atomic Fact Decomposition Prompt

Used in Stage 1 of factual accuracy evaluation. The judge receives a generated summary and decomposes it into independent, verifiable atomic facts. The resulting list feeds Stage 2 (Listing 3). The placeholder `{summary}` is replaced with the full generated summary text at runtime.

```
You are performing atomic fact decomposition for factual accuracy
evaluation.

Your task:
Break the given review summary into a complete list of independent,
minimal, verifiable atomic facts.

STRICT REQUIREMENTS:

1. Complete Coverage:
Extract ALL verifiable factual claims in the summary.
Continue extracting facts until no additional verifiable claims remain.
Do not stop prematurely.

2. Atomicity:
Each fact must contain exactly ONE predicate (one independent piece
of information).

3. No Redundancy:
If two potential facts express the same underlying information, include
only ONE version using the most specific wording. Do not include
paraphrased duplicates.

4. No Over-Splitting:
Do NOT split stylistic modifiers or descriptive adjectives unless they
represent independent verifiable claims.

5. Numeric, Comparative, and Temporal Handling:
If a sentence contains multiple independent predicates, separate each
into its own atomic fact. However, preserve temporal, comparative, or
directional claims as a single fact when splitting would remove the
relationship being asserted.
```

```
  Example -- split:
    Input: "The food was cold and the wait was 40 minutes."
    Output: ["The food was cold.", "The wait was 40 minutes."]

  Example -- preserve as one fact:
    "The rating rose from 3.5 to 4.2 last month."
    "Recent reviews were more negative than earlier reviews."

6. Self-Contained:
Each fact must stand alone. Resolve all pronouns and implicit
references. Replace words like "it", "they", "this branch" with
explicit subjects.

7. Verifiability -- Review Domain Context:
Opinion-based claims about food, service, atmosphere, and pricing ARE
verifiable atomic facts because they can be checked against whether
the review data contains supporting evidence.

8. Preserve Claim Strength Exactly:
Do not weaken or strengthen the wording. Preserve quantifiers and
hedging exactly as written, including words such as "many", "most",
"some", "a few", "often", "rarely", "may", "appears", "seems".

  Example:
    "Many customers complained about slow service."
    must remain
    "Many customers complained about slow service."
    NOT "Customers complained about slow service."

9. No Inference or Expansion:
Extract ONLY what is explicitly stated. Do NOT infer unstated details
or add hidden assumptions.

  Example:
    "Overall sentiment is mixed." -> "The overall sentiment is mixed."
    NOT: "Some reviews were positive." / "Some reviews were negative."

Include as facts:
- Qualitative claims about food, service, atmosphere, pricing.
- Claims about what customers said or experienced.
- Numeric or statistical claims.
- Broad syntheses, overall conclusions, or pattern claims.

Place in non_verifiable_claims ONLY:
- Conversational filler (e.g., "Here is a summary of the reviews:").

---

Output format (strict JSON):
{
  "facts": ["Atomic fact 1", "Atomic fact 2"],
  "non_verifiable_claims": ["Filler statement 1"]
}

---

EXAMPLES:

Input: "Despite the friendly staff, the food was cold and the rating
dropped to 3.5."
```

```
Output:
{
  "facts": ["The staff was friendly.", "The food was cold.",
            "The rating dropped to 3.5."],
  "non_verifiable_claims": []
}

Input: "Here is a summary of the reviews. The food was always fresh
and the service was slow."
Output:
{
  "facts": ["The food was always fresh.", "The service was slow."],
  "non_verifiable_claims": ["Here is a summary of the reviews."]
}

---

Now decompose the following summary:

Summary:
{summary}
```

Listing 2: Atomic fact decomposition prompt.

D.3 Atomic Fact Verification Prompt

Used in Stage 2, once for each atomic fact produced by Stage 1. The judge receives the full source data and one atomic claim, then returns a Supported / Not Supported verdict. The placeholder `{source_data}` is replaced with the rendered review and statistics text; `{claim}` is replaced with one fact at a time.

```
You are a strict fact-checker.

The SOURCE DATA consists of raw customer reviews, and MAY also include
a section of pre-calculated statistical data at the top.

A claim may be supported by the reviews, the statistics, or both.
- Qualitative claims: verify against review text.
- Numeric/aggregated claims: verify against statistics if present,
  or by aggregating from raw review ratings if not.

DOMAIN CONTEXT:
Customer reviews are opinions. Different customers have different
experiences and naturally disagree.
- ALL qualitative claims about food, service, atmosphere, or pricing
  are opinion-grounded. They are supported if at least one review
  semantically entails the claim.
- "The food is good" and "Reviewers described the food as good" are
  equivalent and evaluated using the same standard.

SEMANTIC ENTAILMENT RULE:
Do NOT require exact wording. A claim is "Supported" if the source
data conveys the same underlying factual meaning or opinion.
  Example: "The staff was overwhelmed." is supported by "Waiters were
  running around in a panic and forgot our drinks."

CRITICAL VERIFICATION STANDARD:
Apply one of two standards based on the claim's wording:
```

```
PATTERN CLAIMS -- use language such as "multiple customers",
"consistently", "the majority", "many", "recurring". These require
evidence from MORE THAN ONE review. A single review is not enough.

ALL OTHER CLAIMS -- require at least one supporting review or
statistic. Contradicting evidence in other reviews reflects natural
disagreement and does not make the claim Not Supported.

Use the claim's actual wording to determine the standard; then use
semantic equivalence to find the evidence.

SOURCE DATA:
{source_data}

CLAIM:
"{claim}"

Task:
1. Identify which verification standard applies.
2. Search the source data for evidence using that standard.
3. State your reasoning in one concise sentence.
4. Return JSON.

Return JSON only:
{
  "reasoning": "One sentence explaining what evidence was found or
    why it is missing",
  "result": "Supported" or "Not Supported"
}

Rules:
- Return JSON only.
- Use exactly the keys: reasoning, result.
- Base the verdict only on the provided source data.
- Do not use prior knowledge.
- If support is missing, return Not Supported.
```

Listing 3: Atomic fact verification prompt.

D.4 Information Coverage Evaluation Prompt

Used in Stage 3, once for each golden insight per scenario. The judge receives the full generated summary and one golden insight, then returns a Covered / Not Covered verdict. The placeholder `{generated_summary}` is replaced with the full summary; `{golden_fact}` is replaced with one insight at a time.

```
You are an expert evaluator measuring information coverage in
AI-generated business summaries.

You will be given a Generated Summary and one Key Insight from a
verified golden dataset. Your task is to determine whether the Key
Insight is represented in the Generated Summary.

EVALUATION CRITERIA -- SEMANTIC MATCHING RULE:
1. Do NOT require exact wording or lexical matches.
2. A claim is "Covered" if the generated summary conveys the same
   underlying factual meaning or business insight as the golden
   insight, even if phrased differently.
3. A partial mention still counts as Covered if the core factual
```

```
   claim is present.
4. Only mark Not Covered if the claim is completely absent or
   contradicted.

Generated Summary:
"{generated_summary}"

Evaluation Steps:
1. Read the Key Insight carefully.
2. Scan the summary for evidence that conveys this insight semantically.
3. If the core meaning is present (even partially), mark Covered.
4. State your reasoning in one sentence, then give your verdict.

Return JSON only:
{
  "reasoning": "one sentence explaining your decision",
  "result": "Covered" or "Not Covered"
}

EXAMPLES:

Example 1 (Semantic equivalence):
  Key Insight: "Multiple customers explicitly complain the pasta was
    overly spicy."
  Summary: "Several diners noted that the spaghetti and penne
    contained far too much red chilli."
  Output: {"reasoning": "The summary semantically conveys the pasta
    spice complaint.", "result": "Covered"}

Example 2 (Partial coverage):
  Key Insight: "A single customer noted the outdoor seating was
    pleasant but complained that the tables were too close together."
  Summary: "One reviewer mentioned they enjoyed the outdoor patio."
  Output: {"reasoning": "The outdoor seating praise is present;
    the crowding complaint is omitted but the core claim is covered.",
    "result": "Covered"}

Example 3 (Not covered):
  Key Insight: "Customers consistently praise prompt and attentive
    table service."
  Summary: "Diners frequently complimented the food and the wine
    list."
  Output: {"reasoning": "The summary discusses food and wine but
    omits any mention of service quality.", "result": "Not Covered"}

Key Insight:
"{golden_fact}"
```

Listing 4: Information coverage evaluation prompt.

D.5 Attribution Faithfulness Evaluation Prompt

Used in Stage 4, once for each cited claim in a generated summary. The judge receives the full source data, one claim, and the cited review IDs or statistic fields, and evaluates only those citations without searching for new supporting evidence. The placeholders `{source_data}`, `{claim}`, `{review_ids}`, and `{stat_fields}` are replaced at runtime.

```
You are evaluating attribution fidelity for a generated summary.
```

```
You are given:
1. The FULL SOURCE DATA (reviews and statistics).
2. A CLAIM from the summary.
3. A set of CITED SOURCES (review IDs and/or statistical fields).

Your task:
Verify whether the cited sources actually support the claim.

Evaluation Steps:
1. Read the claim carefully.
2. For each cited review ID: check if the review text semantically
   entails the factual meaning of the claim. Mark valid or invalid.
3. For each cited stat field: check if the statistic semantically
   entails the factual meaning of the claim. Mark valid or invalid.
4. Do NOT add any new sources.
5. The claim is supported if at least one cited source is valid.

IMPORTANT RULES:
- DO NOT search for new sources.
- ONLY evaluate the provided cited sources.
- SEMANTIC ENTAILMENT: A source is valid if it semantically entails
  the exact factual meaning of the claim, even if phrased differently.
- STRICTNESS RULE: A source is invalid if it only mentions the general
  topic without entailing the specific fact. For example, "the food was
  bad" is NOT a valid citation for "the soup was cold" because the
  specific fact is not entailed.

SOURCE DATA:
{source_data}

Return JSON only:
{
  "valid_review_ids": [list of review IDs that support the claim],
  "valid_stat_fields": [list of stat fields that support the claim],
  "supported": true | false,
  "explanation": "one sentence explaining why the valid sources support
    or why none do"
}

CLAIM:
"{claim}"

CITED SOURCES:
Review IDs: {review_ids}
Stat Fields: {stat_fields}
```

Listing 5: Attribution faithfulness evaluation prompt.

D.6 Template-Constrained Output Schema

The Template-Constrained Generation strategy required the generator to produce valid JSON matching this fixed schema. The schema has five top-level sections corresponding to the four evaluation themes of the base task instruction (Section 7.2) plus an open-ended overflow field. The template was provided to the model as part of the prompt, while the API was configured to require a valid JSON object.

```
{
  "overall_sentiment": {
```

```
    "claims": [{"statement": "string",
                "sentiment": "Positive|Negative|Neutral|Mixed|Unknown",
                "review_ids": ["integer"]}]
  },
  "praise_theme": {
    "claims": [{"statement": "string",
                "review_ids": ["integer"]}]
  },
  "complaint_theme": {
    "claims": [{"statement": "string",
                "review_ids": ["integer"]}]
  },
  "specific_signal": {
    "claims": [{"statement": "string",
                "review_ids": ["integer"]}]
  },
  "other_insights": {
    "claims": [{"statement": "string",
                "aspect": "string (max 3 words)",
                "sentiment": "Positive|Negative|Neutral|Mixed",
                "review_ids": ["integer"]}]
  }
}
```

Listing 6: Template-Constrained output schema. Each `statement` field contains a natural-language claim; `review_ids` lists the review IDs cited as evidence.

D.7 RAG Keyword Extraction Prompt

The following prompt is not part of the evaluation pipeline. It is part of the RAG generation strategy.

The RAG strategy uses a two-step retrieval process. Before BM25 retrieval, the generator extracts keyword topics from the base task instruction. These topics are used to score all text-bearing reviews, and the top $k = 8$ reviews are selected. The user message passed alongside this system prompt is the base task instruction itself.

```
You are a keyword extraction assistant for a RAG system analysing
business reviews. Your goal is to bridge the gap between a user's
general question and specific details found in reviews. Given a user
question, identify a maximum of 8 key topics that cover all relevant
dimensions. For each topic, generate up to 6 concrete words that
customers would realistically use in a review. The keywords should
include:
  1. Synonyms (e.g., "food" -> "dish", "meal", "cuisine", "portion").
  2. Specific entities relevant to restaurants (e.g., "waiter",
     "server", "bartender", "kitchen").

Additional keyword rules:
- Prefer lowercase single-word keywords.
- Include common surface forms separately (e.g., "wait", "waited",
  "waiting"; "allergy", "allergic").
- Avoid abstract words unlikely to appear literally in reviews.
- Focus on words likely to maximize lexical overlap with review text.

Output only valid JSON:
{
  "topics": [
    {"topic": "string", "keywords": ["word1", "word2"]}
```

```
  ]
}
Return JSON only (no markdown or code fences).
```

Listing 7: RAG keyword extraction system prompt.

D.8 Shared Generation Contract Blocks

The shared system prompt described in Section 7 attaches three XML contract blocks after the role definition. Listing 8 reproduces these blocks verbatim as used by Plain Prompting and RAG. Aggregated-Statistics Grounding uses the same three blocks with wording adapted to also reference the statistical report (Section 7). Template-Constrained Generation replaces `<output_contract>` with a `<structured_output_contract>` enforcing the JSON schema and extends `<grounding_rules>` with an explicit instruction to state missing evidence as a claim rather than omitting it, also described in Section 7.

```
<output_contract>
- Provide a complete summary based on the reviews.
- Ensure all claims follow the citation rules.
</output_contract>

<grounding_rules>
- Every claim MUST be directly supported by the provided input.
- Do not use prior knowledge.
- If no supporting evidence exists, DO NOT produce the claim.
- Do not include unrelated or external information.
</grounding_rules>

<completeness_contract>
- Ensure the summary reflects the full range of information present in the
    reviews.
- Do not omit important themes or observations.
</completeness_contract>
```

Listing 8: Shared generation contract blocks (Plain Prompting and RAG).

E Per-Scenario Result Heatmaps

This appendix reports the per-scenario, per-strategy values behind the results in Section 8. Each heatmap gives the detailed cell-level breakdown for one metric, complementing the aggregate patterns discussed in the main text: factual accuracy (Figure 20), information coverage (Figure 21), citation precision (Figure 22), citation recall (Figure 23), fully supported claims ratio (Figure 24), and inference latency (Figure 25).

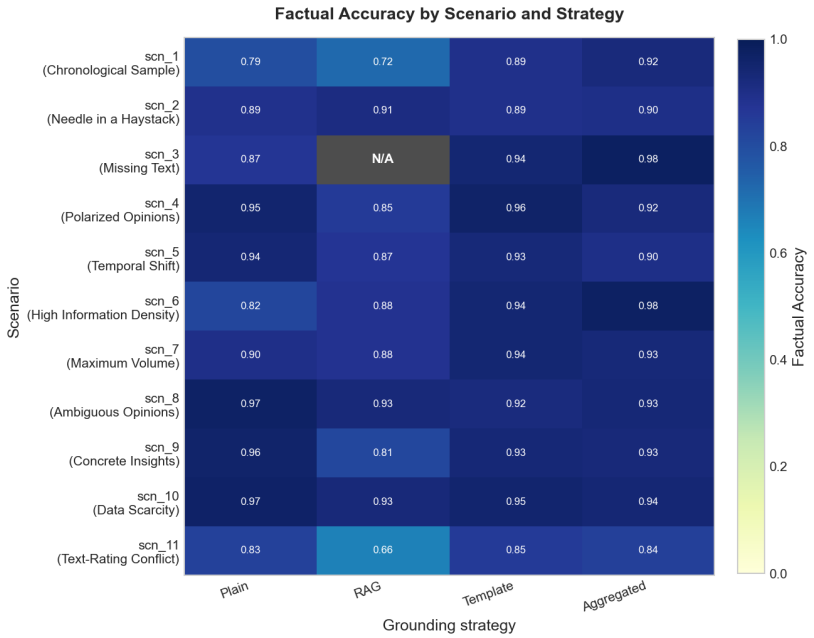

Figure 20: Factual Accuracy by scenario and strategy. RAG's weakest valid outputs appear in Scenario 1 (Chronological Sample) and Scenario 11 (Text-Rating Conflict). The N/A cell for RAG in Scenario 3 (Missing Text) reflects the missing output.

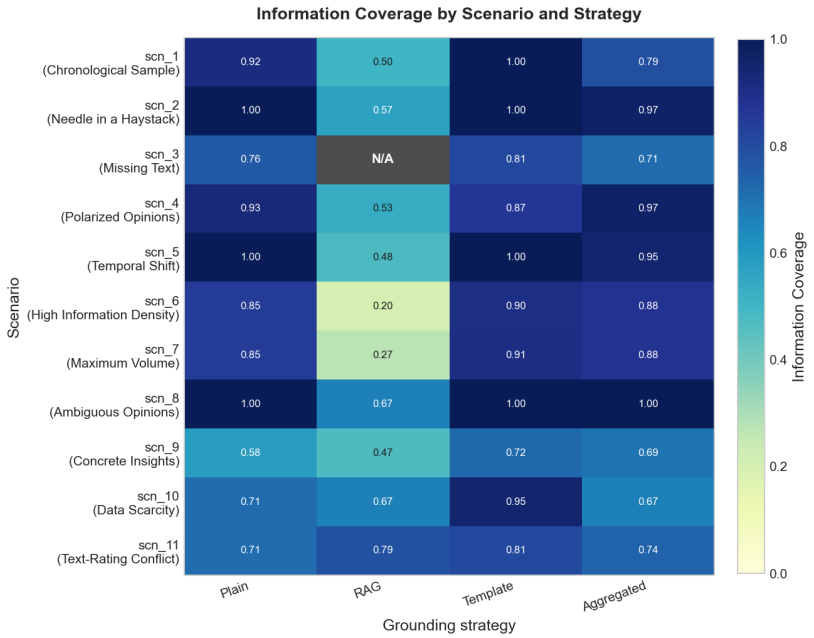

Figure 21: Information Coverage by scenario and strategy. RAG's coverage is lowest where important insights are spread across many distinct reviews.

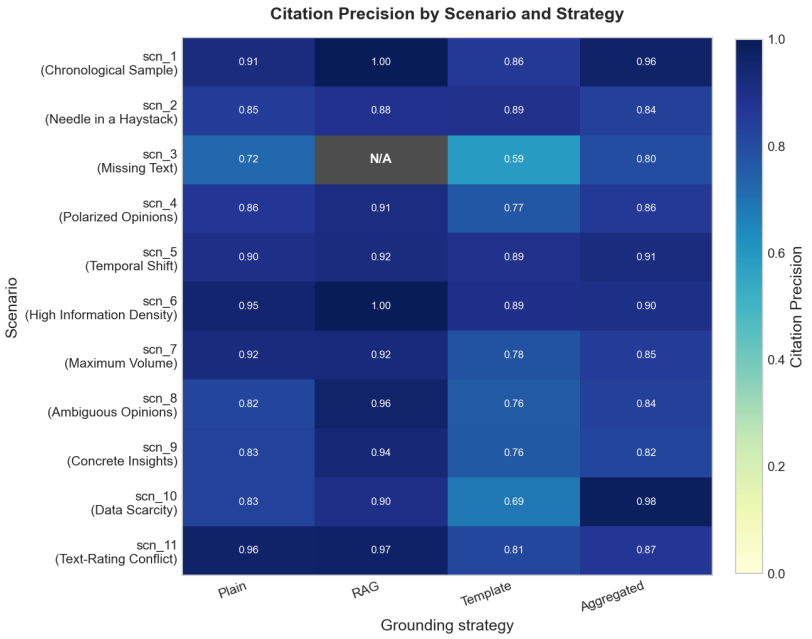

Figure 22: Citation Precision by scenario and strategy. RAG achieves perfect precision (1.00) in Scenario 1 (Chronological Sample) and Scenario 6 (High Information Density). Template-Constrained scores as low as 0.59 in Scenario 3 (Missing Text) and 0.69 in Scenario 10 (Data Scarcity).

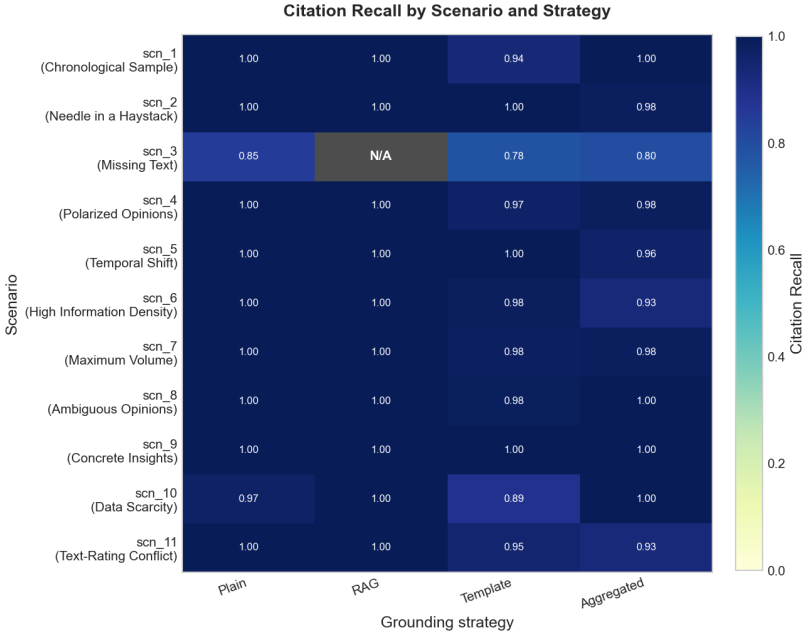

Figure 23: Citation Recall by scenario and strategy. Values are uniformly high (0.93–1.00) across almost all combinations. Template-Constrained shows the most sub-ceiling results in Scenario 3 (Missing Text), Scenario 6 (High Information Density), and Scenario 10 (Data Scarcity).

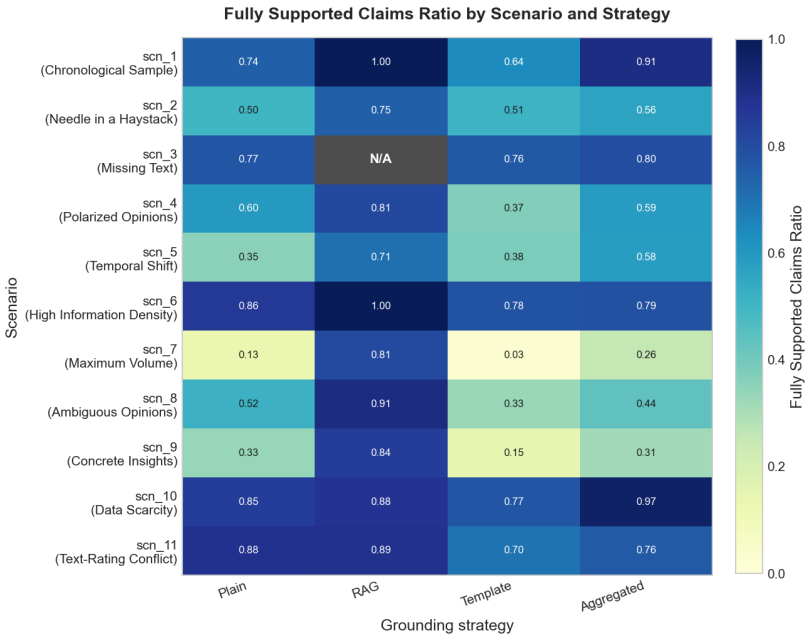

Figure 24: Fully Supported Claims Ratio by scenario and strategy. RAG achieves 1.00 in Scenario 1 (Chronological Sample) and Scenario 6 (High Information Density). Template-Constrained's worst results occur in Scenario 7 (Maximum Volume), where review volume is highest, and Scenario 9 (Concrete Insights), where each review carries dense specific detail.

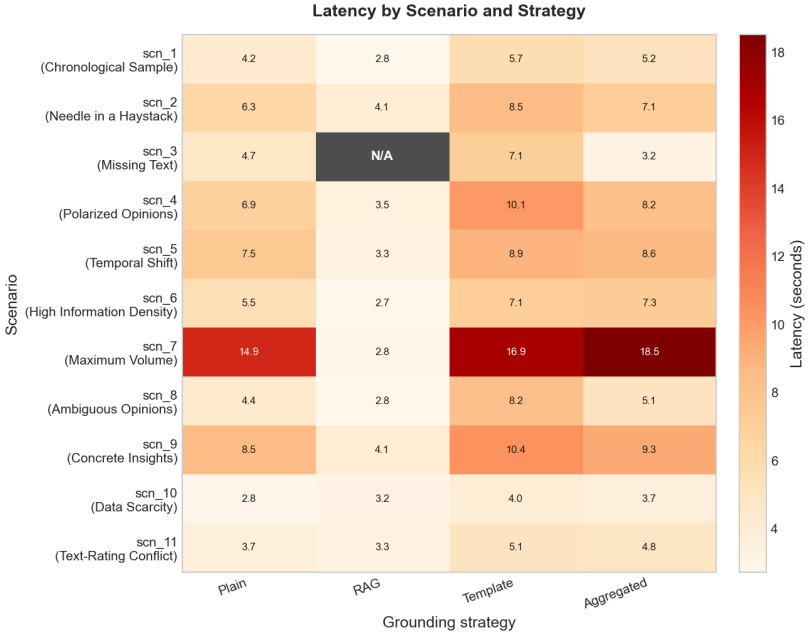

Figure 25: Inference Latency by scenario and strategy (seconds). Darker red indicates higher latency. Scenario 7 (Maximum Volume) is the most expensive for all strategies except RAG, which remains consistently low.

F Use of AI Tools

AI tools, including OpenAI's ChatGPT and the Prism tool, were used for LaTeX editing, proofreading, sentence restructuring, and readability improvements. The research idea, implementation, analysis, and final decisions are the author's own work. All AI-assisted content was reviewed and edited by the author, who takes full responsibility for the content of this thesis.