Rules Lawyer: An NLP-Based Question-Answering System for Board Game Rulebooks

Fred Hayes

Abstract

Understanding complex board game rulebooks poses significant challenges, especially during gameplay when quick, accurate rule clarifications are essential. This paper introduces Rules Lawyer, a question-answering (QA) system leveraging state-of-the-art models—T5 (Base and Large), Flan-T5, BERT, and UL2—to address rule-based QA in a niche domain.

The system utilizes retrieval-augmented methods and fine-tuned language models to handle nuanced, multi-hop queries. A custom dataset of 500+ question-answer pairs and representative sample questions were developed from *Dune: Imperium*, *Carcassonne*, and *Star Wars: The Deckbuilding Game*. Evaluations using ROUGE, BLEU, and BERTScore metrics showed that the optimized T5-Large model outperformed others in coherence and fluency, while UL2 excelled in semantic reasoning.

This work contributes a novel dataset, retrieval framework, and insights into hyperparameter tuning for rule-intensive QA tasks. Future directions include expanding the dataset, refining retrieval methods, and exploring real-time applications.

Introduction

Interpreting the rules of modern board games can be a daunting task. As games grow more complex, with expansions and nuanced interactions, players often struggle to understand rulebooks during gameplay. Static FAQs and online resources fail to address dynamic, context-sensitive questions that arise. This study introduces Rules Lawyer, a question-answering (QA) system designed to deliver quick and accurate rule clarifications.

The project leverages state-of-the-art NLP models, including T5 (Base and Large), Flan-T5, BERT, and UL2, along with retrieval-augmented methods to handle rule-based QA in a niche domain. Unlike traditional QA tasks, board game rules span multiple sections, requiring multi-hop reasoning and advanced retrieval techniques to navigate structured yet ambiguous language.

Contributions:

1. Dataset Creation: A custom corpus of 500+ Q&A pairs and cleaned rulebooks for three popular games.
2. Comparative Model Analysis: Evaluation of multiple NLP models in both baseline and optimized configurations.

3. Insights into Retrieval-Augmented QA: Exploration of dense retrieval, reranking, and hyperparameter optimization.

This paper presents the development and evaluation of Rules Lawyer, providing insights into addressing rule-based QA tasks effectively.

Background and Related Work

Recent advancements in natural language processing (NLP) have significantly enhanced the capabilities of question-answering (QA) systems. Early work like **SQuAD (Rajpurkar et al., 2016)** established extractive approaches for QA tasks but struggled with multi-hop reasoning, where answers span multiple contexts. **Retrieval-augmented generation (RAG)**, introduced by **Lewis et al. (2020)**, addressed this by integrating dense retrieval systems like FAISS with generative models to provide context-aware responses.

Models such as **T5 (Raffel et al., 2020)** further advanced QA capabilities through a text-to-text framework, excelling in both generative and extractive tasks. Dense retrieval techniques, including **FAISS (Johnson et al., 2019)**, have become essential for ranking semantically relevant passages, particularly in multi-hop contexts where lexical matching (e.g., BM25) falls short.

Building on these advancements, this study focuses on applying dense retrieval, reranking, and fine-tuned language models to the niche domain of board game rules. Unlike datasets such as SQuAD or HotpotQA, rulebooks often require synthesizing information from structured yet ambiguous text, presenting unique challenges for retrieval and reasoning.

Methodology

Dataset Creation

The corpus for this project comprised cleaned rulebooks, supplemental documentation, and online FAQs for *Dune: Imperium*, *Carcassonne*, and *Star Wars: The Deckbuilding Game*. Additionally, a custom dataset of 500+ question-answer pairs was created, covering gameplay scenarios, edge cases, and multi-step interactions. These pairs were split into training (90%) and validation (10%) sets.

Preprocessing Steps:

1. Sliding Window Splits: Rulebook text was divided into overlapping passages (200-word windows with 50-word overlaps) to preserve context.
2. Normalization: Text was lowercased, punctuation standardized, and abbreviations expanded (e.g., *"EM" → "Emperor"*).
3. Corpus Embeddings: Semantic embeddings were generated using the all-mpnet-base-v2 model for dense retrieval.

Retrieval Approach

Baseline retrieval relied on the sliding window approach, which selected passages based on keyword matches. However, this struggled with long or multi-hop queries. To improve context relevance, the pipeline included:

- Semantic Retrieval: Dense embeddings ranked passages by semantic similarity.
- Reranking: A cross-encoder (MiniLM) reranked the retrieved passages, optimizing relevance for QA inputs.

Models and Training

This study evaluated four NLP models:

1. T5-Base and T5-Large: Pretrained transformers fine-tuned for text-to-text QA tasks.
2. Flan-T5-Base: An instruction-tuned variant optimized for multi-task learning.
3. BERT: An extractive transformer used for span-based QA.
4. UL2: A unified framework capable of both generative and extractive tasks.

Training Setup:

- Baseline models were trained with fixed hyperparameters (e.g., learning rate = 3e-5, batch size = 4).
- Optimized configurations were derived using Optuna, which conducted 20 hyperparameter trials per model. ROUGE-L was used as the primary metric for tuning due to its emphasis on sentence fluency and coherence.

Evaluation Metrics

Model performance was assessed using:

1. ROUGE (1, 2, L): Evaluates n-gram overlap for word, bigram, and sentence fluency.
2. BLEU: Measures lexical precision, particularly for generative models.
3. BERTScore F1: Computes semantic similarity using embeddings of generated and reference answers.

These metrics provided a balanced assessment for both extractive and generative tasks. In addition, sample Q&A outputs were evaluated qualitatively to analyze model strengths, limitations, and error patterns.

Here’s the revised **Results and Analysis** section, integrating sample Q&A insights and highlighting key findings:

Results and Analysis

Quantitative Results

Baseline and optimized models were evaluated using ROUGE, BLEU, and BERTScore metrics. The optimized T5-Large model achieved the best overall performance, with scores of 81.47 (ROUGE-1), 74.19 (ROUGE-2), and 90.57 (BERTScore F1). Other models, such as UL2 and Flan-T5-Base, also demonstrated strong performance, particularly in handling complex multi-hop queries. Refer to Appendix A for detailed comparisons of baseline vs. optimized metrics.

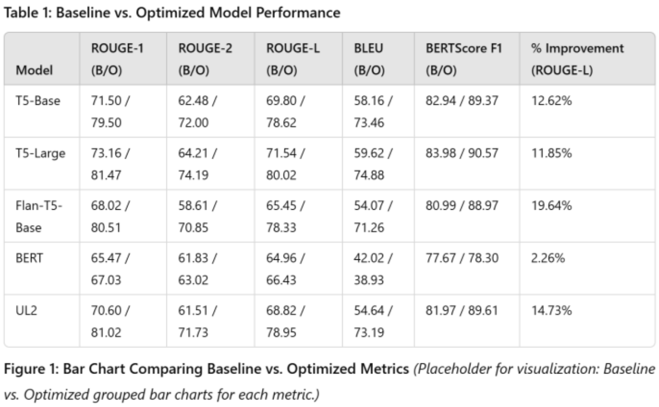


Table 1: Baseline vs. Optimized Model Performance

Model	ROUGE-1 (B/O)	ROUGE-2 (B/O)	ROUGE-L (B/O)	BLEU (B/O)	BERTScore F1 (B/O)	% Improvement (ROUGE-L)
T5-Base	71.50 / 79.50	62.48 / 72.00	69.80 / 78.62	58.16 / 73.46	82.94 / 89.37	12.62%
T5-Large	73.16 / 81.47	64.21 / 74.19	71.54 / 80.02	59.62 / 74.88	83.98 / 90.57	11.85%
Flan-T5-Base	68.02 / 80.51	58.61 / 70.85	65.45 / 78.33	54.07 / 71.26	80.99 / 88.97	19.64%
BERT	65.47 / 67.03	61.83 / 63.02	64.96 / 66.43	42.02 / 38.93	77.67 / 78.30	2.26%
UL2	70.60 / 81.02	61.51 / 71.73	68.82 / 78.95	54.64 / 73.19	81.97 / 89.61	14.73%

Figure 1: Bar Chart Comparing Baseline vs. Optimized Metrics *(Placeholder for visualization: Baseline vs. Optimized grouped bar charts for each metric.)*

Qualitative Analysis

Sample questions provided insights into model strengths and limitations. For example:

- Question: *“What happens if you lose influence with a faction and hold their Alliance token in Dune: Imperium?”*
  - UL2: Provided a detailed response including token transfer specifics: *“If you lose Influence and fall below the required level, you lose the Alliance token, which must be given to the player with the highest Influence.”*
  - T5-Large: Delivered a comprehensive, context-aware answer, also mentioning faction conflicts and Victory Points.
  - BERT: Failed entirely, producing an unrelated answer, highlighting limitations in handling nuanced queries.

However, all models struggled with certain edge cases:

- Question: *“What is the ability of the Twisted Mentat in Dune: Imperium?”*
  - All models incorrectly referred to the Mentat, a recurring resource, instead of the Twisted Mentat, a unique card, reflecting corpus gaps and entity confusion.

4.3. Error Analysis

Three key challenges emerged:

1. Ambiguity in Rules: Implicit terms and unclear definitions led to incomplete or inconsistent answers.
2. Disconnected Passages: Multi-hop reasoning failed when critical information was spread across non-adjacent sections.
3. Dataset Size: Limited examples constrained model generalization, especially for rare gameplay scenarios.

Despite these challenges, retrieval improvements from semantic embeddings and reranking significantly enhanced context selection, particularly for generative models. Appendix A highlights the importance of hyperparameters in influencing model performance.

Discussion

Key Insights

This study highlights the potential of integrating dense retrieval, reranking, and fine-tuned NLP models to address complex, rule-based QA tasks. Among the models evaluated:

- T5-Large achieved the best overall performance, excelling in coherence and fluency, particularly for multi-hop reasoning tasks.
- UL2 showcased strong semantic reasoning, delivering detailed answers in scenarios requiring nuanced understanding.
- BERT often failed in producing accurate answers for complex queries, confirming its limitations as an extractive-only model.

Optimizations primarily improved sentence-level fluency (e.g., ROUGE-L) but did not significantly alter the quality of model outputs. The type of model had a greater impact on performance than hyperparameter tuning alone.Comparative performance of optimized models is visualized in Appendix A.

Challenges

Several challenges emerged during the study:

1. Corpus Gaps: Errors in entity distinction (e.g., “Twisted Mentat” vs. “Mentat”) reflect the limitations of the dataset in representing unique game elements comprehensively.
2. Nuanced Reasoning: Even top-performing models missed critical details in ~50% of cases, highlighting the need for expanded datasets with greater diversity and depth.
3. Resource Constraints: The size and complexity of models like UL2 posed computational challenges, limiting the scope of hyperparameter optimization.

Implications

The findings from this project have both practical and research implications:

1. Practical Applications: Systems like Rules Lawyer can serve as real-time assistants during gameplay, providing players with dynamic and accurate rule clarifications. This could enhance the accessibility of complex board games and reduce learning curves.
2. Contributions to QA Research: This study demonstrates the feasibility of applying retrieval-augmented QA techniques to niche, rule-intensive domains. The methodology and insights offer a foundation for exploring similar structured text contexts, such as legal or policy documents.

Addressing the identified challenges through dataset expansion, hybrid retrieval methods, and task-specific tuning strategies could further improve the system’s accuracy and generalizability.

Conclusion and Future Work

This study introduced Rules Lawyer, a retrieval-augmented QA system designed to tackle the complexities of interpreting modern board game rulebooks. By integrating dense retrieval, cross-encoder reranking, and fine-tuned NLP models, the system demonstrated significant improvements in rule-based QA tasks. Among the models evaluated, T5-Large achieved the best overall performance, excelling in coherence and fluency, while UL2 showcased strong semantic reasoning capabilities.

The findings highlight the impact of retrieval augmentation and hyperparameter tuning in enhancing QA accuracy, with models like Flan-T5-Base achieving substantial percentage improvements over baseline metrics. Despite these successes, challenges such as computational constraints and dataset size limitations underscore the need for further refinement.

Future Work

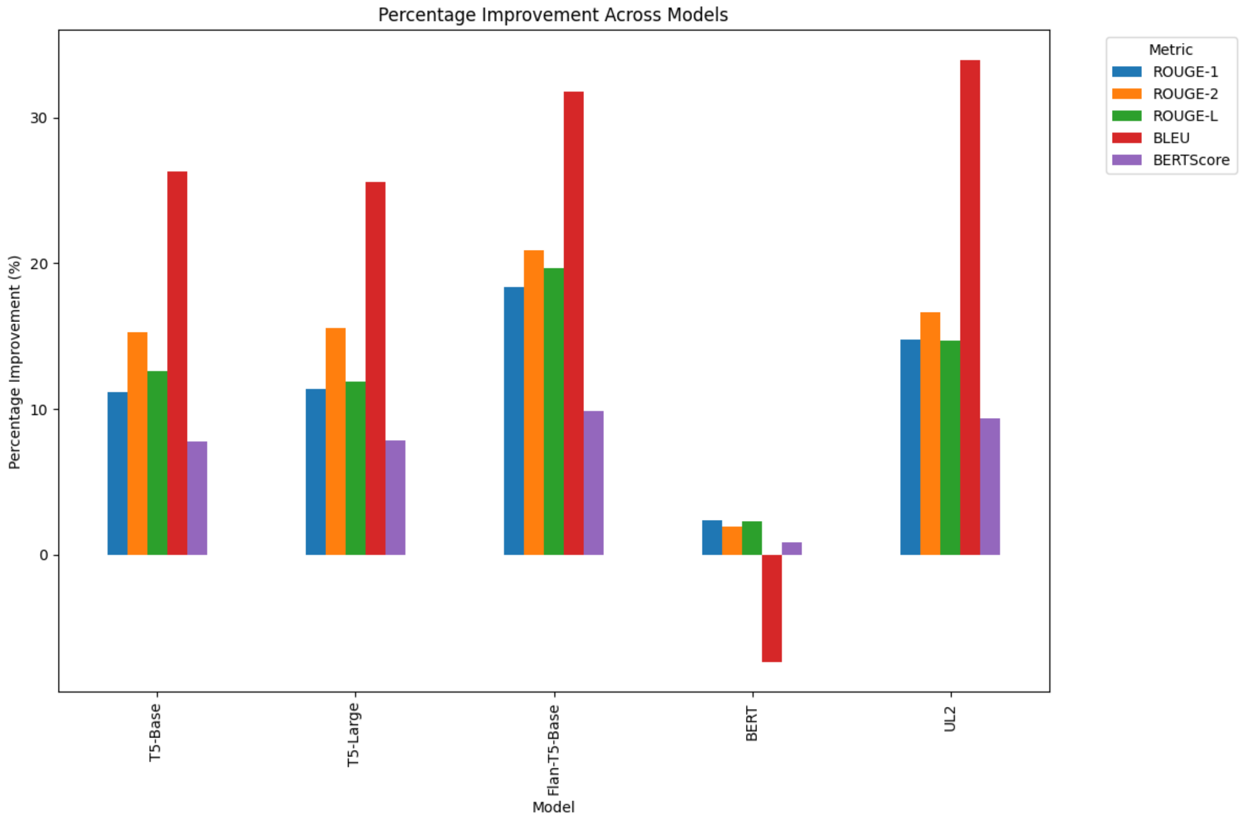
Future iterations of Rules Lawyer will focus on:

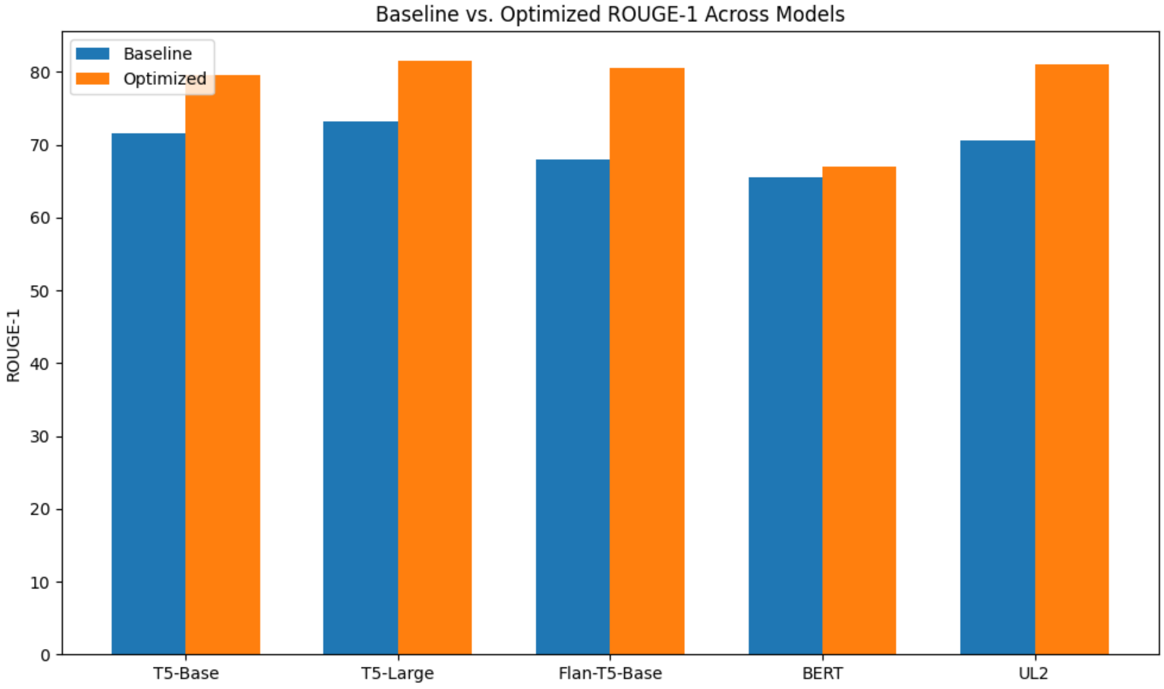
1. Dataset Expansion: Incorporating community-generated questions and additional rulebooks to improve model generalization and handle edge cases.
2. Hybrid Retrieval Methods: Combining lexical (e.g., BM25) and semantic retrieval to balance precision and recall.
3. Advanced Reranking: Exploring architectures like ColBERT for more robust passage scoring.
4. Real-Time RAG Pipelines: Developing end-to-end systems capable of dynamic retrieval and generation during gameplay.

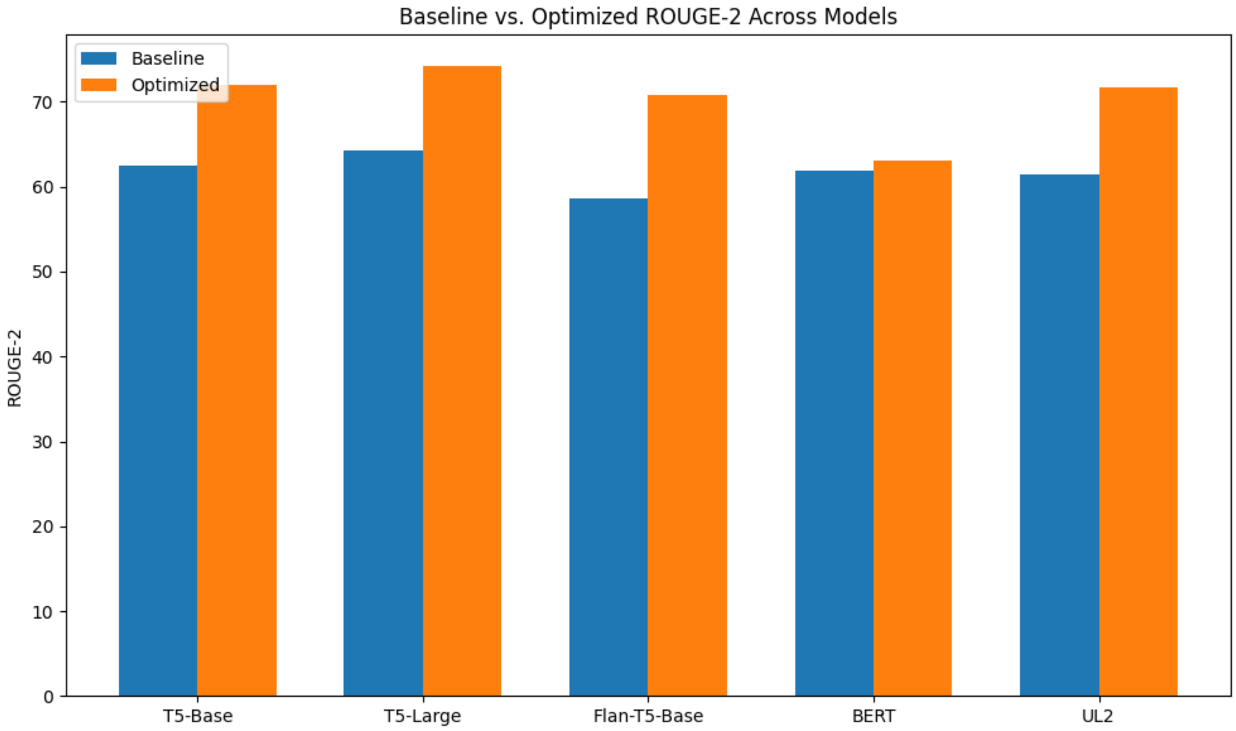
These advancements will extend Rules Lawyer into a comprehensive, real-time assistant for players, with potential applications in other rule-intensive domains such as legal or policy documentation. This work represents a step toward making complex rule systems more accessible and comprehensible through the power of NLP.

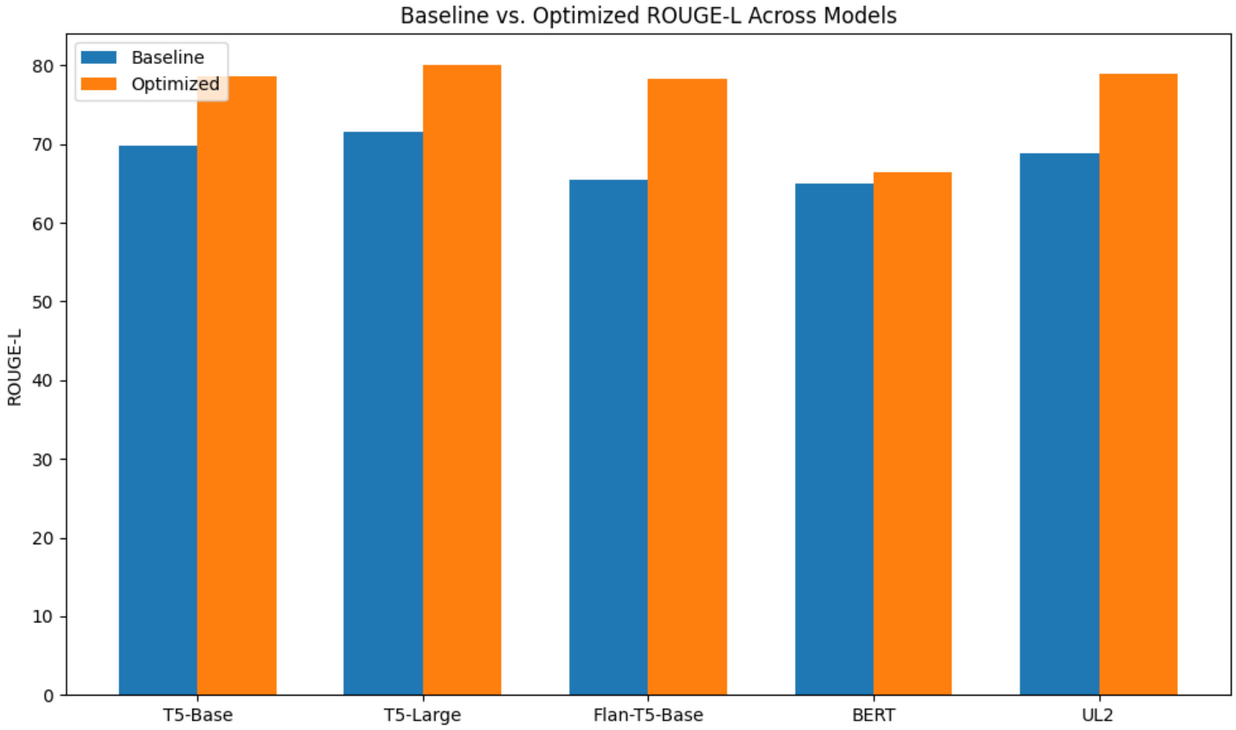
Appendix A: Visualizations

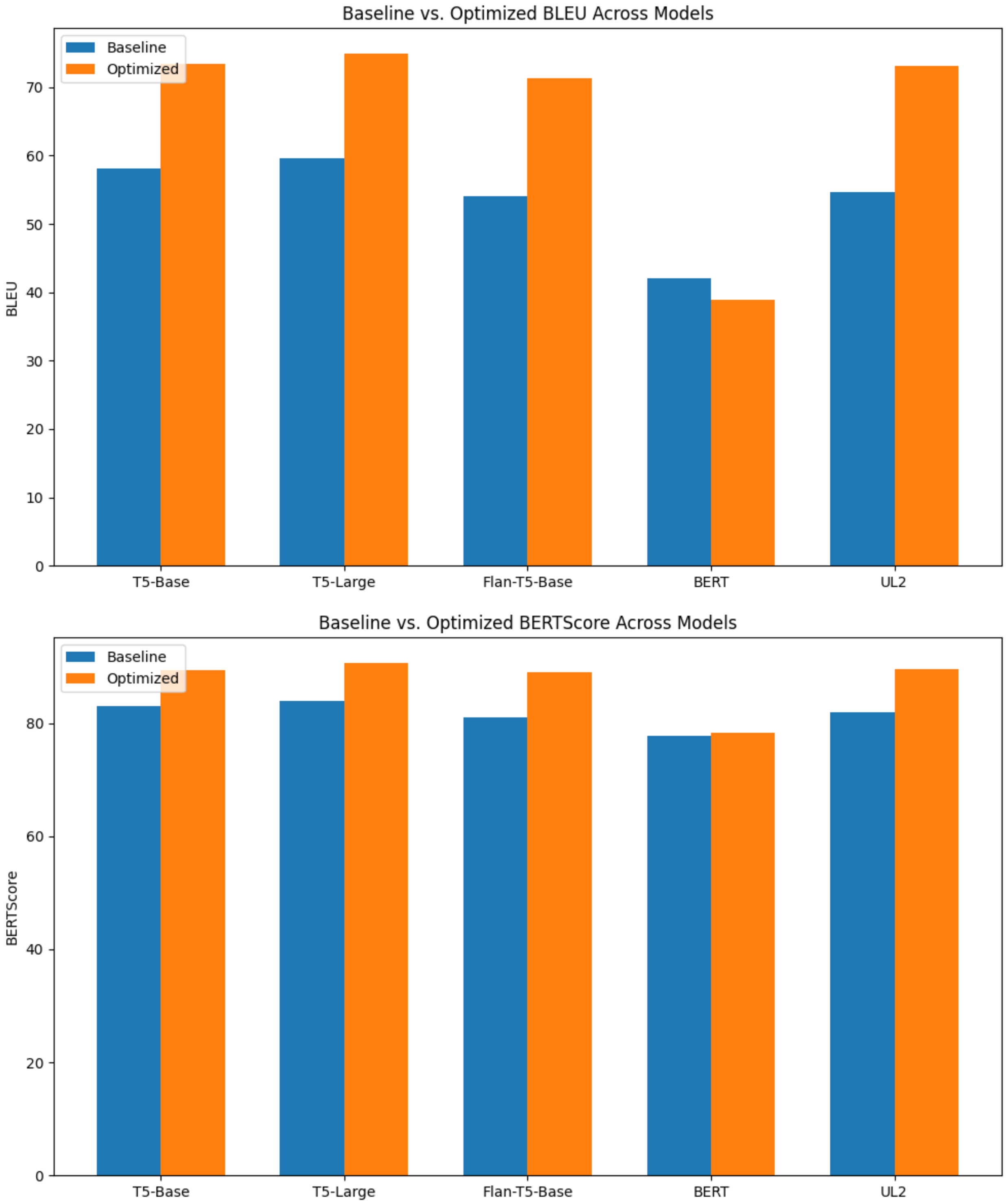
Figure A1: Baseline vs. Optimized Metrics






Baseline vs. Optimized ROUGE-2 Across Models
Baseline
Optimized
ROUGE-2
0
10
20
30
40
50
60
70
T5-Base
T5-Large
Flan-T5-Base
BERT
UL2

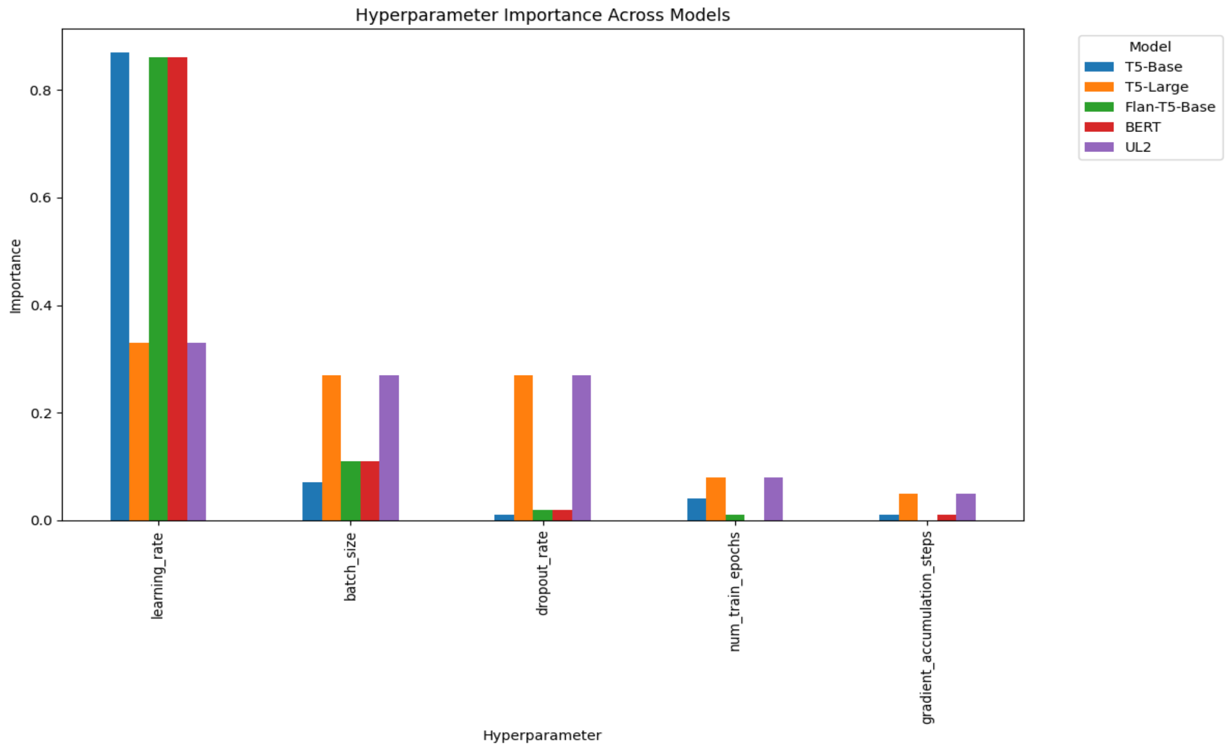

Baseline vs. Optimized ROUGE-L Across Models
Baseline
Optimized
ROUGE-L
0
10
20
30
40
50
60
70
80
T5-Base
T5-Large
Flan-T5-Base
BERT
UL2



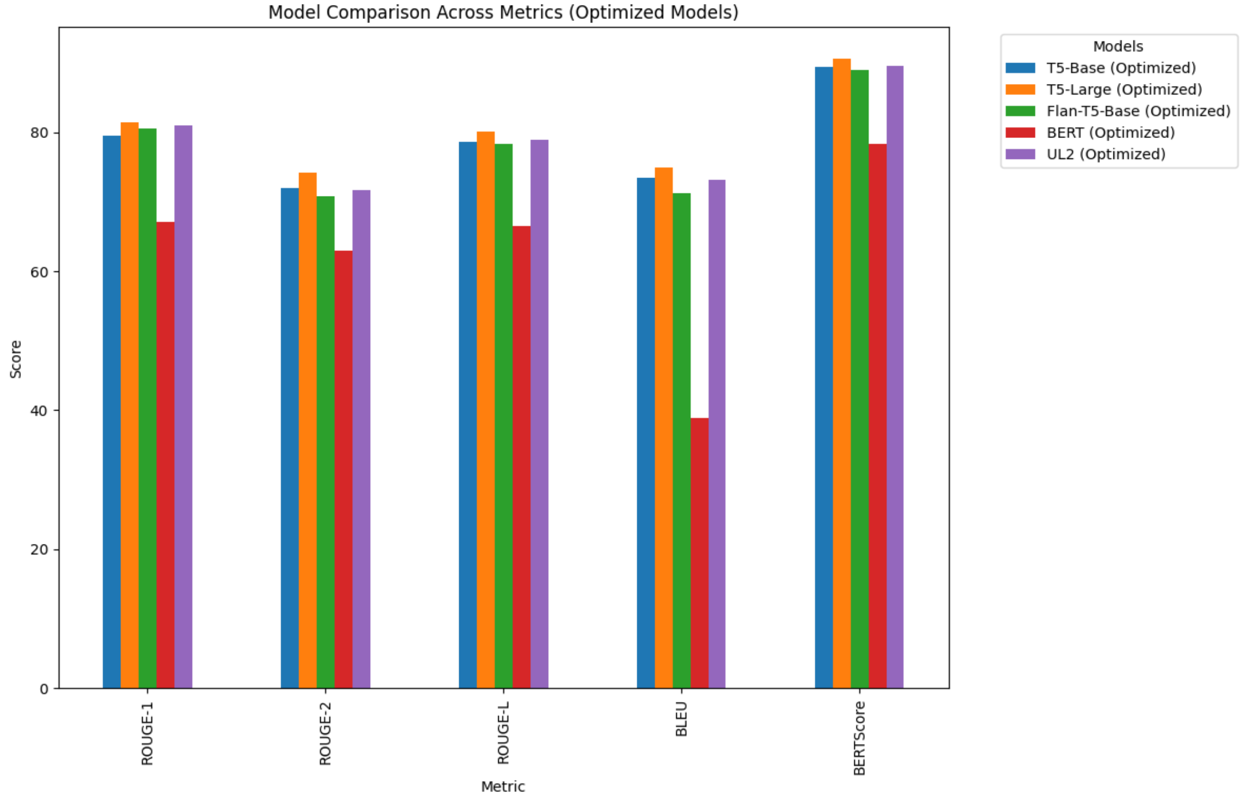
Baseline vs. Optimized BLEU Across Models

Baseline vs. Optimized BERTScore Across Models

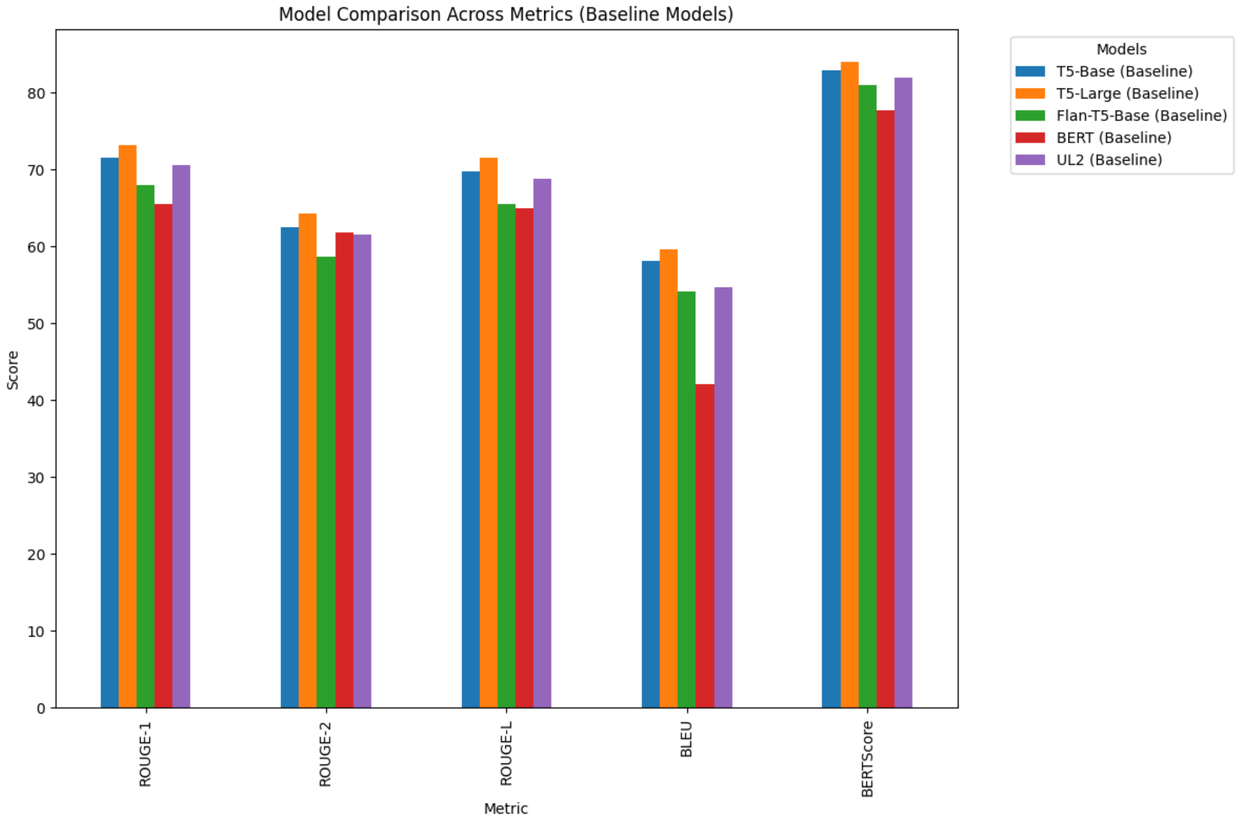
Figure A2: Hyperparameter Importance Across Models



Optimized Model Performance Comparison



Baseline Model Performance Comparison


Model Comparison Across Metrics (Baseline Models)
Models
T5-Base (Baseline)
T5-Large (Baseline)
Flan-T5-Base (Baseline)
BERT (Baseline)
UL2 (Baseline)
Score
0
10
20
30
40
50
60
70
80
ROUGE-1
ROUGE-2
ROUGE-L
BLEU
BERTScore
Metric

Works Cited and References

1. Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv preprint arXiv:1606.05250*. Retrieved from https://arxiv.org/abs/1606.05250
2. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., … Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint arXiv:2005.11401*. Retrieved from https://arxiv.org/abs/2005.11401
3. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., … Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research, 21*(140), 1–67. Retrieved from https://arxiv.org/abs/1910.10683
4. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*. doi:10.1109/TBDATA.2019.2921572. Retrieved from https://arxiv.org/abs/1702.08734
5. Tay, Y., Tran, V. Q., Ruder, S., Gupta, J., Dehghani, M., Qin, Z., … Metzler, D. (2022). Unifying Language Learning Paradigms. *arXiv preprint arXiv:2205.05131*. Retrieved from https://arxiv.org/abs/2205.05131
6. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*. Retrieved from https://arxiv.org/abs/1908.10084
7. Robertson, S. E., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval, 3*(4), 333–389. doi:10.1561/1500000019. Retrieved from https://dl.acm.org/doi/10.1561/1500000019