Chapter 4

Experimentation & Validation

The evaluation of retrieval-augmented generation (RAG) systems over structured knowledge graphs introduces a distinct set of challenges, especially when the task involves generating accurate and contextually grounded responses from data stored in formats such as the Internet Yellow Pages (IYP) graph. This chapter presents the experimental methodology, validation process, and performance analysis used to assess the proposed RAG model for the IYP domain. The overarching objective is to develop a reproducible, data-driven framework that measures the model's ability to interpret natural language inputs, retrieve relevant graph information, and produce high-quality responses.

Unlike open-domain systems, where evaluation often focuses on broad textual similarity, this work operates within a structured graph environment, where natural language queries must be translated into executable Cypher queries. As such, effective evaluation must consider both the syntactic accuracy of the generated queries and the semantic fidelity of the resulting outputs.

This chapter is organized into three key sections. Section 4.1 begins by detailing the creation of a custom validation dataset tailored to the IYP domain. Given the absence of publicly available benchmarks for this specific use case, an iterative process was followed, starting with synthetic data generation and culminating in the adoption and refinement of a curated dataset known as CypherEval [14]. This dataset includes a wide range of prompts, associated Cypher queries, and metadata necessary for structured evaluation.

Next, Section 4.2 describes the architecture of the experimental setup. The core RAG model incorporates multiple retrieval mechanisms through the LlamaIndex framework, combining symbolic and semantic retrieval components before passing the context to a generative language model. In parallel, a validation model is employed that uses ground truth Cypher queries from the validation dataset to retrieve ideal graph responses. These are then used as reference outputs to evaluate the performance of the RAG model under controlled conditions.

Finally, Section 4.3 presents a detailed analysis of model performance across several automated evaluation metrics, including BLEU, ROUGE, BERTScore, and G-Eval. These metrics offer complementary perspectives on output quality, ranging from surface-level overlap to semantic relevance. The results are further broken down by prompt complexity and domain type, offering insight into how the model performs across different levels of difficulty. Representative examples are included to illustrate key findings and highlight the strengths and weaknesses of different evaluation methods.

Together, these components provide a comprehensive framework for experimentation and validation, enabling both qualitative and quantitative insights into the capabilities and limitations of the proposed RAG system within the structured context of the IYP graph.

4.1 Dataset Setup

The development and evaluation of retrieval-augmented generation (RAG) models require a high-quality validation dataset to serve as a benchmark for performance assessment and model comparison. In particular, for applications involving structured knowledge graphs such as the Internet Yellow Pages (IYP) graph, a validation dataset must contain not only natural language prompts and corresponding answers, but also accurate Cypher queries that reflect the graph's structure and semantics. Such a dataset is essential for identifying model limitations, quantifying improvements, and ensuring the reproducibility of experimental results. However, no publicly available dataset exists that meets the specific requirements of the IYP domain, and prior approaches in the literature typically rely on simplified graphs or domain-agnostic methods that are not directly applicable to this context. As a result, a custom validation dataset had to be constructed to support the evaluation of the proposed system.

This section outlines the three-stage process followed to create this validation dataset.

Section 4.1.1 describes the initial approach of generating a synthetic dataset using ChatGPT. By providing the model with the IYP graph schema and a few example pairs of prompts and Cypher queries, an attempt was made to produce realistic validation examples. While this approach offered a rapid means of dataset creation, it was ultimately found to be inadequate due to inconsistencies, incorrect assumptions about the graph, and a lack of domain-specific understanding.

To address these shortcomings, the CypherEval dataset, introduced in Section 4.1.2, was acquired from the IYP team. CypherEval is a curated benchmark designed specifically to evaluate large language models on Cypher query generation for the IYP graph. It includes a diverse range of natural language prompts, corresponding Cypher queries, and difficulty annotations that classify each prompt along two dimensions: complexity (easy, medium, hard) and

type (general, technical). With two variants containing approximately 100 examples each, the dataset provides a robust foundation for model evaluation.

Finally, Section 4.1.3 details the post-processing of the CypherEval dataset to produce the final validation set. This involved filtering out examples that returned empty results on the reduced IYP graph, merging the dataset variants, and executing canonical queries to generate accurate ground truth responses. Each example was enriched with detailed metadata, including both generated and reference outputs, to support comprehensive evaluation. Although the primary evaluation focuses on comparing generated responses with ground truth responses, alternative evaluation strategies based on Cypher query structure or graph-level outputs are identified as future work.

4.1.1 ChatGPT-Generated Validation Dataset

In the early stages of model development, a significant challenge arose due to the absence of publicly available validation datasets tailored to the Internet Yellow Pages (IYP) graph. Specifically, there was no existing dataset that met the requirements of the Retrieval-Augmented Generation (RAG) pipeline or that accurately reflected the structure and semantics of the IYP knowledge graph. To overcome this limitation, an initial version of the validation dataset was generated using ChatGPT, a general-purpose large language model.

The dataset creation process was prompt-driven: ChatGPT was instructed to generate natural language questions along with corresponding Cypher queries executable on the IYP graph. To ensure that the generated queries adhered to the structure of the underlying graph, the full graph schema was provided as context. This schema included node labels, relationship types, and property names relevant to the IYP domain. In addition, a few example pairs of natural language questions and Cypher queries were included as part of the prompt to guide the model's understanding of the expected format and style.

The final version of this initial dataset consisted of 20 examples: 4 manually curated samples derived from the IYP Gallery and 16 ChatGPT-generated question-query pairs. Below are two representative examples from the ChatGPT-generated portion:

1. **Natural Language Query**:

   Which organization manages AS15169?

   Cypher Query:

```
MATCH (as:AS {asn: 15169}) -[:MANAGED_BY]->(org:Organization)
RETURN as, org
```

2. **Natural Language Query**:

   Which Internet Exchange Points (IXPs) is AS6453 a member of?

   Cypher Query:

```
MATCH (as:AS {asn: 6453}) -[:MEMBER_OF]->(ixp:IXP)
RETURN as, ixp
```

Despite these efforts to constrain and steer the generation process, the resulting dataset exhibited several limitations. Most notably, the quality and correctness of the generated queries varied significantly. The underlying issue was twofold: first, ChatGPT, while proficient in language generation, lacks specialized domain knowledge and cannot reason over the actual contents of a large, unseen knowledge graph. Second, without access to the complete instance-level data in the IYP graph, the model often made incorrect assumptions about the existence or structure of certain nodes and relationships. Consequently, many of the generated Cypher queries, though syntactically valid, did not reflect the true semantics or connectivity of the data.

These limitations highlighted the inherent difficulty of generating graph-specific query datasets without grounding the generation process in the full graph context. While the inclusion of schema information and few-shot examples provided a useful starting point, they were insufficient for producing high-quality queries at scale. As a result, the ChatGPT-generated dataset was ultimately deemed inadequate for reliable validation purposes.

To address this issue, contact was made with the IYP development team to inquire about the availability of internal resources. This outreach resulted in the acquisition of the CypherEval dataset, a curated collection of validated Cypher queries and corresponding natural language questions developed specifically for the IYP graph. This dataset, which is described in the following section, offered a more robust and accurate foundation for validation and evaluation of the RAG model.

4.1.2 CypherEval

To address the limitations encountered with the ChatGPT-generated dataset, a more robust and domain-specific alternative was required. This need was met through the CypherEval [14] dataset, a curated benchmark specifically designed to evaluate the performance of large language models (LLMs) in generating Cypher queries from natural language prompts on the Internet Yellow Pages (IYP) knowledge graph.

CypherEval consists of three primary components: a natural language Prompt, which represents a user's query; the corresponding Canonical Solution, a Cypher query that accurately

retrieves the intended information from the graph; and a Difficulty label, which annotates each example with a classification of its complexity. The difficulty annotation serves a dual purpose: it captures both the level of difficulty, categorized as easy, medium, or hard, and the type of the question, which is either general or technical. This additional metadata enables more fine-grained evaluations of model performance across distinct dimensions of query complexity and abstraction.

The dataset is provided in two distinct variants, each containing approximately 100 examples, resulting in a total of 200 samples. These examples span a diverse range of query structures and use cases, making the dataset sufficiently comprehensive for the scope of this research. Given their quality and variety, these 200 examples are used as the primary benchmark for evaluating the RAG model's performance in generating responses as well as in question-to-Cypher translation.

Below are two representative examples from the CypherEval dataset:

1. **Difficulty Level**:

   Easy general prompt

   Prompt:

   Give me the names of the IXPs where AS2497 is member.

   Canonical Solution:

```
MATCH (:AS {asn: 2497}) -[:MEMBER_OF]->(ixp:IXP)
RETURN DISTINCT ixp.name
```

2. **Difficulty Level**:

   Hard technical prompt

   Prompt:

   Find all the AS nodes that have peer to peer relationship with the AS with asn 2497.

   Canonical Solution:

```
MATCH (:AS {asn:2497}) -[:PEERS_WITH {rel: 0}]-(a:AS)
RETURN a
```

The availability of difficulty annotations is particularly valuable for diagnostic purposes, as it allows for targeted analysis of a model's strengths and weaknesses under varying levels of prompt complexity. For instance, as discussed in Section 4.3.3, this categorization was used to assess how the retrieval and generation components of the RAG model perform when exposed

to general versus technical prompts, as well as to queries of increasing syntactic and semantic complexity.

In contrast to the synthetic dataset generated via ChatGPT, CypherEval offers a higher degree of accuracy, coverage, and relevance. Its examples are aligned with the actual structure and vocabulary of the IYP graph, and the Cypher queries have been validated to ensure correctness. As such, it serves not only as a reliable benchmark for model evaluation but also as a guiding reference for future work involving question-to-Cypher translation in the IYP domain.

4.1.3 Final Validation Dataset Construction

Following the acquisition of the CypherEval dataset, a series of preprocessing steps were performed to construct the final validation dataset used for model evaluation. The goal of this process was to ensure that all examples in the dataset were executable against the reduced version of the IYP graph and would yield meaningful, non-empty results, thereby providing reliable evaluation signals.

The first step involved filtering both variants of the CypherEval dataset to retain only the examples whose canonical Cypher queries produced non-empty results when executed against the reduced IYP knowledge graph. This filtering was essential to avoid including cases where a query, although syntactically and semantically correct, did not correspond to any existing data in the deployed graph version. After filtering, the two dataset variants, each originally containing approximately 100 examples, were merged to form a single, unified dataset.

This final dataset was then used as the benchmark for evaluating the RAG model. For each example, the model was run end-to-end on the user prompt. To define a reference output for comparison, the canonical Cypher query provided by CypherEval was executed directly against the IYP graph using the validation model described in Section 4.2.2. The resulting graph output was then passed to ChatGPT for natural language generation, producing a ground truth response that serves as the gold standard for evaluation. This approach enabled a fair and consistent comparison between model-generated outputs and the expected system behavior.

The final validation dataset consists of 190 examples. Each example in the final validation dataset retains the following metadata and outputs. An example value is given for each of its fields:

- **Prompt**: The user's natural language question to the model.

  Example: Give me the names of the IXPs where AS2497 is member.

- **Difficulty**: The complexity label provided by the CypherEval dataset, including both difficulty level (Easy, Medium, Hard) and question type (General, Technical).

Example: Easy general prompt

- **Ground Truth Cypher Query**: The canonical Cypher query associated with the prompt, as specified in the CypherEval dataset.

  Example (Cypher):

```
MATCH (:AS {asn: 2497})-[:MEMBER_OF]->(ixp:IXP)
RETURN DISTINCT ixp.name
```

- **Ground Truth Cypher Query Source**: The origin of the Cypher query (e.g., CypherEval, ChatGPT, IYP GitHub).

  Example: CypherEval

- **RAG Version**: The specific version or configuration of the RAG architecture used for response generation.

  Example: v1.0

- **Generated Cypher Query**: The Cypher query produced by the RAG model in response to the input prompt.

  Example (Cypher):

```
MATCH (a:AS {asn: 2497})-[:MEMBER_OF]->(i:IXP)
RETURN i.name
```

- **Generated Response**: The final natural language answer generated by the RAG model.

  Example: The names of the IXPs where AS2497 is a member are 'NYIIX New York', 'DE-CIX Frankfurt', 'DIX-IE', 'NSPIXP6', 'SIX', and 'SIX Seattle'.

- **Intermediate Cypher Query Output**: The structured result returned from Neo4j for the model-generated Cypher query, used internally by the RAG pipeline.

  Example (JSON):


```
[{"i.name":"NYIIX New York"},{"i.name":"DE-CIX Frankfurt"},{"i.
    name":"DE-CIX Frankfurt"},{"i.name":"DE-CIX Frankfurt"},{"i.
    name":"DE-CIX Frankfurt"},{"i.name":"NYIIX New York"},{"i.
    name":"DIX-IE"},{"i.name":"NSPIXP6"},{"i.name":"SIX"},{"i.
    name":"SIX Seattle"}]
```


- **Neo4j Cypher Query Output**: The raw response to the generated Cypher query as retrieved from the Neo4j browser.

  Example (JSON):

```
[{"i.name":"NYIIX New York"},{"i.name":"DE-CIX Frankfurt"},{"i.name":"DE-CIX Frankfurt"},{"i.name":"DE-CIX Frankfurt"},{"i.name":"DE-CIX Frankfurt"},{"i.name":"NYIIX New York"},{"i.name":"DIX-IE"},{"i.name":"NSPIXP6"},{"i.name":"SIX"},{"i.name":"SIX Seattle"},{"i.name":"Network Service Provider IXP-6"},{"i.name":"SGIX"},{"i.name":"SGIX"},{"i.name":"Equinix New York"},{"i.name":"JPNAP Tokyo"},{"i.name":"Equinix San Jose"},{"i.name":"Equinix Los Angeles"},{"i.name":"Equinix Palo Alto"},{"i.name":"HKIX"},{"i.name":"HKIX"},{"i.name":"Equinix Ashburn"},{"i.name":"Equinix Singapore"},{"i.name":"DE-CIX Frankfurt"},{"i.name":"LINX LON1"},{"i.name":"Equinix Hong Kong"},{"i.name":"JPNAP Osaka"},{"i.name":"Equinix IBX San Jose"},{"i.name":"Equinix Hong Kong"},{"i.name":"JPNAP Tokyo"},{"i.name":"DE-CIX Frankfurt"},{"i.name":"SIX Seattle"},{"i.name":"DIX-IE"},{"i.name":"SGIX"},{"i.name":"Equinix Singapore"},{"i.name":"HKIX"},{"i.name":"JPNAP Osaka"},{"i.name":"LINX LON1"},{"i.name":"Equinix New York"},{"i.name":"Equinix San Jose"},{"i.name":"Equinix Palo Alto"},{"i.name":"Equinix Ashburn"},{"i.name":"Equinix Los Angeles"}]
```


- **Ground Truth Generated Response**: The natural language output generated by the RAG model using the canonical Cypher query as input.

  Example: The names of the IXPs where AS2497 is a member are: 'NYIIX New York', 'DE-CIX Frankfurt', 'DIX-IE', 'NSPIXP6', 'SIX', 'SIX Seattle', 'Network Service Provider IXP-6', 'SGIX', 'Equinix New York', and 'JPNAP Tokyo'.

- **Ground Truth Intermediate Cypher Query Output**: The Neo4j output for the canonical Cypher query used as intermediate input in the RAG pipeline.

  Example (JSON):


```
[{"ixp.name":"NYIIX New York"},{"ixp.name":"DE-CIX Frankfurt"},{"ixp.name":"DIX-IE"},{"ixp.name":"NSPIXP6"},{"ixp.name":"SIX"},{"ixp.name":"SIX Seattle"},{"ixp.name":"Network Service Provider IXP-6"},{"ixp.name":"SGIX"},{"ixp.name":"Equinix New York"},{"ixp.name":"JPNAP Tokyo"}]
```


- **Ground Truth Neo4j Cypher Query Output**: The raw Neo4j response to the canonical Cypher query, captured via the Neo4j browser interface.

  Example (JSON):

```
[{"ixp.name":"NYIIX New York"},{"ixp.name":"DE-CIX Frankfurt"},{"ixp.name":"DIX-IE"},{"ixp.name":"NSPIXP6"},{"ixp.name":"SIX"},{"ixp.name":"SIX Seattle"},{"ixp.name":"Network Service Provider IXP-6"},{"ixp.name":"SGIX"},{"ixp.name":"Equinix New York"},{"ixp.name":"JPNAP Tokyo"},{"ixp.name":"Equinix San Jose"},{"ixp.name":"Equinix Los Angeles"},{"ixp.name":"Equinix Palo Alto"},{"ixp.name":"HKIX"},{"ixp.name":"Equinix Ashburn"},{"ixp.name":"Equinix Singapore"},{"ixp.name":"LINX LON1"},{"ixp.name":"Equinix Hong Kong"},{"ixp.name":"JPNAP Osaka"},{"ixp.name":"Equinix IBX San Jose"}]
```


For the purposes of evaluation, the primary focus is on comparing the *Generated Response* with the *Ground Truth Generated Response*, as this captures the end-to-end effectiveness of the RAG model in answering natural language questions over the IYP graph. While alternative evaluation strategies, such as comparing the *Generated Cypher Query* with the *Ground Truth Cypher Query* or analyzing the structural similarity between the query outputs at the graph level, are possible and potentially valuable, they were not explored within the scope of this work. These approaches are identified as promising directions for future research, particularly for gaining deeper insights into intermediate model behavior and Cypher generation accuracy.

4.2 Experiment Setup

This section outlines the experimental framework employed to evaluate the Retrieval-Augmented Generation (RAG) model developed for interacting with the Internet Yellow Pages (IYP) knowledge graph. The purpose of the experimental setup is to assess the model's ability to accurately interpret natural language queries, retrieve relevant graph information, and generate coherent, context-aware responses. The evaluation strategy involves configuring two complementary models and applying a suite of automated metrics to compare generated outputs against reference responses.

The evaluation process for evaluating the ChatIYP pipeline across its retrieval and generation methodology stages as depicted in Figure 4.1 proceeds as follows. The validation dataset, which includes natural language prompts and corresponding ground truth Cypher queries, is used as the basis for all experiments. Each prompt is first processed by the RAG model, which generates both a candidate Cypher query and a corresponding natural language response. In parallel, the validation model uses the ground truth Cypher query to retrieve accurate information from the knowledge graph and generates a reference response using the same language model. This ensures that the validation output reflects the ideal response grounded in verified data. After both models have produced responses for all prompts in the dataset, a suite of evaluation

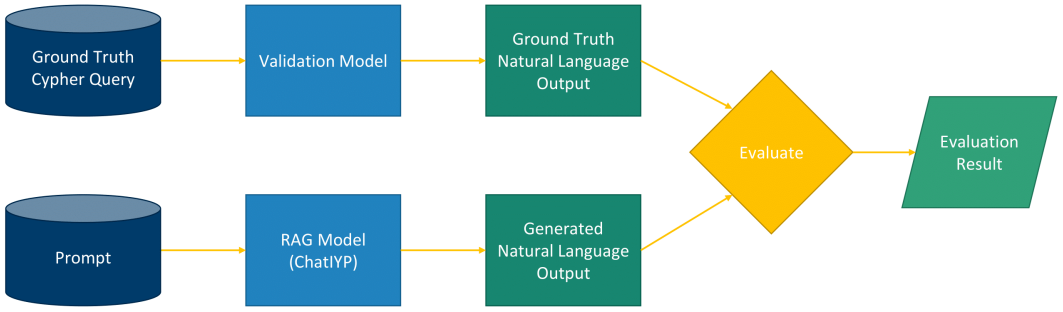

FIGURE 4.1: Experimental setup for evaluating the ChatIYP pipeline across its retrieval and generation methodology stages. Given the same natural language input, a validation model executes the ground truth Cypher query to produce a reference response, while ChatIYP performs its full RAG process to generate an answer. The two outputs are then compared to assess the correctness and quality of ChatIYP's response.

metrics is applied to compare the RAG-generated outputs against the reference responses from the validation model. These comparisons provide a comprehensive assessment of the RAG model's performance in terms of relevance, correctness, and completeness.

Section 4.2.1 introduces the primary RAG model configuration. This model integrates three retrieval components from the LlamaIndex framework, namely the *TextToCypherRetriever*, *VectorContextRetriever*, and *LLMReranker*, which work in tandem to extract both symbolic and semantic information from a curated subgraph of the IYP dataset (as defined in Section 4.1.3). The outputs of these retrievers, along with the user's original natural language query, are passed to a generative language model, GPT-3.5-Turbo, which produces the final response.

To enable consistent and accurate evaluation, a validation model is introduced in Section 4.2.2. Unlike the RAG model, the validation model does not perform query generation. Instead, it utilizes the ground truth Cypher queries provided in the Validation Dataset (Section 4.1.3) to retrieve the correct information from the graph. These results are then passed to the same generative model used in the RAG pipeline, ensuring that the validation outputs represent ideal responses grounded in verified data. These outputs serve as the gold standard for subsequent evaluation.

Section 4.2.3 describes the methodology used to evaluate the model's responses. Four metrics are applied: BLEU, ROUGE, BERTScore, and G-Eval, as introduced in Section 2.2.4.2. The first three are conventional similarity metrics computed directly between the generated and

ground truth responses. In contrast, G-Eval is an LLM-based evaluator that requires prompt engineering to guide the model in assessing responses along dimensions such as relevance, factual correctness, and completeness. Evaluation prompts are carefully constructed to ensure consistency and reliability in G-Eval's scoring behavior.

Together, these components constitute a robust and reproducible experiment pipeline. The RAG and validation models provide complementary outputs that allow for precise, controlled comparisons, while the evaluation metrics offer both surface-level and semantic perspectives on model performance. The results of this experimental setup are presented and analyzed in the following section.

4.2.1 Model Setup

To enable natural language interaction with the Internet Yellow Pages (IYP) knowledge graph, a Retrieval-Augmented Generation (RAG) model was developed and deployed over a curated subgraph of the IYP dataset, as detailed in Section 3.1. The RAG architecture combines multiple retrieval strategies with a generative language model, allowing it to retrieve relevant information from the graph and compose coherent, context-aware responses in natural language.

The retrieval phase of the model integrates three core components from the LlamaIndex framework, each described in detail in Section 3.3.1:

- The **TextToCypherRetriever**, which translates a natural language query into an executable Cypher query aimed at extracting precise graph data;
- The **VectorContextRetriever**, which leverages dense embedding similarity to retrieve semantically relevant nodes and their context from the graph;
- The **LLMReranker**, which reorders the retrieved results based on relevance using a cross-encoder model, ensuring that the most pertinent information is prioritized for the generation step.

Following the retrieval phase, the generation component uses **GPT-3.5-Turbo** to synthesize the final output. As outlined in Section 3.3.2, this model conditions its response on both the original user query and the retrieved content. The generative model is responsible for producing fluent, informative answers that include both the Cypher query output and summarize the information retrieved.

The system is designed to accept a user's natural language question as input. Upon receiving a query, the retrievers work in tandem to surface relevant entities, relationships, and paths from the IYP subgraph, while also generating a Cypher query that formalizes the user's intent. This

intermediate representation and context-aware information are then passed to the generative model, which produces the final response.

By combining symbolic query generation with semantic retrieval and neural generation, this RAG pipeline offers a robust approach for accessing structured graph data through natural language, thereby enhancing the accessibility and usability of the IYP knowledge base.

4.2.2 Validation Model

To establish a reliable benchmark for evaluating the performance of the RAG model, a validation model was constructed to generate reference answers based on the ground truth Cypher queries provided in the Validation Dataset (Section 4.1.3). This model is purposefully constrained in order to isolate the generation quality from retrieval variability and ensure alignment with the canonical Cypher solutions included in the dataset.

Unlike the full RAG model, the validation model employs only a single retrieval component: the **TextToCypherRetriever**. However, in this configuration, the retriever does not perform query generation. Instead, it bypasses the query formulation step entirely and utilizes the *Ground Truth Cypher Query* associated with each validation sample. This query is executed against the IYP subgraph to retrieve the exact result set that the user's prompt is expected to produce.

The retrieved results are then passed directly to the generation component—again, GPT-3.5-Turbo—along with the original natural language query. By constraining the input to the generation model in this way, the validation model produces a deterministic, reference output that reflects the intended response for each user query, grounded in the verified Cypher query from the dataset.

This setup plays a crucial role in the evaluation pipeline. Since the outputs of the validation model are conditioned on correct Cypher queries and accurate graph data, they serve as the *gold standard* against which the responses generated by the RAG model are compared. In particular, this reference output allows for a consistent and fair assessment of the RAG system's end-to-end performance, with respect to both factual accuracy and linguistic coherence.

By decoupling retrieval uncertainty from generation quality, the validation model enables precise diagnostic analysis, especially when evaluating how well the RAG model approximates the target response across varying prompt types and complexity levels, as explored in the next Section 4.3.

4.2.3 Application of Validation Metrics

To quantify the performance of the RAG model described in Section 4.2.1, a series of automatic evaluation metrics were applied to its outputs on the Validation Dataset (Section 4.1.3). These outputs, referred to as the *Generated Responses*, were assessed against the corresponding *Ground Truth Responses* produced by the validation model described in Section 4.2.2. The objective of this evaluation is to measure the extent to which the RAG model is able to replicate the ideal response, both in content and in form.

Four evaluation metrics were employed for this purpose: **BLEU**, **ROUGE**, **BERTScore**, and **G-Eval**, each introduced in Section 2.2.4.2. The first three, BLEU, ROUGE, and BERTScore, are applied in a straightforward manner: the metric is computed between the generated and ground truth responses, yielding a numerical score that reflects similarity. These metrics capture different aspects of text overlap: BLEU emphasizes precision of n-gram matches, ROUGE prioritizes recall of key phrases, and BERTScore computes semantic similarity using contextual embeddings.

The fourth metric, **G-Eval**, requires a more nuanced setup. As a large language model-based evaluator, G-Eval does not rely solely on lexical similarity but instead assesses the quality of a response from multiple linguistic and semantic dimensions. In order to guide the LLM toward consistent and meaningful evaluations, a carefully designed prompting strategy was employed. The prompt provides step-by-step instructions for evaluating key aspects of response quality, including:

- **Relevance**: Whether the generated response appropriately addresses the user's question;
- **Factual Correctness**: Whether all presented facts are consistent with the graph-derived answer;
- **Completeness**: Whether important details from the ground truth are omitted or misrepresented;
- **Coherence and Fluency**: Whether the response is well-structured and grammatically sound.

These instructions ensure that the G-Eval model performs holistic evaluation rather than relying solely on surface-level similarity. Each evaluated pair, Generated Response and Ground Truth Response, is passed to G-Eval along with the evaluation criteria embedded in the prompt, and the model returns a scalar score along with optional qualitative feedback.

Together, these four metrics offer a multi-faceted view of model performance. While BLEU, ROUGE, and BERTScore provide standardized, reproducible scores for large-scale benchmarking, G-Eval adds an LLM-informed perspective that is more aligned with human judgments of quality and correctness. The results obtained from applying these metrics are presented and analyzed in Section 4.3.

4.3 Results

This section presents a comprehensive analysis of the RAG model's performance on the IYP dataset, structured in two parts to provide both a metric-centric and difficulty-aware perspective. The goal is to evaluate not only how well the model performs but also how the chosen evaluation metrics and prompt characteristics affect our understanding of model behavior.

Before delving into the detailed analysis, a dedicated subsection, *Representative Examples for Metric Evaluation* in Section 4.3.1, introduces two illustrative examples drawn from the IYP dataset. These examples are referenced throughout the Results section to provide concrete, qualitative insight into how different evaluation metrics respond to specific types of model outputs, such as factual errors or informative elaborations. By grounding the discussion in these real examples, the analysis aims to enhance interpretability and support more nuanced comparisons between metrics.

The first part, *IYP Dataset Results Analysis* in Section 4.3.2, focuses on a comparative evaluation of several automated scoring metrics, BLEU, ROUGE-L, BERTScore, and G-Eval, used to assess the quality of generated answers. These metrics are categorized based on their methodological foundations: BLEU and ROUGE represent traditional statistical approaches, G-Eval is model-based, and BERTScore functions as a hybrid by leveraging both pretrained embeddings and cosine similarity. By contrasting these metrics, we highlight key differences in their sensitivity to semantic fidelity, factual grounding, and linguistic overlap, ultimately motivating the selection of G-Eval as the primary metric for subsequent evaluations.

The second part, *Prompt Complexity Evaluation* in Section 4.3.3, builds on these findings by examining how prompt difficulty influences model performance. Prompts in the IYP dataset are stratified by difficulty (Easy, Medium, Hard) and domain type (Technical, General), allowing for a detailed analysis of how complexity impacts response quality. G-Eval and ROUGE-L scores are analyzed across these dimensions, revealing clear trends in model capability degradation as task complexity increases. This difficulty-aware perspective provides critical insights into where the model struggles most, particularly with hard technical prompts, and underscores the importance of refining evaluation strategies and model design to better handle real-world, high-complexity inputs.

Together, these analyses offer a multidimensional view of performance, shedding light on both the effectiveness of different evaluation metrics and the challenges posed by varying levels of prompt difficulty.

4.3.1 Representative Examples for Metric Evaluation

To support a clearer understanding of how various evaluation metrics respond to different types of model behavior, two representative examples from the IYP dataset are presented. These examples are referenced throughout the analysis to illustrate key distinctions in metric sensitivity to factual correctness, semantic alignment, and surface-level similarity.

Example 1 — Numerical Factual Error:

- **Prompt**: Find the total number of distinct AS in the Country with country_code 'US' as specified by the reference_org NRO.
- **Ground Truth Response**: The total number of distinct AS in the Country with the country_code 'US' as specified by the reference_org NRO is 30827.
- **Generated Response**: The total number of distinct AS in the Country with the country_code 'US' as specified by the reference_org NRO is 0.

This example illustrates a critical factual error in the generated response, despite its close lexical resemblance to the ground truth. Evaluation metrics based on lexical overlap, such as BLEU and ROUGE-L, may assign relatively high scores due to the structural similarity, failing to reflect the semantic inaccuracy. In contrast, BERTScore and G-Eval, which incorporate semantic understanding or model-based judgment, are more likely to penalize this type of factual inconsistency.

Example 2 — Extended but Correct Answer:

- **Prompt**: Find the Country of the Prefixes originated by AS node with asn 2497. Return the country and prefixes nodes.
- **Ground Truth Response**: The prefixes originated by AS node with ASN 2497 belong to Japan and the United States of America.
- **Generated Response**: The query results show the countries and prefixes associated with the AS node having ASN 2497. The countries are Japan and the United States of America, while the prefixes originate from these countries include IPv4 and IPv6 addresses such as 157.8.16.0/23, 240a:71:9000::/36, 2001:48b0::/32, and others.

In this case, the generated response is factually accurate and semantically aligned with the ground truth but includes additional information not explicitly present in the reference. Lexical metrics may penalize the response for its verbosity or lack of direct overlap, whereas model-aware metrics such as G-Eval are better positioned to recognize the validity and informativeness of the extended content.

These examples serve as qualitative anchors throughout the evaluation. They help demonstrate how each metric captures, or fails to capture, important aspects of answer quality, particularly in the context of knowledge-intensive tasks. As such, they provide essential context for interpreting metric-based results presented in the subsequent analyses.

4.3.2 IYP Dataset Results Analysis

This subsection presents a comprehensive evaluation of the RAG model's performance on the IYP dataset using a diverse suite of automatic evaluation metrics, each aligned with a distinct methodological category. BLEU and ROUGE represent statistical metrics, relying on n-gram precision and sequence-based recall to quantify lexical overlap between generated and reference responses. BERTScore is classified as a hybrid metric: it combines statistical structure with semantic analysis by computing cosine similarity over contextual embeddings derived from a pretrained language model (typically BERT). This fusion enables BERTScore to capture both surface-level and deeper semantic similarities. In contrast, G-Eval belongs to the class of model-based metrics, employing a large language model to directly assess the factual correctness, completeness, and coherence of generated outputs. This approach allows for nuanced judgments that go beyond token-level comparisons.

Given the fact-sensitive and graph-grounded nature of the IYP question answering task, traditional statistical metrics often fall short in capturing the semantic and factual fidelity of model outputs. As such, this analysis leverages all four metrics to construct a multi-faceted understanding of the RAG model's capabilities and limitations. The subsequent sections explore the scoring distributions, illustrative examples, and inter-metric correlations, ultimately motivating the adoption of G-Eval as the most informative and domain-appropriate evaluation metric.

4.3.2.1 BLEU Score Analysis

To assess the performance of the RAG model on the custom IYP dataset, the BLEU (Bilingual Evaluation Understudy) metric was employed to measure the lexical similarity between generated responses and ground truth answers. The distribution of BLEU scores, shown in Figure 4.2, provides several insights into the model's behavior. In this histogram, the x-axis represents BLEU score values ranging from 0 to 1 split into discrete score bins (e.g., 0.0–0.2, 0.2–0.4,

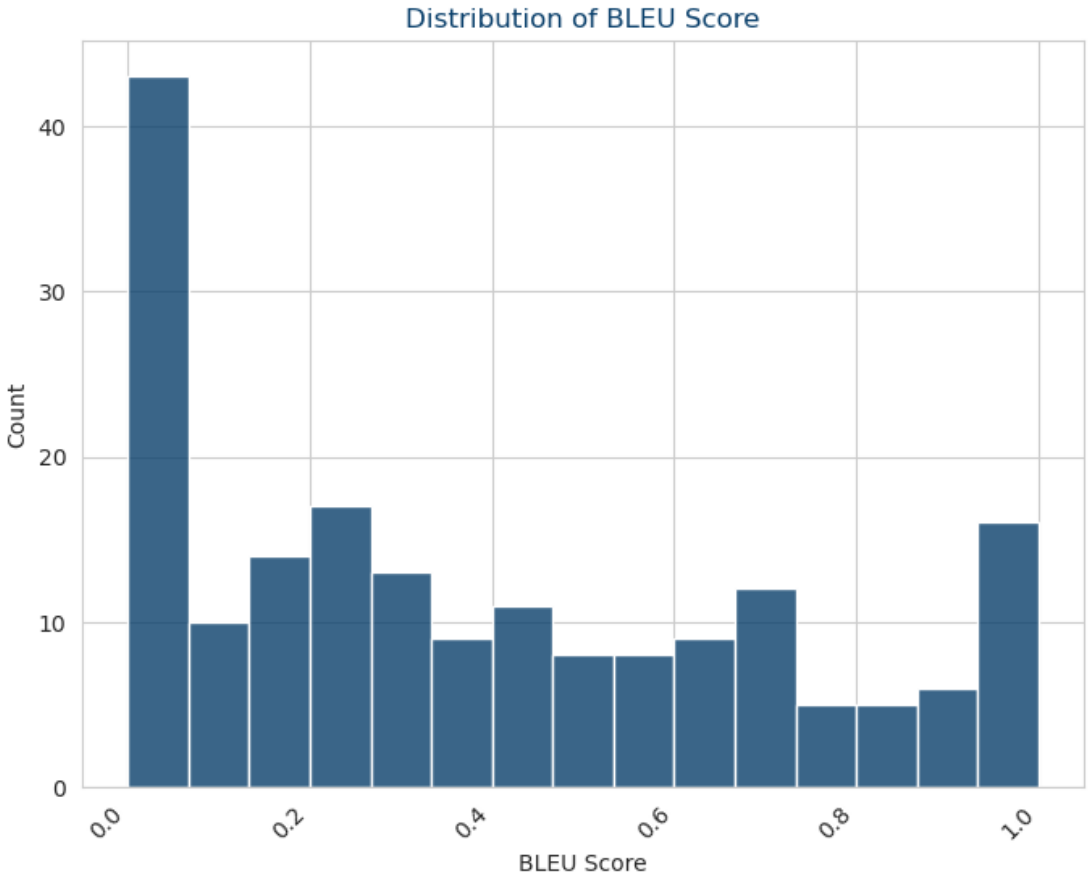

FIGURE 4.2: Histogram of BLEU scores for responses generated by the RAG model on the custom IYP dataset. The x-axis represents score intervals from 0 to 1, indicating the degree of lexical overlap with ground truth answers, while the y-axis shows the number of responses falling into each interval.

up to 1.0), where 0 indicates no lexical overlap with the reference and 1 indicates an exact match. The y-axis shows the number of examples within each score interval. The distribution appears relatively flat, indicating no clear trend toward higher or lower quality outputs. However, a prominent spike at a BLEU score of 0.0 reveals that a considerable number of generated responses were entirely incorrect or bore little resemblance to the expected answers.

Although a number of examples achieved a BLEU score of 1.0, implying a perfect lexical match with the ground truth, such outcomes must be interpreted with caution. High BLEU scores do not necessarily reflect factual correctness. For example, one generated response stated that the number of Autonomous Systems (AS) in the US was "0," while the reference answer was "30827." Despite this contradiction, the BLEU score was 0.95 due to the strong structural similarity between the two sentences. This case highlights a fundamental limitation of BLEU: its inability to detect semantic errors, particularly in numerical or factual content.

On the other hand, low BLEU scores are not always indicative of incorrect or poor responses. In several instances, the model produced outputs that were accurate and even enriched with

additional relevant details, yet received low BLEU scores due to differences in phrasing. For example, a response that correctly identified the countries ”Japan” and ”United States of America” and included specific IP prefixes scored only 0.12, simply because the added information and alternative wording reduced its n-gram overlap with the reference.

In conclusion, while BLEU provides a convenient way to quantify surface-level similarity, it proves to be an unsuitable metric for evaluating the quality of responses in this graph-grounded, fact-sensitive question answering task. Its inability to capture semantic correctness, factual accuracy, and the effects of valid rephrasing significantly limits its effectiveness.

4.3.2.2 ROUGE Score Analysis

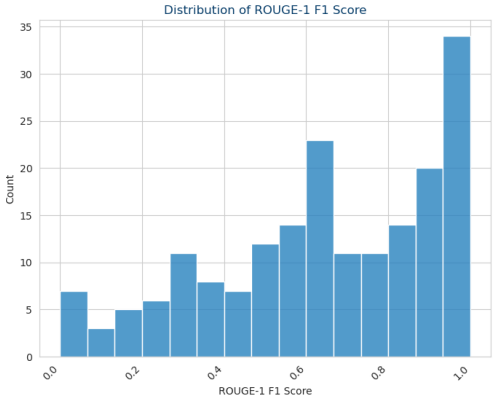

FIGURE 4.3: Histogram of ROUGE-1 scores for the RAG model's responses on the IYP dataset. The x-axis shows score intervals from 0 to 1, representing unigram overlap with reference answers, while the y-axis indicates the number of responses per interval.

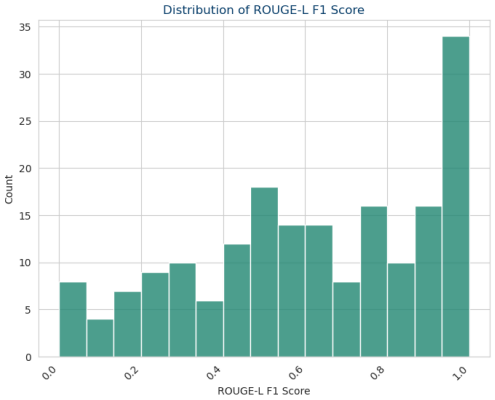

FIGURE 4.4: Histogram of ROUGE-L scores, capturing the longest common subsequence between generated and reference responses. The distribution closely resembles that of ROUGE-1, indicating consistent trends across different lexical similarity metrics.

In addition to the BLEU analysis, the evaluation of the RAG model on the IYP dataset included the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric to further assess lexical similarity. Both ROUGE-1 and ROUGE-L variants were initially computed. As illustrated in Figures 4.3 and 4.4, the distributions of these two metrics are highly similar. This similarity is also evident in the corresponding scatter plot (Figure 4.5), where a consistent alignment between ROUGE-1 and ROUGE-L scores can be observed. Based on this consistency, the analysis proceeded with ROUGE-L, which focuses on the longest common subsequence and is better suited for evaluating structural similarity and rephrased outputs.

The distribution of ROUGE-L scores differs significantly from that of BLEU. While BLEU scores (Figure 4.2) exhibit a flat distribution with no discernible pattern, the ROUGE-L histogram demonstrates a rising tendency, with a greater number of examples achieving higher scores.

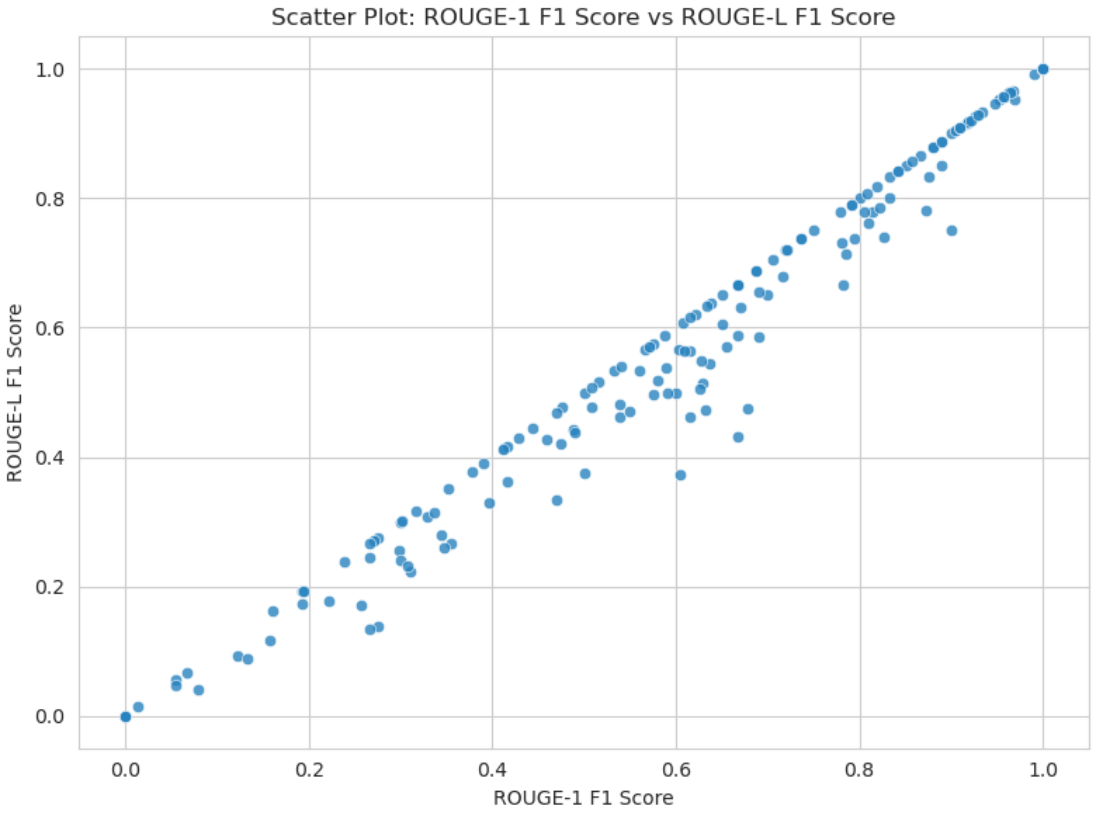

FIGURE 4.5: Scatter plot comparing ROUGE-1 and ROUGE-L scores for each generated response. The plot shows a strong positive correlation at higher score ranges, indicating agreement between the two metrics

This suggests that the RAG model's outputs tend to preserve key sequences from the reference texts, even in cases where the phrasing varies. This contrast in distribution arises from the differing computational strategies of the two metrics: BLEU is based on n-gram precision and is highly sensitive to exact wording and token order, whereas ROUGE, particularly ROUGE-L, emphasizes recall and sequence overlap, allowing for a more flexible interpretation of similarity that better accommodates rephrasings and partial matches.

Despite its advantages, ROUGE-L shares several limitations with BLEU, particularly in its treatment of factual and semantic correctness. For instance, in one example, the generated response inaccurately stated that the number of Autonomous Systems (ASes) in the United States was "0," while the reference answer correctly indicated "30827." Despite this clear factual error, ROUGE-L assigned a high score of 0.96, closely mirroring BLEU's 0.95, due to the high degree of lexical and structural overlap between the two sentences. This illustrates that ROUGE, like BLEU, fails to penalize factually incorrect but lexically similar outputs.

Conversely, in another instance, the model generated a response that was semantically accurate and contained additional relevant information, including the correct identification of the countries "Japan" and "United States of America" along with their associated IP prefixes.

However, due to differences in phrasing and the inclusion of extra content not found in the reference answer, ROUGE-L scored the response at 0.36. While this is an improvement over BLEU's score of 0.12 for the same example, it still significantly underrepresents the true quality of the answer. These cases highlight ROUGE-L's relative advantage in handling rephrased or enriched content, while also underscoring its continued difficulty in distinguishing lexical variation from semantic divergence.

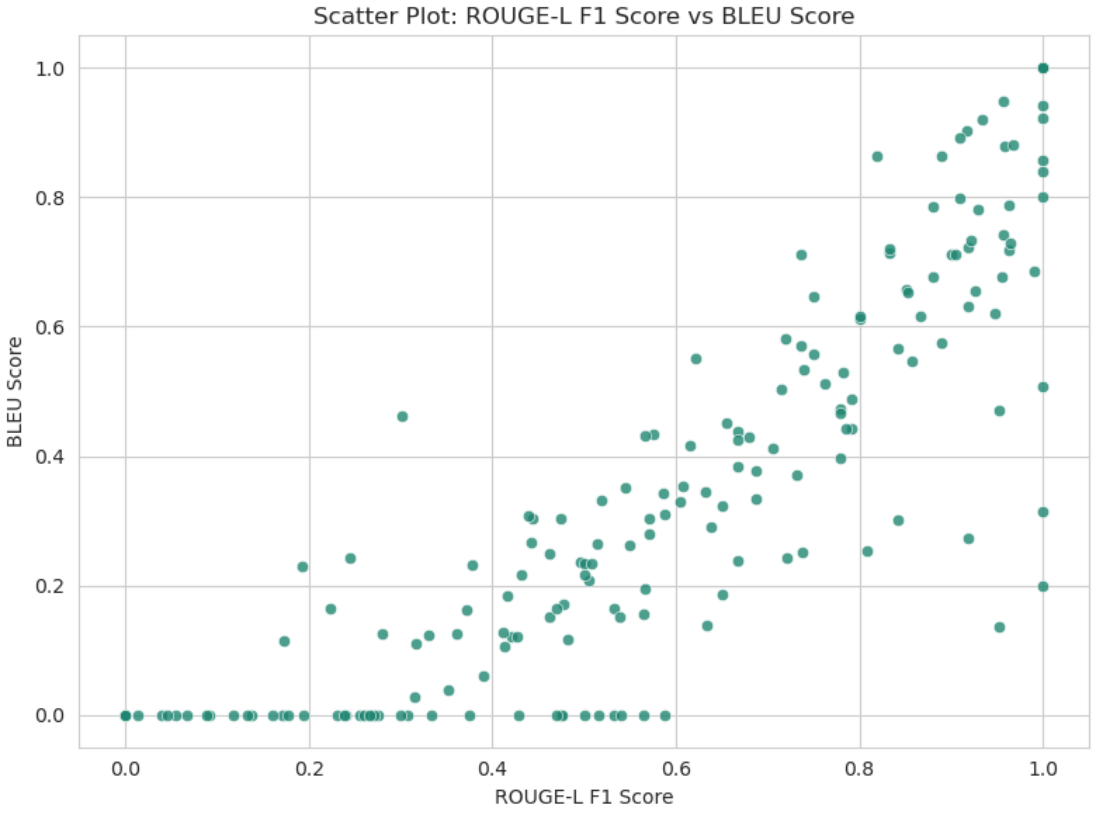

FIGURE 4.6: Scatter plot showing the relationship between ROUGE-L and BLEU scores for the RAG model's generated responses.

The scatter plot in Figure 4.6 further illustrates the relationship between the two metrics. A positive correlation is observed at higher score values, indicating agreement in cases where the model closely matches the reference. However, this correlation breaks down at lower scores, where the two metrics diverge, highlighting their differing sensitivities to content structure and phrasing.

Ultimately, ROUGE-L was selected as the preferred metric due to its more informative distribution and improved handling of paraphrased content. While it does not fully address the limitations of automatic lexical metrics in evaluating semantic correctness or factual detail, it provides a more nuanced reflection of the model's performance in contexts where linguistic variation and structural fidelity are important.

4.3.2.3 BERTScore Analysis

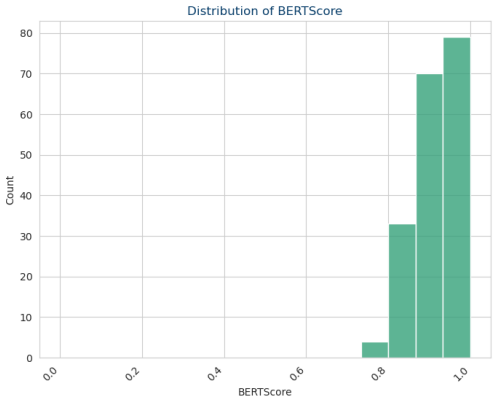

FIGURE 4.7: Histogram of BERTScore scores for the RAG model's responses on the IYP dataset. The x-axis shows score intervals from 0 to 1, while the y-axis indicates the number of responses per interval.

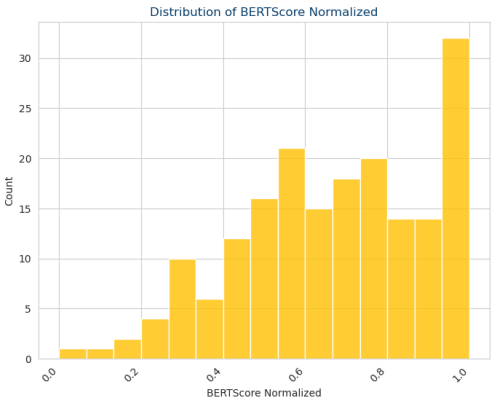

FIGURE 4.8: Histogram of normalized BERTScore values for the RAG model's responses on the IYP dataset.

To complement the lexical evaluations conducted with BLEU and ROUGE, BERTScore was also employed to assess the quality of the RAG model's outputs on the IYP dataset. BERTScore operates by leveraging contextual embeddings generated from a pretrained BERT model to measure semantic similarity between generated and reference texts. Unlike BLEU and ROUGE, which rely on exact lexical matches and sequence overlap, BERTScore evaluates meaning by comparing the cosine similarity of token embeddings, thereby providing a deeper representation of semantic alignment.

As shown in Figure 4.7, the distribution of BERTScore values demonstrates a pronounced rising tendency, with most scores skewing towards the upper end of the scale. The lowest observed score was 0.74, indicating that even the least similar outputs maintained a relatively high degree of semantic resemblance to their references. This inflation in values can be attributed to the nature of the pretrained BERT model, which was trained on large-scale, generic corpora. Because the IYP dataset is composed exclusively of domain-specific network data—particularly uniform in structure and terminology—the resulting embeddings tend to exhibit high inter-sentence similarity. This limits BERTScore's capacity to differentiate between high- and low-quality outputs within this specific domain, as even structurally divergent outputs may still produce similar contextual vectors.

To enable more meaningful comparisons and mitigate this compression effect, the BERTScore values were normalized to the [0,1] range, as illustrated in Figure 4.8. The normalized scores preserve the rising tendency observed in the raw values and reveal a more interpretable gradient of output quality. When plotted against ROUGE-L scores in a scatter plot (Figures 4.9, 4.10),

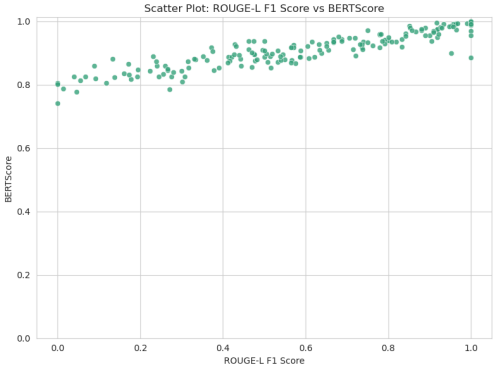

FIGURE 4.9: Scatter plot comparing raw BERTScore values with ROUGE-L scores for the RAG model's generated responses.

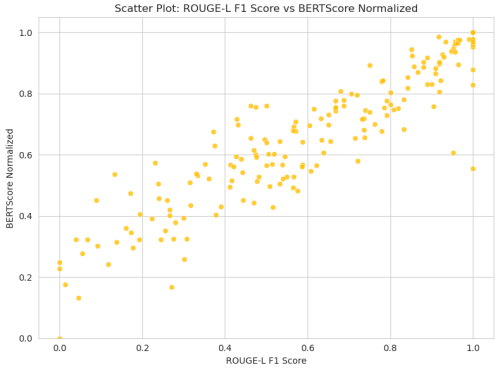

FIGURE 4.10: Scatter plot comparing normalized BERTScore values with ROUGE-L scores for the RAG model's generated responses.

both raw and normalized BERTScore values display a clear linear correlation with ROUGE, suggesting a consistent relationship between the two metrics in evaluating output fidelity.

Despite its semantic orientation, BERTScore exhibits many of the same limitations encountered with BLEU and ROUGE. In particular, it fails to penalize factual inaccuracies if the surrounding semantic context remains similar. For instance, in the example where the generated response erroneously states that the number of Autonomous Systems (ASes) in the United States is "0" (as opposed to the correct value "30827"), BERTScore still assigns a notably high value of 0.98 (normalized: 0.94), aligning closely with ROUGE-L's 0.96 and BLEU's 0.95. This reflects the metric's tendency to over-reward lexical similarity, even when the semantic integrity of the content is compromised by factual errors.

Similarly, BERTScore demonstrates insensitivity in the opposite case—underestimating outputs that are semantically accurate but rephrased or enriched with additional relevant information. In the previously discussed example, where the generated output correctly identifies the countries "Japan" and "United States of America" along with associated IP prefixes, but includes phrasings and details not present in the reference, BERTScore assigns a moderate value of 0.88 (normalized: 0.52). Although higher than BLEU's 0.12 and ROUGE's 0.36, this still underrepresents the true semantic and factual quality of the response.

Given these observations, ROUGE-L was selected as the preferred evaluation metric over BERTScore. The primary rationale lies in the linear correlation observed between the two metrics, indicating that they generally rank outputs similarly. However, unlike BERTScore, ROUGE-L produces a more informative and interpretable distribution of values, and does not suffer from the ceiling effect introduced by BERT's pretrained embedding space on highly domain-specific data. Additionally, the domain mismatch between BERT's training data and the IYP corpus diminishes BERTScore's reliability as a standalone metric in this context. Therefore, while

BERTScore provides valuable insight into semantic similarity, its limitations in domain adaptation, factual accuracy sensitivity, and score dispersion make ROUGE-L a more suitable choice for evaluating the RAG model's performance in this experiment.

4.3.2.4 G-Eval Analysis

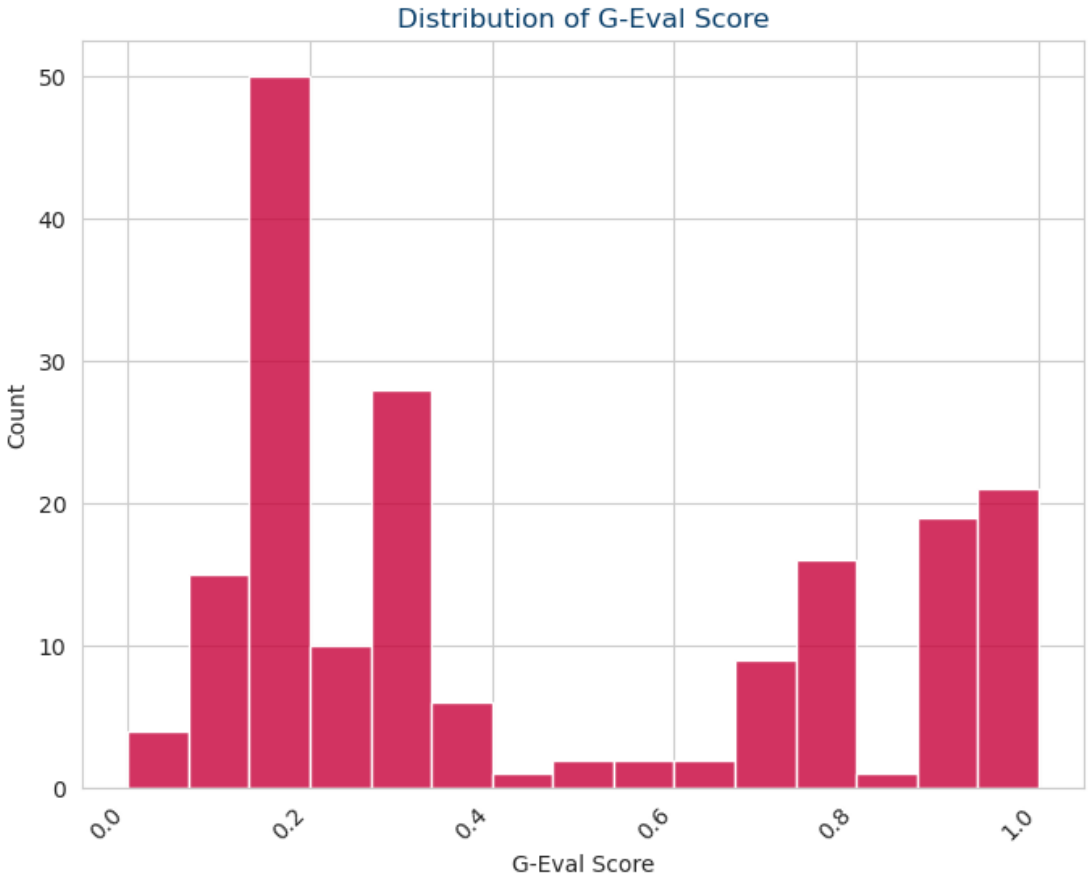

FIGURE 4.11: Histogram of G-Eval scores for the RAG model's responses on the IYP dataset. The x-axis shows score intervals from 0 to 1, while the y-axis indicates the number of responses per interval.

To address the limitations of traditional and embedding-based evaluation metrics—namely BLEU, ROUGE, and BERTScore—a more robust semantic and factual accuracy metric, G-Eval, was incorporated into the evaluation of the RAG model. G-Eval operates by prompting a large language model to assess the correctness, completeness, and coherence of a generated response with respect to its reference, returning a score between 0 and 1 alongside a justification. This qualitative grounding allows G-Eval to capture nuanced distinctions in output quality that are often missed by surface-level or embedding-based similarity measures.

As illustrated in Figure 4.11, the distribution of G-Eval scores exhibits a distinctive bimodal pattern, characterized by concentrations at both the lower and upper ends of the scale, with

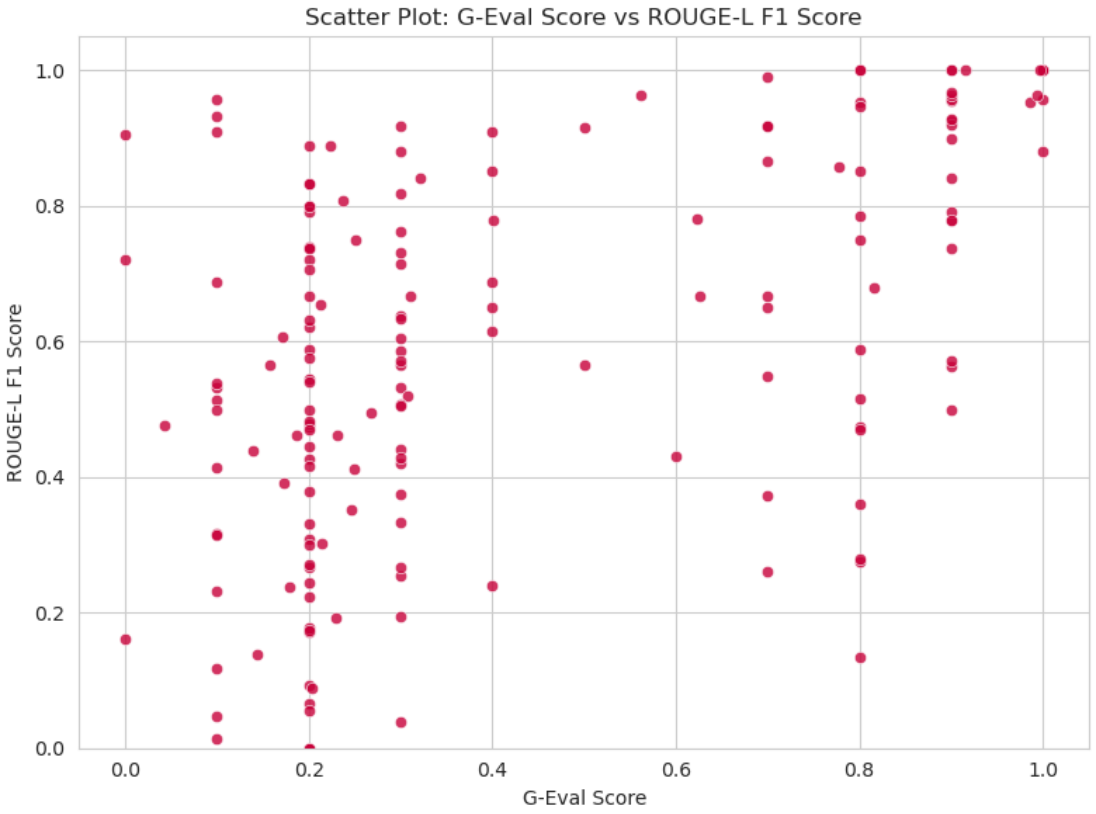

FIGURE 4.12: Scatter plot comparing normalized G-Eval values with ROUGE-L scores for the RAG model's generated responses.

relatively few values near the midpoint (0.5). This distribution indicates that the metric decisively classifies outputs as either high-quality or low-quality, with minimal ambiguity. This behavior is a direct consequence of the tailored prompts used to guide the evaluation model, which explicitly instructed the system to make firm judgments regarding correctness and informativeness rather than offering hedged or uncertain assessments.

Furthermore, when G-Eval scores are plotted against ROUGE-L in a scatter plot (Figure 4.12), no correlation is observed. This stark divergence underscores a key strength of G-Eval: it remains unaffected by superficial textual overlap and is instead more attuned to semantic correctness and factual precision. For example, in a case where the generated output incorrectly states that the number of Autonomous Systems (ASes) in the United States is "0" instead of the correct value "30827," the traditional metrics yield disproportionately high scores—BLEU: 0.95, ROUGE-L: 0.96, and BERTScore: 0.98 (normalized: 0.94). In contrast, G-Eval assigns a low score of 0.1, accurately reflecting the critical factual error despite lexical similarity.

Conversely, in an instance where the generated text includes accurate information that has been rephrased and supplemented with additional context—identifying countries such as "Japan" and "United States of America" along with associated IP prefixes—most metrics fail to capture the full semantic fidelity of the response. BLEU scores this output at 0.12, ROUGE-L at 0.36, and

BERTScore at 0.88 (normalized: 0.52). G-Eval, however, recognizes the factual correctness and coherence of the rephrased content and assigns a high score of 0.8, demonstrating its capacity to reward outputs that diverge lexically but remain semantically aligned.

Crucially, G-Eval avoids the pitfalls of over-rewarding surface similarity or penalizing semantic rephrasing—common issues with BLEU, ROUGE, and BERTScore—by focusing on interpretive and factual alignment rather than token overlap or embedding proximity. Moreover, unlike other metrics, G-Eval provides detailed reasoning for each evaluation, making it a valuable tool not only for scoring but also for error analysis and model refinement. These qualitative insights can guide targeted fine-tuning of the RAG model on failure cases, ultimately improving the robustness and factual reliability of generated outputs.

Given these advantages, G-Eval is selected as the most appropriate evaluation metric for this study. Its ability to discern semantic quality across a wide range of output types, handle rephrased or enriched content, penalize factual inaccuracies, and provide explanatory feedback makes it superior to BLEU, ROUGE, and BERTScore within the context of evaluating outputs on the IYP dataset.

4.3.2.5 Evaluation Summary

To provide a comprehensive overview of the metric-based evaluation results, a comparative boxplot (Figure 4.13) consolidates the distributions of BLEU, ROUGE-1, ROUGE-L, BERTScore (both raw and normalized), and G-Eval scores. This visualization allows for a high-level comparison of the behavior and tendencies of each metric. The BLEU score distribution is notably skewed toward the lower end of the scale, reaffirming earlier observations that BLEU struggles to accommodate rephrased, semantically correct outputs. Its strict dependence on n-gram precision makes it overly sensitive to minor lexical and structural deviations, which leads to underestimation of true response quality in many cases.

In contrast, ROUGE-1 and ROUGE-L exhibit nearly identical distributions in the boxplot, which is consistent with their high mutual correlation and their shared focus on sequence overlap. Both variants reflect a tendency to assign higher scores than BLEU, particularly in cases where the generated responses retain structural similarity to the reference even if exact phrasings differ. This further supports the decision to use ROUGE-L as the representative ROUGE metric in subsequent analyses.

The BERTScore distributions, especially the raw version, are heavily skewed toward the upper end of the scale, indicating a ceiling effect. This is a direct consequence of the use of pretrained language models, which tend to generate high semantic similarity scores for structurally similar content, even when such content contains factual inaccuracies. The normalized version of

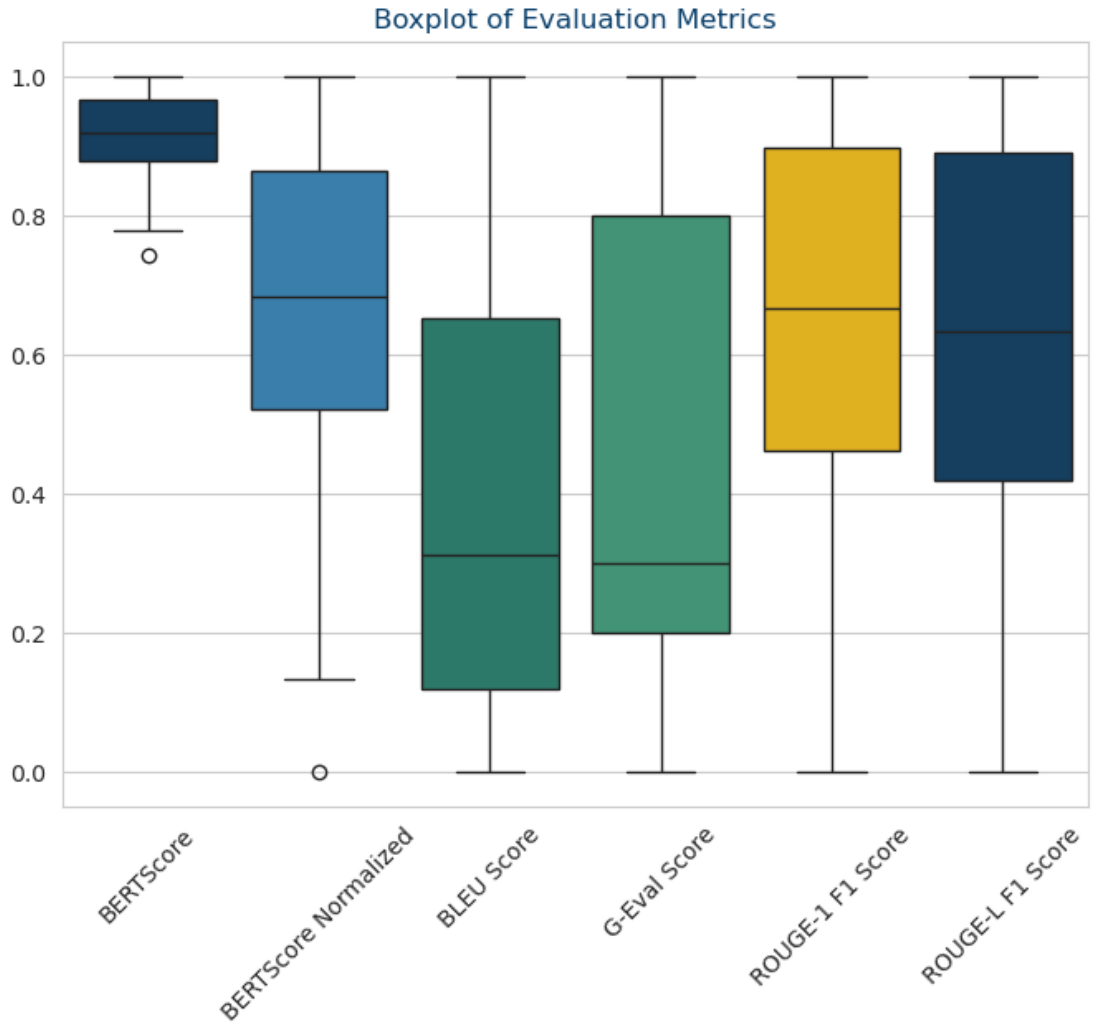

FIGURE 4.13: Comparative boxplot displaying the distributions of BLEU, ROUGE-1, ROUGE-L, BERTScore (raw and normalized), and G-Eval scores for the RAG model's outputs on the IYP dataset. This visualization highlights differences in metric behaviors.

BERTScore mitigates this compression effect to some extent and produces a score distribution more comparable to that of ROUGE-L. However, both versions of BERTScore show an overall preference for assigning higher values, potentially obscuring critical distinctions in response quality—particularly in a domain with narrow linguistic variety but high factual specificity like IYP.

The G-Eval score distribution presents a stark contrast to the other metrics. It features a distinctive bimodal pattern, with clusters of scores at both the low and high ends, and very few scores concentrated near the midpoint (0.5). This bimodality indicates that G-Eval tends to make clear-cut assessments of response quality, classifying outputs as either strong or weak with minimal ambiguity. This decisive behavior stems from the design of the evaluation prompts, which explicitly instructed the model to prioritize judgments of factual accuracy and informativeness over hedged or uncertain evaluations. Importantly, the G-Eval distribution spans the

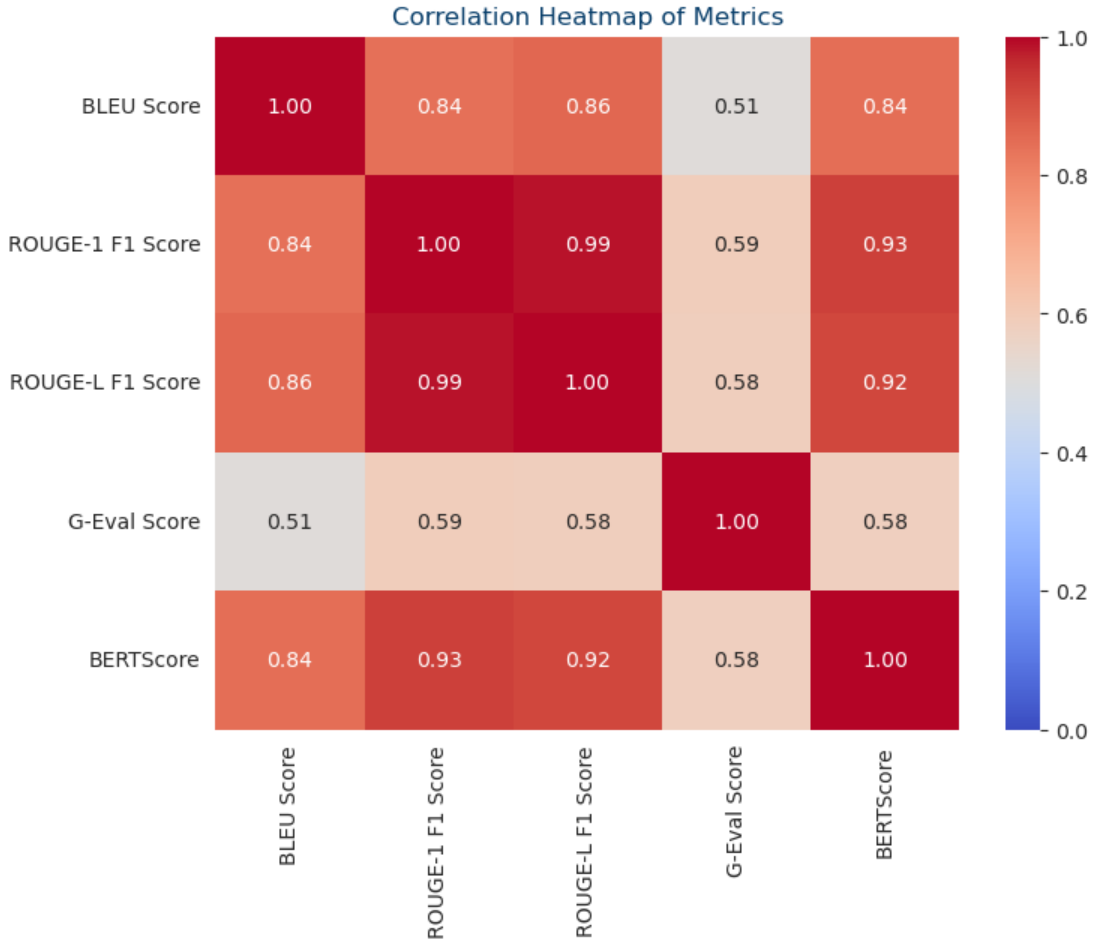

FIGURE 4.14: Correlation matrix illustrating relationships among evaluation metrics for the RAG model's outputs on the IYP dataset.

full range of scores without the pronounced skew seen in BLEU or BERTScore, suggesting a more balanced and discerning evaluation capability.

To further understand how these metrics relate to one another, a correlation matrix (Figure 4.14) was computed. ROUGE-1 and ROUGE-L, as expected, exhibit near-perfect correlation with a coefficient of 0.99, reflecting their nearly interchangeable behavior in this context. BERTScore (normalized) is also strongly correlated with both ROUGE variants, showing a correlation coefficient of 0.92. BLEU, while generally more conservative in its scoring, maintains a moderate-to-strong correlation with ROUGE and BERTScore (around 0.84 for both), indicating some shared sensitivity to lexical and structural features despite its harsher scoring tendencies.

G-Eval, however, demonstrates weak correlation with all other metrics, with coefficients in the range of 0.5 to 0.6. This relative independence from traditional lexical and embedding-based similarity metrics highlights G-Eval's unique evaluative lens, one that places greater emphasis on semantic fidelity, factual accuracy, and interpretive quality. Unlike the other metrics, G-Eval is not overly influenced by surface-level text similarity or token overlap, making it a valuable

complementary measure—especially in a fact-sensitive, domain-specific context where lexical similarity can be misleading.

Given these findings, G-Eval emerges as the most suitable metric for evaluating generated responses in the IYP dataset. It avoids the pitfalls of traditional metrics that tend to either over-reward surface-level matches or underappreciate correct yet paraphrased responses. Additionally, G-Eval offers qualitative justifications for each score, enabling deeper error analysis and offering actionable insights for model refinement. This interpretability is a critical advantage in iterative development cycles, as it allows researchers to identify failure modes, such as specific types of factual errors or hallucinations, and tailor future training or prompt engineering efforts accordingly.

Quantitatively, the mean G-Eval score for the RAG model across all evaluated examples is 0.46. While this average suggests that the model has room for improvement—particularly in ensuring factual correctness and handling edge cases—it also reflects a promising baseline. With no existing models tailored for this dataset or domain, achieving a G-Eval score near 0.5 implies that the system can produce high-quality answers for approximately one out of every two queries. This result is particularly encouraging in the context of Internet Yellow Pages (IYP)-style queries, where even partial coverage of questions with correct, informative responses can significantly enhance usability and practical value.

In conclusion, the metric comparisons reveal both the strengths and limitations of existing evaluation tools for graph-grounded question answering. Traditional lexical metrics such as BLEU and ROUGE offer surface-level insights, while embedding-based metrics like BERTScore improve semantic sensitivity but remain affected by domain mismatches and limited score dispersion. G-Eval, by contrast, introduces a robust, semantically aware, and interpretable alternative that aligns more closely with the goals of factual, context-aware response generation. As such, it forms the cornerstone of evaluation for this study and provides a foundation for future fine-tuning and benchmarking efforts in the IYP domain.

4.3.3 Prompt Complexity Evaluation

Building on the previous section's metric-based assessment of the RAG model's performance on the IYP dataset, this subsection shifts focus to explore how prompt complexity influences response quality. While the prior analysis revealed that G-Eval offers a more nuanced and semantically rich evaluation than traditional overlap-based metrics like ROUGE or BLEU, it did not yet account for variation in input difficulty. Given that real-world user queries span a spectrum of complexity, from straightforward factual lookups to abstract or multi-step reasoning tasks, it is crucial to understand how well the model handles this diversity.

To this end, prompts in the IYP dataset were annotated with difficulty levels across two dimensions: cognitive complexity (Easy, Medium, Hard) and domain type (Technical vs. General). This classification enables a more fine-grained evaluation of the model's strengths and weaknesses across different query types. By applying both G-Eval and ROUGE-L to this stratified prompt set, we assess how response quality degrades—or potentially shifts—under increasing challenge, and whether domain-specific patterns emerge.

The following analyses offer insight into not only how much performance changes with difficulty, but also why such changes occur. They reinforce the advantages of using G-Eval for difficulty-aware evaluation, while also highlighting the limits of traditional metrics like ROUGE-L in capturing deeper semantic or reasoning-related shortcomings. Ultimately, this section aims to identify key pain points in model behavior that can inform targeted improvements in future iterations of grounded, retrieval-augmented QA systems.

4.3.3.1 G-Eval Analysis

To better understand how prompt complexity influences model performance, G-Eval scores were analyzed across varying difficulty levels. Figure 4.15 presents a boxplot showing the distribution of G-Eval scores for each prompt category. The x-axis displays the six prompt categories in increasing order of difficulty: Easy Technical, Easy General, Medium Technical, Medium General, Hard Technical, and Hard General. The y-axis represents the G-Eval score, ranging from 0 (poor quality) to 1 (high quality). This visualization reveals a clear trend: as difficulty increases, the distribution of scores shifts downward. Prompts labeled as "easy" yield higher and more consistent scores, while the "medium" and "hard" categories show broader dispersion and generally lower values.

Notably, no significant differences are observed between technical and general prompts within the same difficulty level. This suggests that prompt difficulty—rather than the domain type—is the primary factor influencing model performance in this evaluation.

Type	Easy	Medium	Hard
General	0.52	0.38	0.34
Technical	0.58	0.40	0.30

Table 4.1: Mean G-Eval scores across different difficulty levels and types

Table 4.1 complements this analysis by reporting the mean G-Eval scores for each prompt category. These numerical results quantify the trends seen in the boxplot. On average, the model performs approximately 8% better on technical prompts for both the easy and medium difficulty levels. However, this pattern reverses at the hard level, where general prompts yield better scores than technical ones by roughly 12%. This reversal may reflect the greater factual

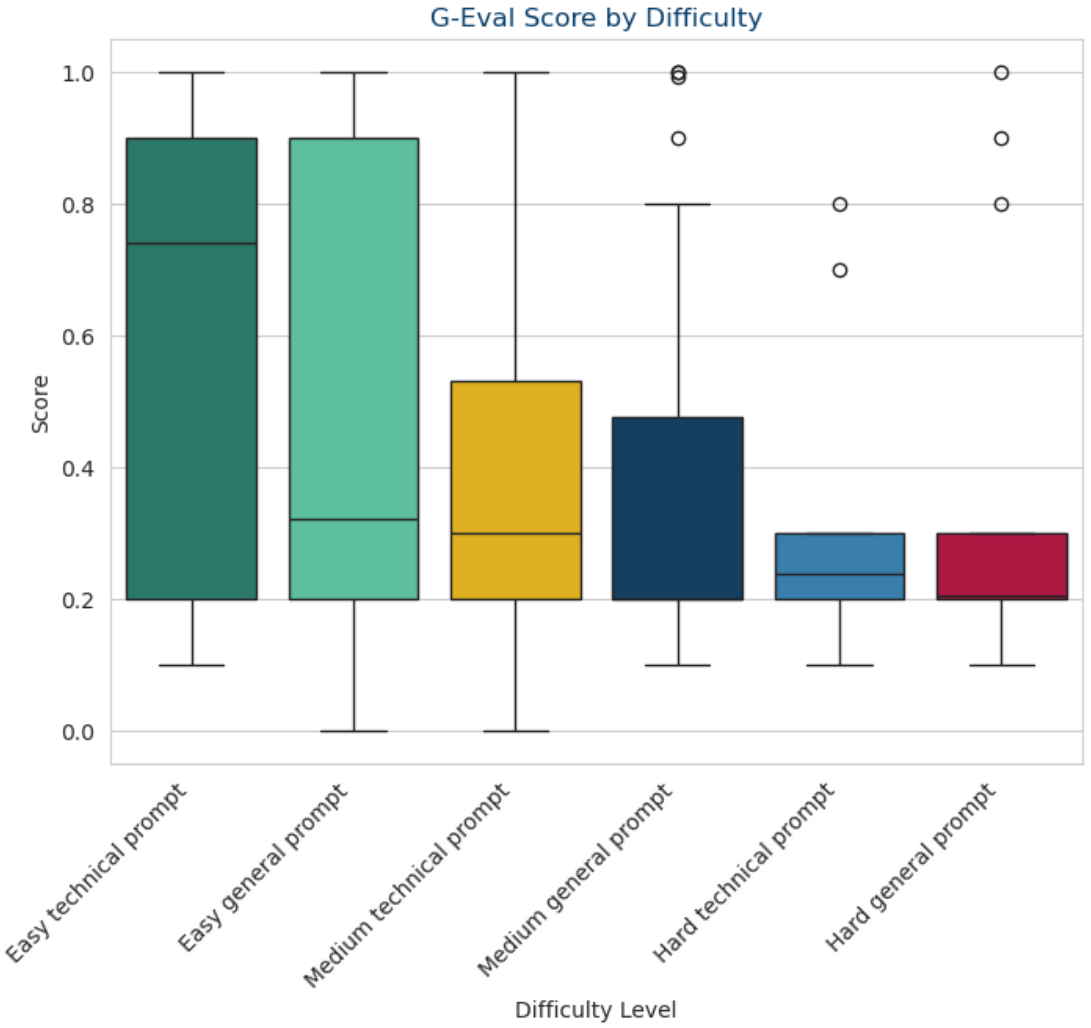

FIGURE 4.15: Boxplot of G-Eval scores across six prompt difficulty categories, ranging from Easy Technical to Hard General. The x-axis orders prompt categories by increasing difficulty, and the y-axis shows G-Eval scores from 0 (low quality) to 1 (high quality).

precision required for hard technical prompts, which present a higher challenge for the model's grounding and reasoning capabilities.

As difficulty increases, a notable degradation in average performance is observed. Specifically, there is an approximate 30% drop in mean G-Eval score when moving from easy to medium prompts, followed by an additional 18% reduction from medium to hard. These stepwise declines underscore the model's growing struggle with more complex or nuanced queries and signal the importance of targeting these segments for future refinement.

Finally, Figure 4.16 presents a stacked bar plot showing the distribution of G-Eval scores across all prompt difficulties. The x-axis represents discrete score bins (e.g., 0.0–0.2, 0.2–0.4, up to 1.0), and the y-axis indicates the number of responses (count) that fall into each score bin. Each bar is color-coded according to prompt difficulty, allowing a visual breakdown of how score frequencies differ by category. The plot highlights the bimodal nature of the G-Eval scoring: most

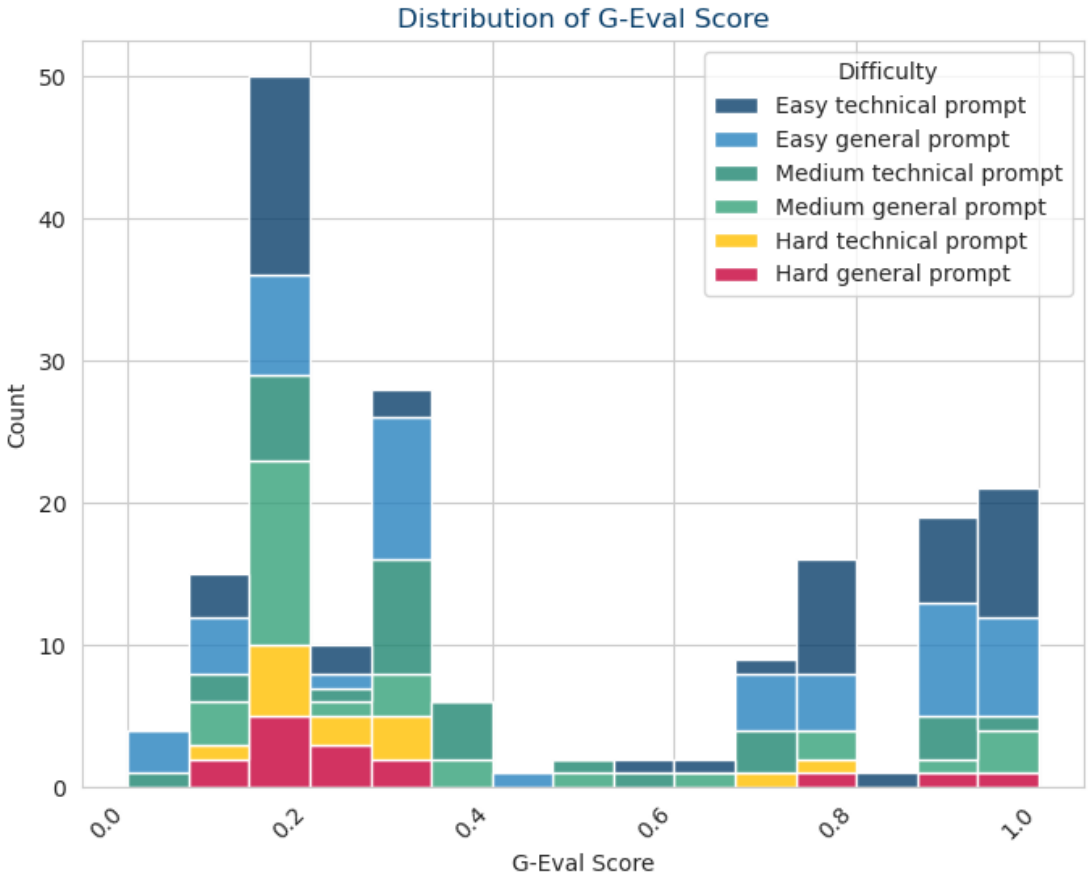

FIGURE 4.16: Stacked bar plot showing the distribution of G-Eval scores across all prompt difficulty levels. The x-axis displays discrete score bins from 0.0 to 1.0, and the y-axis indicates the count of responses per bin. Bars are color-coded by prompt difficulty.

responses are either rated low (around 0.2–0.4) or high (around 0.8–1.0), with relatively few falling in the midrange. This further illustrates G-Eval's decisive evaluative behavior. Easier prompts are heavily represented in the higher score bins, while medium and especially hard prompts cluster in the lower ranges, confirming the model's struggle with more challenging inputs.

In conclusion, prompt difficulty has a profound effect on the model's ability to generate high-quality answers, as measured by G-Eval. While technical prompts slightly outperform general ones at lower difficulties, this advantage disappears at higher levels, where general prompts yield better scores. Overall, the clear degradation in score distributions and averages across difficulty levels reflects the current limitations of the RAG model in handling complex, information-dense questions. These results highlight the value of difficulty-aware evaluation and suggest that future efforts should prioritize improving performance on medium and hard prompts—particularly in the technical domain.

4.3.3.2 ROUGE-L Analysis

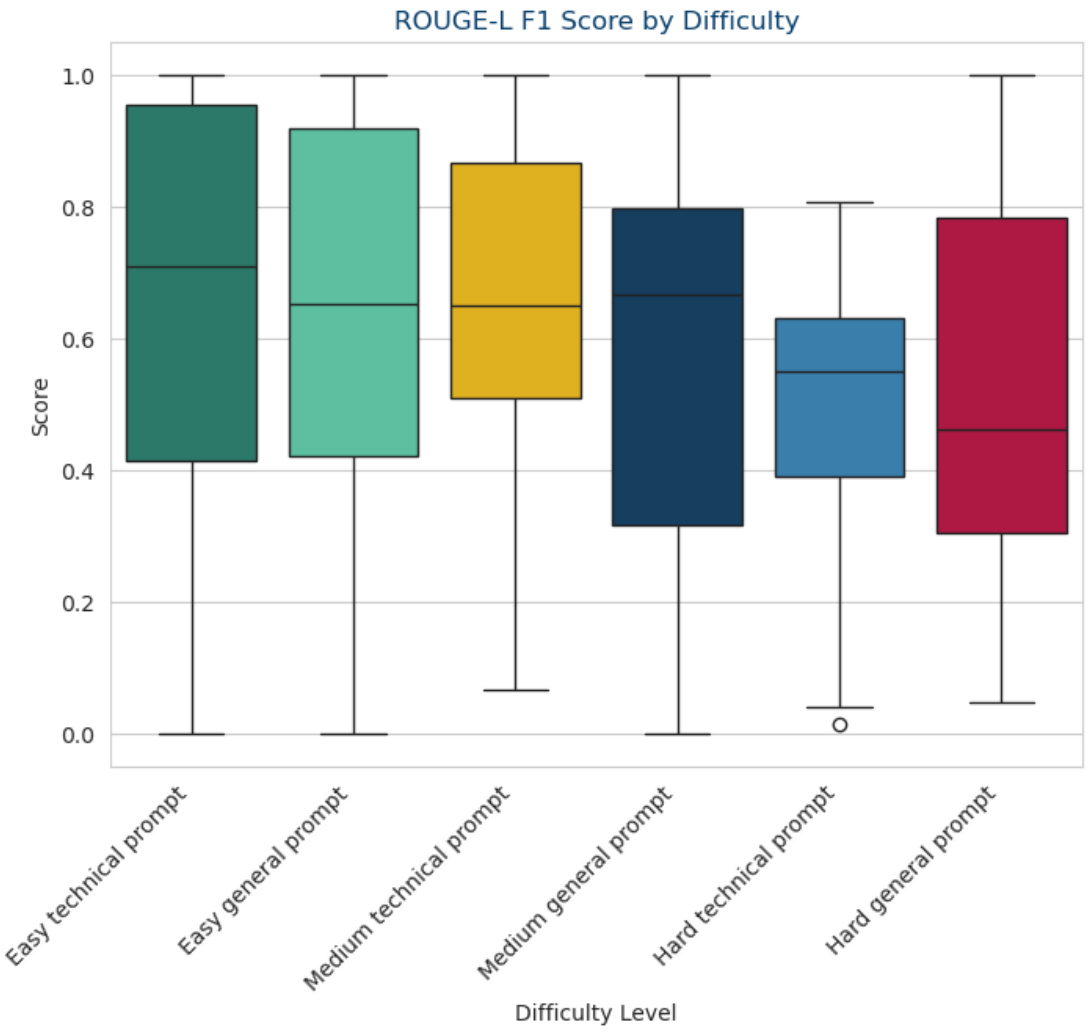

FIGURE 4.17: Boxplot of ROUGE-L scores across six prompt difficulty categories, ranging from Easy Technical to Hard General. The x-axis orders prompt categories by increasing difficulty, and the y-axis shows G-Eval scores from 0 (low quality) to 1 (high quality).

To complement the G-Eval findings and provide a broader view of evaluation outcomes, ROUGE-L scores were also analyzed across the same set of prompt difficulty levels. Figure 4.17 presents a boxplot depicting the distribution of ROUGE-L scores for each of the six prompt categories, arranged from easiest to hardest: Easy Technical, Easy General, Medium Technical, Medium General, Hard Technical, and Hard General. The y-axis captures the ROUGE-L score, which ranges from 0 (no overlap with reference) to 1 (perfect overlap).

While a general downward trend is still visible as prompt difficulty increases, the decline is less pronounced than in the G-Eval results. In particular, the difference between easy and medium prompts is subtle in the boxplot, suggesting a more gradual reduction in overlap-based performance. The distributions for each category also appear more overlapped, indicating greater variability and less clear separation between difficulty levels in this metric.

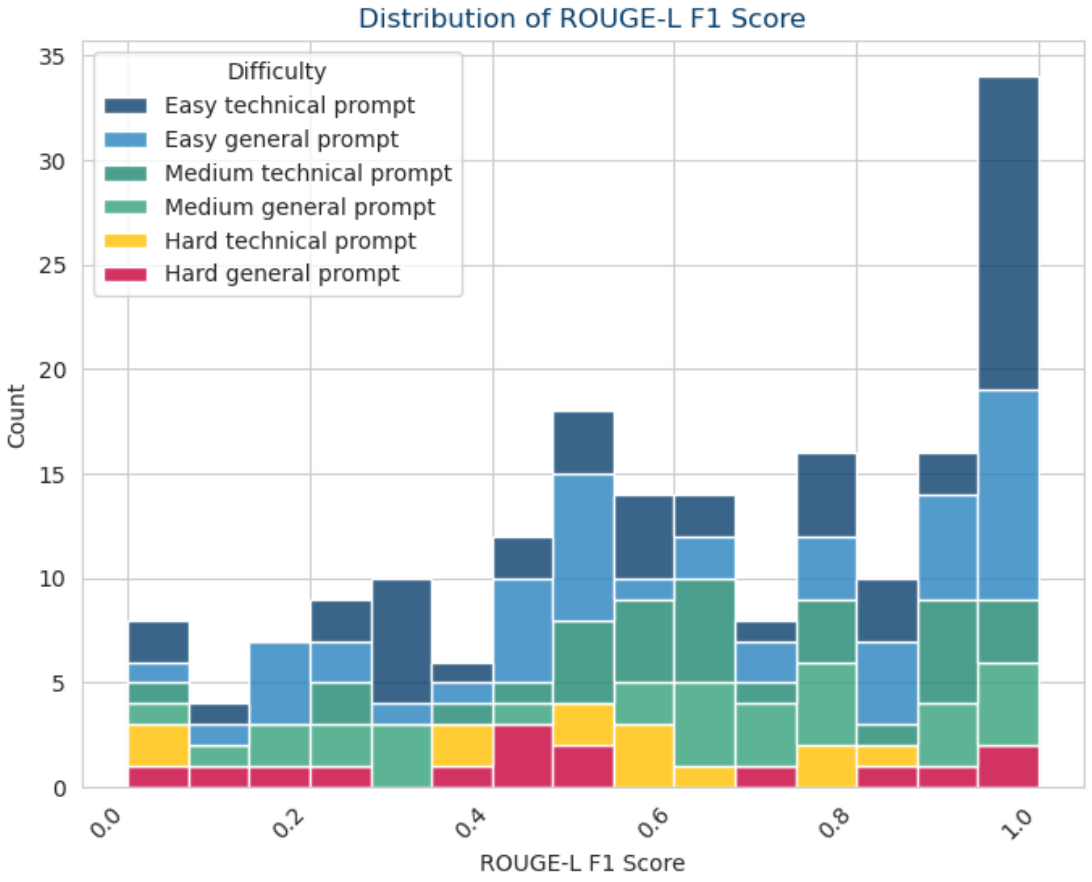

FIGURE 4.18: Stacked bar plot showing the distribution of ROUGE-L scores across all prompt difficulty levels. The x-axis displays discrete score bins from 0.0 to 1.0, and the y-axis indicates the count of responses per bin. Bars are color-coded by prompt difficulty.

Type	Easy	Medium	Hard
General	0.64	0.59	0.51
Technical	0.66	0.65	0.50

TABLE 4.2: Mean ROUGE-L scores across different difficulty levels and types

Consistent with this visualization, Table 4.2 reports the mean ROUGE-L scores for each prompt category and reveals a nuanced pattern. For both the easy and hard difficulty levels, general and technical prompts yield nearly identical performance. However, at the medium level, technical prompts outperform general ones by approximately 10%, suggesting a domain-specific advantage at this intermediate stage. Overall, the average score drops only slightly—around 5%—when moving from easy to medium difficulty, but the decline from medium to hard prompts is more substantial, at roughly 18%. This confirms that the ROUGE-L metric still captures the growing challenge posed by harder prompts, albeit with a less dramatic gradient than G-Eval.

Further insights are offered by the stacked bar plot shown in Figure 4.18, which displays the distribution of ROUGE-L scores across all prompt difficulties. The x-axis represents score bins (e.g., 0.0–0.2, 0.2–0.4, etc.), while the y-axis indicates the number of model responses

per bin. Color-coding by difficulty level enables an intuitive view of how different categories contribute to the overall score landscape. The distribution highlights a shift in response quality across difficulty levels: easy prompts cluster strongly toward higher scores (particularly above 0.8), medium prompts exhibit a wide and dispersed spread across the full score range, and hard prompts show a bimodal distribution—concentrated both at lower scores and around the midrange (especially near 0.5). This pattern reflects increased inconsistency in the model's ability to match reference answers as prompt complexity rises.

In summary, ROUGE-L does capture the general trend of declining performance with increasing prompt difficulty, particularly evident in the mean score reductions and the changing shape of the score distributions. However, the differences across difficulty levels are less pronounced compared to G-Eval, and the metric shows limited sensitivity to semantic nuances, factual accuracy, and reasoning depth. As a surface-level, overlap-based measure, ROUGE-L provides a useful but partial view of model performance. In contrast, G-Eval appears better suited to this evaluation context, offering more decisive and discriminative insights into how well the model handles prompts of varying complexity.

4.3.3.3 Prompt Difficulty Level Evaluation Summary

The evaluation results clearly demonstrate that prompt difficulty significantly affects model performance, as reflected by both G-Eval and ROUGE-L metrics. Across all analyses, increasing difficulty levels, from Easy to Medium to Hard, lead to progressive declines in performance.

G-Eval results show a sharp downward trend in both score distributions and means as difficulty increases. Easy prompts consistently yield high, stable scores, while medium and hard prompts show greater variability and lower overall quality, especially in technical domains. Interestingly, while technical prompts slightly outperform general ones at lower difficulty levels, this advantage reverses at the hard level, likely due to the higher factual and reasoning demands of hard technical prompts.

ROUGE-L, although showing a similar overall decline, exhibits less pronounced sensitivity to difficulty. The overlap in score distributions and smaller drops in mean values indicate that ROUGE-L captures surface-level differences less decisively. Medium prompts, in particular, show only a modest performance drop compared to easy ones, and category distinctions are less distinct.

Taken together, these findings emphasize that G-Eval is the more reliable and discriminative metric for assessing model responses to prompts of varying complexity. It better captures semantic depth, factual correctness, and reasoning challenges than ROUGE-L.

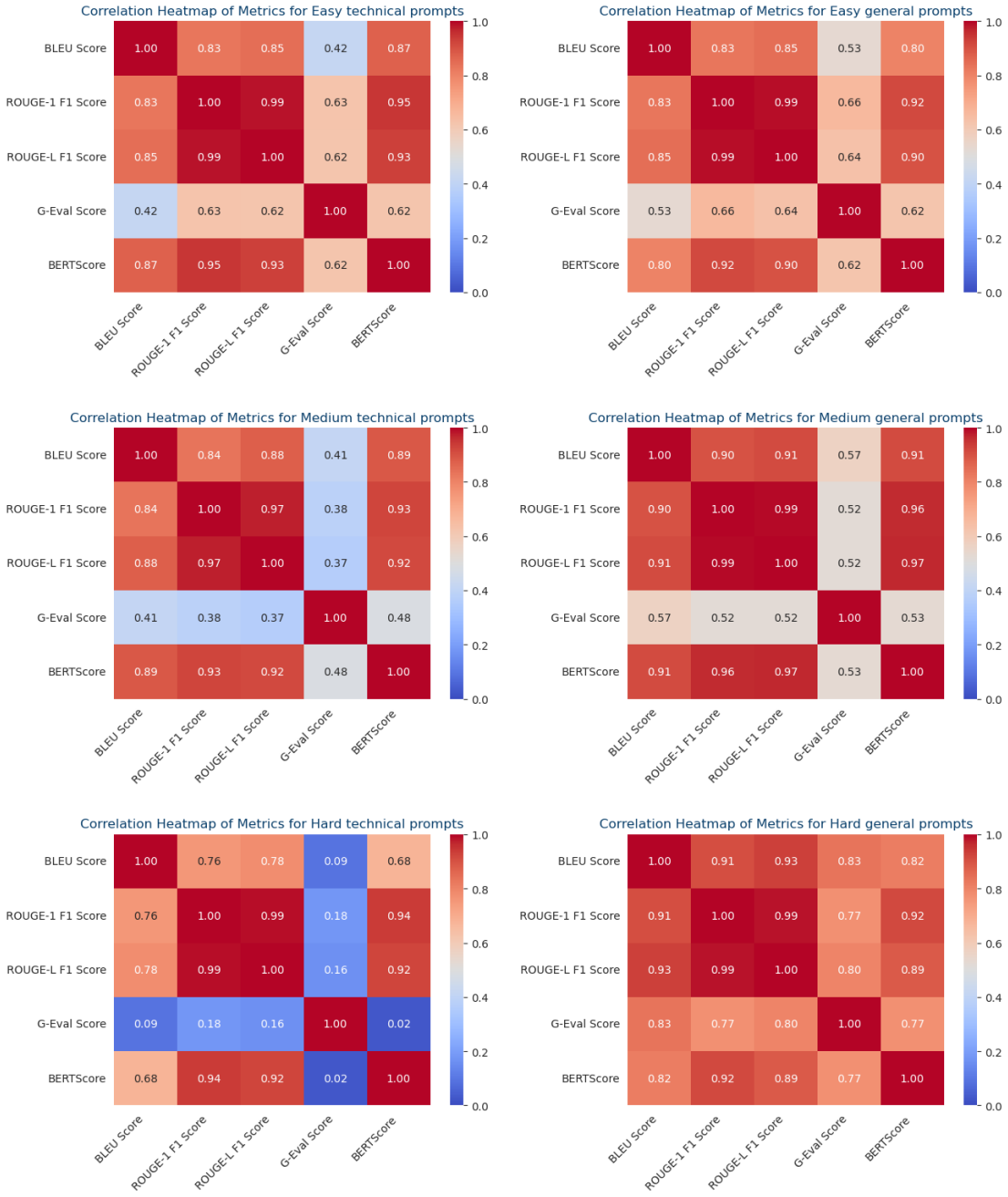

FIGURE 4.19: Correlation matrices of evaluation metrics across prompt difficulties and types.

Further insights are revealed in Figure 4.19, which presents correlation matrices between all evaluation metrics across prompt difficulties and types. For easy and medium prompts, G-Eval shows only moderate correlation (approximately 0.4–0.6) with other metrics, which aligns with the earlier findings. However, for hard prompts, correlation patterns diverge significantly. In hard general prompts, G-Eval aligns strongly with other metrics (around 0.80), indicating that poor performance is consistently captured across evaluation dimensions. In contrast, hard technical prompts exhibit a weak or even negative correlation (around 0.15), implying that

while lexical metrics may register surface similarity, G-Eval detects substantial semantic or factual errors, possibly due to the model producing answers that are lexically similar but conceptually incorrect. This discrepancy highlights the unique value of G-Eval in diagnosing deeper quality issues, especially in challenging technical scenarios. Nevertheless, further targeted analysis is warranted to fully unpack these trends and their implications.

For future model development and evaluation, prompt difficulty must be treated as a central factor. The consistent degradation in performance with increasing complexity—especially in technical domains—highlights the need to prioritize improvements in handling medium and hard prompts. Difficulty-aware evaluation strategies will be essential for building more robust and capable models.

4.3.4 Results Summary

The analyses presented in this section collectively highlight the importance of both metric selection and prompt difficulty awareness in evaluating the performance of the RAG model on the IYP dataset. Among the diverse set of evaluation metrics considered, G-Eval emerged as the most reliable and informative, consistently capturing semantic accuracy, factual grounding, and reasoning quality across varying prompt complexities. Its superior discriminative power, particularly in contrast to traditional overlap-based metrics like BLEU and ROUGE-L, underscores its alignment with the nuanced demands of the IYP task.

This finding represents a key scientific contribution of this work. No prior study was found to have systematically evaluated multiple automated scoring methods within this specific context. By rigorously comparing statistical, hybrid, and model-based metrics, it becomes more evident that model-based approaches like G-Eval provide the most accurate and meaningful assessment of model responses in complex, information-seeking tasks. This insight not only validates the use of G-Eval for subsequent evaluations but also offers a valuable methodological guideline for future research in similar domains.

Another key finding is that both the difficulty and type of prompts exert a significant influence on model performance, further emphasizing the need for nuanced evaluation frameworks. The data clearly show that as prompt difficulty increases, model outputs deteriorate not only in surface-level coherence but also in deeper semantic and factual dimensions, with G-Eval revealing sharper performance declines than traditional metrics. Moreover, the interaction between difficulty and prompt type, particularly the reversal of performance trends between technical and general prompts at higher difficulty levels, suggests that the model struggles more with domain-specific reasoning and knowledge synthesis under complex conditions. This pattern highlights a critical challenge for RAG systems operating in specialized knowledge domains: general prompt understanding may be manageable at all levels, but technical prompts impose

compounding burdens on retrieval quality, factual grounding, and reasoning depth. These findings underscore the importance of factoring in prompt type and difficulty when evaluating or developing RAG systems on the IYP graph, as they directly affect both model behavior and the interpretability of evaluation results.

Ultimately, these results demonstrate that accurate evaluation of the RAG model on the IYP dataset requires careful consideration of both metric selection and prompt characteristics. G-Eval emerges as the most reliable metric, offering meaningful insights into semantic and factual quality, especially in complex scenarios. Additionally, prompt difficulty and type significantly impact model performance, with harder and more technical prompts exposing deeper limitations. Together, these findings underscore the importance of using robust evaluation methods and structured prompt analysis to assess and improve RAG systems in the IYP domain.

The insights derived from these evaluations not only deepen the understanding of model behavior across varying prompt conditions but also directly inform the design and development of the ChatIYP system. The next chapter shifts focus from experimental analysis to system implementation, detailing how the core methodological concepts, such as hybrid retrieval, semantic reasoning, and prompt-aware generation, were translated into functional components. It presents the architecture and interactive interface of ChatIYP, illustrating how these elements work together to enable grounded, explainable, and contextually relevant responses over the IYP knowledge graph.