Performance Evaluation of RAG Models: A Comparative Study

Adam Fittler – s4696807

University of Queensland

1. INTRODUCTION

In modern information retrieval and natural language generation tasks, ranking models are essential for sorting search results or generated text based on relevance or other criteria. Re-ranking models are employed to refine the initially ranked results, ensuring improved relevance, coherence, or other qualities. The interplay between these rankers and rerankers is becoming increasingly crucial in large language models (LLMs) used for text generation, especially when fine-tuning outputs for specific contexts or user preferences.

The primary goal of this report is to evaluate the performance of various ranking and reranking model combinations in conjunction with state-of-the-art text generators. Specifically, this study focuses on evaluating the effectiveness of three rankers and three rerankers when applied to four different text generators. These models are assessed on their performance in Precision, Recall, F1 *[1],* METEOR *[13]* as well as ROUGE-L F1 *[17]* (longest subsequence) and BERT Score *[4]* with the goal of identifying the most and least effective combinations.

Three widely recognized metrics were chosen for evaluation: ROUGE-L, METEOR, and BERT Score. ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) measures the overlap between the generated and reference texts based on the longest common subsequence. METEOR (Metric for Evaluation of Translation with Explicit ORdering) complements ROUGE-L by focusing on semantic alignment, considering synonymy, stemming, and word order. BERT Score goes beyond surface-level matching by evaluating the similarity of embeddings from pre-trained BERT models, allowing for a more nuanced assessment of semantic similarity. These metrics were selected to accurately reflect the relevance, quality, and semantic alignment of the generated text.

---

2. METHODOLOGY

The three rankers and three rerankers used in this study each bring distinct methodologies for evaluating the relevance of generated text:

- **Ranker A** (llm-embedder) [3]: Leverages LLM embeddings to match semantically relevant outputs.
- **Ranker B** (instructor-base) [6]: A model fine-tuned for instructional prompts, ranking outputs based on alignment with expected instructional content.
- **Ranker C** (all-mpnet-base-v2) [18]: A state-of-the-art sentence transformer model that ranks outputs based on contextual relevance and similarity of terms.
- **Reranker A** (bge-reranker-base) [2]: Focuses on reranking based on contextual embeddings, refining the outputs further after the initial ranker.

- **Reranker B** (jina-reranker-v2): A multilingual reranker designed to evaluate output based on both relevance and diversity across languages *[8]*.
- **Reranker C** (mxbai-rerank-base-v1): Prioritizes novelty and diversity when re-ranking, ensuring that generated text is both relevant and diverse *[15]*.

Four advanced text generators were chosen for this study:

- **Generator A** (Gemma-2-2b): A smaller yet efficient instruction-based model, optimized for multilingual and instructional content generation *[5]*.
- **Generator B** (Llama-2-7b): Meta's conversational language model, specifically fine-tuned for chatbot interactions and coherent, contextually relevant content generation *[12]*.
- **Generator C** (Mistral-7B): A lightweight yet highly capable model, designed to handle instruction-based prompts with enhanced accuracy and responsiveness *[14]*.
- **Generator D** (Qwen2.5-7B): A large-scale instruction-tuned model known for generating high-quality text across diverse contexts and tasks, leveraging an efficient quantized format for faster processing without compromising accuracy *[16]*.

The study evaluated nine combinations of rankers and rerankers in the initial retrieval stage to identify the best and worst performers. Four key metrics were used:

- **Hits@10** and **Hits@4** measure how often a relevant result appears in the top 10 and top 4, respectively *[7]*.
- **MAP@10** (Mean Average Precision) assesses the precision of relevant results within the top 10, considering their ranking *[9]*.
- **MRR@10** (Mean Reciprocal Rank) captures how early the first relevant result appears within the top 10 *[10]*.

The best and worst combinations were chosen to be used as baselines for the generators, to see to see how this affected each. Each combination's outputs were analysed using Precision, Recall, F1, METEOR, ROUGE-L F1 and BERT Score, with the goal of identifying combinations that performed best and worst. Results were analysed based on overall performance as well as per-query variations to evaluate consistency.

---

3. EXPERIMENTAL RESULTS

The Table One summarises the stage one results for the ranker-reranker retrieval combinations.

Ret	Hits@10	Hits@4	MAP@10	MRR@10
A-A	0.620	0.541	0.214	0.493
A-B	0.621	0.544	0.211	0.486
A-C	0.587	0.478	0.179	0.412
B-A	0.621	0.542	0.208	0.480
B-B	0.622	0.546	0.215	0.488
B-C	0.586	0.467	0.176	0.409
C-A	0.565	0.492	0.195	0.443
C-B	0.567	0.502	0.197	0.448
C-C	0.539	0.441	0.168	0.389

Table 1: *Retrieval Results*

Overall, we can see that the Hits@10 and Hits@4 ranged from between 53.9% - 62.2%

and 44.1% and 54.6%. Indicating that for most of the retrievers the majority of the documents within the top ten are relevant to the query, however this is not always the case for the top four documents.

The evaluation results indicate that the B-B retrieval model is the best-performing configuration among the tested models. Approximately 62% of the top ten retrieved with this model documents were relevant to the query. Additionally, over 54% of the top four results were relevant. On average for each query only 20% of the top 10 documents are related to the query. Considering the relatively high average Hits@10 and 4 scores seen earlier this indicates the need to explore the results on a per query or per query type level in order to understand the relationship here. Furthermore, this trend seems to be the case across all of the retrieval results reinforcing the need to evaluate on a deeper level. The MRR reflects that the most relevant documents are on average found around the second position in the ranking.

In contrast, the worst-performing model is the C-C, with 53.9% of the top ten documents being relevant and 44.1% of the top four documents. It was also the worst performing comparatively in terms of MAP and MRR scores being 16.8% and 38.9% respectively.

Moving into stage two the B-B and C-C retrievers will be used as the two baselines for the generators in order to explore how the different generators are affected y the differing quality of retrieval and see the effect this has on their performance. It was observed that the Precision, Recall and F1 scores calcualted based on word overlap directly reflected the trends captured by METEOR and ROUGE-L F1 scores and have therefore been omitted here, but can be found in Appendix One.

Table Two shows the overall results for each of the 8 RAG combinations of retrival and generator stages. It outlines the average METEOR Score (M), ROUGE-L F1 Score (RLF1), as well as the BERT Precision and Recall (BP & BR) across all of the queries. The full results across all query types and metrics can be found in Appendix One.

RAG	M	RLF1	BP	BR
B-B-A	0.17	0.25	0.82	0.87
C-C-A	0.14	0.23	0.81	0.87
B-B-B	0.19	0.23	0.84	0.88
C-C-B	0.17	0.20	0.83	0.87
B-B-C	0.08	0.06	0.80	0.85
C-C-C	0.07	0.06	0.80	0.85
B-B-D	0.29	0.42	0.88	0.92
C-C-D	0.26	0.38	0.87	0.91

Table 2: *Generator Results*

From these results it is evident that the best performing RAG combination was B-B-D Using a combination of the best retriever as well as the Qwen2.5 7B parameter model. It has the best scores across all of the metrics, most import of which being the BERT Precision and Recall with 87% and 91% respectively indicating indicating a high level of semantic alignment between the generated and reference texts. This suggests that the model consistently produces contextually relevant and accurate outputs, closely matching the intended meaning in both precision (correctness of the generated text) and recall (completeness of the response).

However, the story told by the METEOR Score and ROUGE-L F1 Score is quite different with scores of 29% and 42% respectively, indicating that while the generated text may semantically align with the reference, it struggles with exact word

choice, order, and surface-level similarity. This suggests that although the meaning is preserved (as captured by BERT Score), the outputs lack fluency and coherence in terms of structure and specific word overlap, which are critical factors in these metrics.

The worst performing RAG combination was C-C-C leveraging the weaker baseline retieval as well as the Minstral 7b model. It still achieved a 80% BERT Precsion as well as 85% BERT Recall, but the Minstral generators in general had the lowest METEOR and ROUGE-L Scores. In this case the RAG the scores were 7% and 6% respectively. Therefore, highlighting that while sentiment was relatively accurate the specific word and token choices did not align with the desired gold labels.

Overall between all of the RAG results thet BERT Scores are much higher than that of the METEOR and ROUGE Scores. Indicating that a more indepth analysis should be conducted into the specifics of the query results.

When analysing the results on a per query basic the following patterns emerge. Figure One below shows the overall per query box plot for METEOR, ROUGE and BERT Scores for the best RAG pipeline however the highlighted patterns were evident in all of the other tested models.

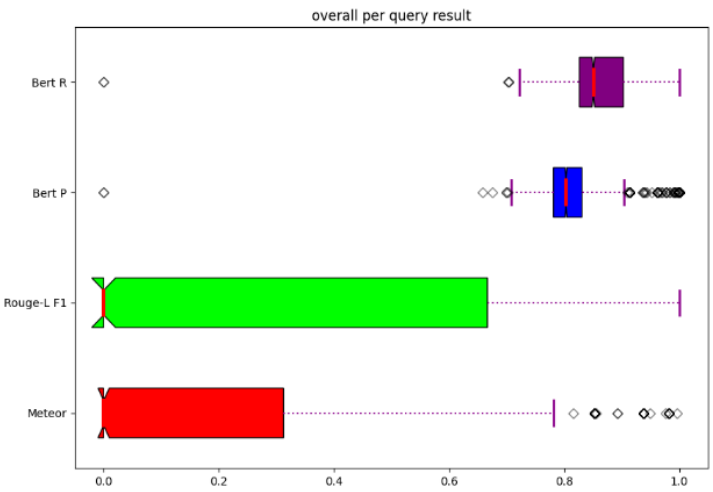

Figure 1: *RAG B-B-D overall per query*

The figure outlines The difference in distributions between the BERT Scores and the METEOR and ROUGE Scores. It can be seen that for the majority of the queries the METEOR and ROUGE Scores are very low with the mean around 0. Indicating low to no overlaps between the predicted and gold standard tokens. However, due to the semantic embedings used by BERT Score it accuractly identifies the accuracy of the response.

Beyond this when looking at the results form the different query types it can be seen that the METEOR and ROUGE Scores for inference queries are much higher than those seen for the comparison, temporal and null queries. Figure Two below highlights that for inference queries the ROUGE-L F1 and METEOR Scores are much more accurate than what was seen in the overall query results, while the BERT Score precision and recall distributions remained relatively similar. Thus, indicating that for the inference queries the models tended to return responses that more accurately

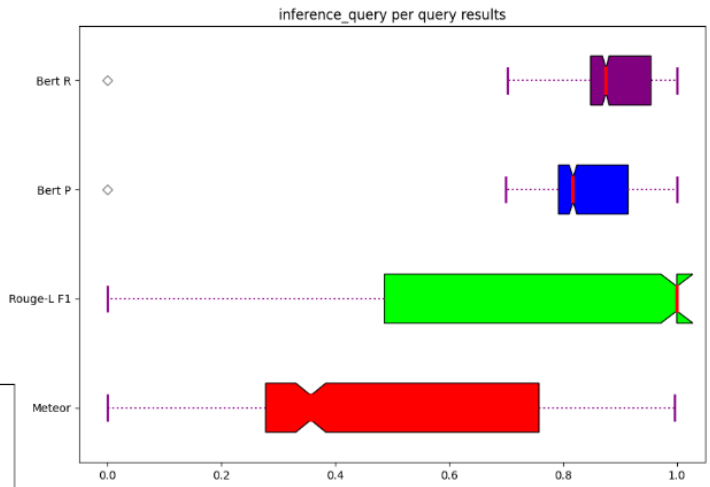

Figure 2: *RAG B-B-D inference per query*

matched the extact characters/token for the gold labels.

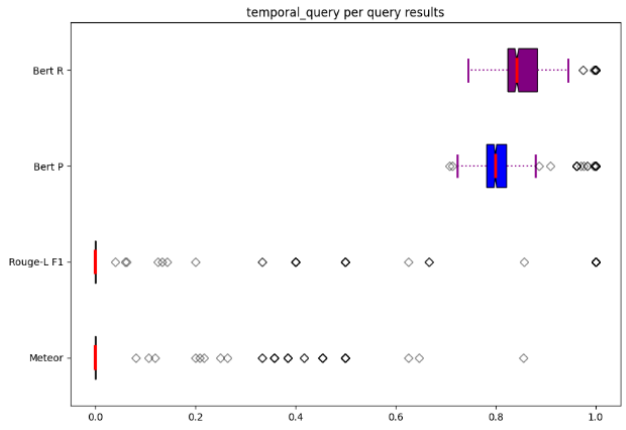

Figure 3: *RAG B-B-D temporal per query*

However, when looking at the other query types for example the temporal query results as seen in Figure Three, the ROUGE-L and METEOR scores effectively drop to 0 while the BERT scores remain relatively similar to the seen in the inference query results. Therefore indicating that for the other query types, that being comparison, null and temportal, the model had the correct sentiment in the majority of the results but the extact words and phrases utilised in the responses drastically different from that seen in the gold label solutions.

In terms of the individual models performances as mentioned Qwen performed the best on both of the retrieval baselines, followed by Llama2, then Gemma and finally Minstral had the weakest performance. It is interesting to highlight that Gemma performed better than Minstral even through it was only the 2B parameter model as opposed to the Minstrals 7B parameters, indicating a strong architecture. Gemma models only have a 2B and 9B varient so it would be interesting to see how the 9B model performed against the rest in a future comparison.

Furthermore, In terms of the retrieval baselines affects on each of the models, the results were indeed as expected with all of the models performing worse with the lower scoring retrieval baseline. Therefore, highlighting the importance of choosing the best ranker reranker pair during the retrieval stage in order to achieve the best generation result possible.

---

4. CONCLUSION

This study evaluated the performance of various ranker, reranker, and large language model (LLM) combinations within a retrieval-augmented generation (RAG) framework. The B-B-D combination (best retriever with Qwen2.5-7B) delivered the strongest results, showing excellent semantic alignment. However, displayed challenges in exact word matching and sentence structure. The worst performer, C-C-C (Minstral-7B with the worst retriever), also displayed poor alignment in word overlap despite reasonable semantic accuracy. Inference queries consistently outperformed comparison, temporal, and null queries, underscoring variability in model effectiveness based on query type.

This study was limited model in selection, reducing the generalisability of the results. The focus on specific query types also highlighted potential bias, with models performing better on inference queries than more complex types. Lastly, the absence of a larger comparison between different model sizes, limits understanding of the relationship between model size and performance. Future research should address these limitations by expanding model selection, incorporating human evaluation, and diversifying datasets to achieve more comprehensive results.

5. REFERENCES

[1] K. P. Shung, "Accuracy, Precision, Recall or F1?," Medium. Accessed: Oct. 14, 2024. [Online]. Available: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

[2] "BAAI/bge-reranker-base · Hugging Face." Accessed: Oct. 14, 2024. [Online]. Available: https://huggingface.co/BAAI/bge-reranker-base

[3] "BAAI/llm-embedder · Hugging Face." Accessed: Oct. 14, 2024. [Online]. Available: https://huggingface.co/BAAI/llm-embedder

[4] "BERT Score - a Hugging Face Space by evaluate-metric." Accessed: Oct. 14, 2024. [Online]. Available: https://huggingface.co/spaces/evaluate-metric/bertscore

[5] "google/gemma-2-2b-it · Hugging Face." Accessed: Oct. 14, 2024. [Online]. Available: https://huggingface.co/google/gemma-2-2b-it

[6] "hkunlp/instructor-base · Hugging Face." Accessed: Oct. 14, 2024. [Online]. Available: https://huggingface.co/hkunlp/instructor-base

[7] "How Hits@k is used for evaluating missing link prediction in Knowledge Graphs? | 5 Answers from Research papers," SciSpace - Question. Accessed: Oct. 14, 2024. [Online]. Available: https://typeset.io/questions/how-hits-k-is-used-for-evaluating-missing-link-prediction-in-o5muhtou40

[8] "jinaai/jina-reranker-v2-base-multilingual · Hugging Face." Accessed: Oct. 14, 2024. [Online]. Available: https://huggingface.co/jinaai/jina-reranker-v2-base-multilingual

[9] "Mean Average Precision (MAP) in ranking and recommendations." Accessed: Oct. 14, 2024. [Online]. Available: https://www.evidentlyai.com/ranking-metrics/mean-average-precision-map

[10] Mean Reciprocal Rank (MRR) explained." Accessed: Oct. 14, 2024. [Online]. Available: https://www.evidentlyai.com/ranking-metrics/mean-reciprocal-rank-mrr

[12] “meta-llama/Llama-2-7b-chat-hf · Hugging Face.” Accessed: Oct. 14, 2024. [Online]. Available: https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[13] “METEOR - a Hugging Face Space by evaluate-metric.” Accessed: Oct. 14, 2024. [Online]. Available: https://huggingface.co/spaces/evaluate-metric/meteor

[14] “mistralai/Mistral-7B-Instruct-v0.1 · Hugging Face.” Accessed: Oct. 14, 2024. [Online]. Available: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1

[15] “mixedbread-ai/mxbai-rerank-base-v1 · Hugging Face.” Accessed: Oct. 14, 2024. [Online]. Available: https://huggingface.co/mixedbread-ai/mxbai-rerank-base-v1

[16] “Qwen/Qwen2.5-7B-Instruct-GPTQ-Int4 · Hugging Face.” Accessed: Oct. 14, 2024. [Online]. Available: https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-GPTQ-Int4

[17] “ROUGE - a Hugging Face Space by evaluate-metric.” Accessed: Oct. 14, 2024. [Online]. Available: https://huggingface.co/spaces/evaluate-metric/rouge

[18] “sentence-transformers/all-mpnet-base-v2 · Hugging Face.” Accessed: Oct. 14, 2024. [Online]. Available: https://huggingface.co/sentence-transformers/all-mpnet-base-v2

6. APPENDIX ONE – RAG RESULTS TABLES

LLM	Retrieval	Precision	Recall	F1	Meteor	Rouge- L F1	Bert Precision	Bert Recall
Gemma	**B-B**	0.28	0.24	0.25	0.17	0.25	0.82	0.87
Gemma	**C-C**	0.24	0.23	0.23	0.14	0.23	0.81	0.87
Llama2	**B-B**	0.42	0.23	0.23	0.19	0.23	0.84	0.88
Llama2	**C-C**	0.39	0.19	0.20	0.17	0.20	0.83	0.87
Minstral	**B-B**	0.29	0.05	0.06	0.08	0.06	0.80	0.85
Minstral	**C-C**	0.25	0.04	0.05	0.07	0.06	0.80	0.85
Qwen	**B-B**	0.42	0.42	0.42	0.29	0.42	0.88	0.92
Qwen	**C-C**	0.38	0.38	0.38	0.26	0.38	0.87	0.91

Table 3: *Overall Generation Results*

LLM	Retrieval	Precision	Recall	F1	Meteor	Rouge- L F1	Bert Precision	Bert Recall
Gemma	**B-B**	0.76	0.65	0.69	0.46	0.69	0.85	0.89
Gemma	**C-C**	0.63	0.63	0.63	0.40	0.63	0.85	0.88
Llama2	**B-B**	0.83	0.68	0.69	0.55	0.69	0.94	0.96
Llama2	**C-C**	0.73	0.57	0.58	0.48	0.59	0.92	0.94
Minstral	**B-B**	0.79	0.06	0.10	0.21	0.11	0.80	0.85
Minstral	**C-C**	0.69	0.05	0.08	0.18	0.09	0.79	0.84
Qwen	**B-B**	0.85	0.86	0.85	0.62	0.85	0.97	0.97
Qwen	**C-C**	0.73	0.73	0.73	0.53	0.73	0.95	0.95

Table 4: *Inference Generation Results*

LLM	Retrieval	Precision	Recall	F1	Meteor	Rouge- L F1	Bert Precision	Bert Recall
Gemma	**B-B**	0.04	0.02	0.03	0.02	0.03	0.79	0.85
Gemma	**C-C**	0.03	0.02	0.02	0.01	0.02	0.80	0.87
Llama2	**B-B**	0.20	0.01	0.02	0.02	0.02	0.78	0.84
Llama2	**C-C**	0.21	0.01	0.02	0.02	0.02	0.78	0.84
Minstral	**B-B**	0.02	0.00	0.00	0.00	0.00	0.78	0.85
Minstral	**C-C**	0.02	0.00	0.00	0.00	0.00	0.78	0.85
Qwen	**B-B**	0.09	0.09	0.09	0.05	0.09	0.81	0.88
Qwen	**C-C**	0.09	0.09	0.09	0.04	0.09	0.81	0.88

Table 5: *Comparison Generation Results*

LLM	Retrieval	Precision	Recall	F1	Meteor	Rouge- L F1	Bert Precision	Bert Recall
Gemma	**B-B**	0.12	0.11	0.11	0.08	0.12	0.80	0.86
Gemma	**C-C**	0.02	0.02	0.02	0.01	0.02	0.79	0.84
Llama2	**B-B**	0.06	0.01	0.01	0.02	0.01	0.81	0.85
Llama2	**C-C**	0.10	0.01	0.02	0.02	0.02	0.81	0.85
Minstral	**B-B**	0.23	0.22	0.22	0.15	0.22	0.85	0.88
Minstral	**C-C**	0.23	0.22	0.22	0.15	0.22	0.85	0.88
Qwen	**B-B**	0.87	0.87	0.87	0.56	0.87	0.96	0.96
Qwen	**C-C**	0.86	0.86	0.86	0.56	0.86	0.96	0.96

Table 6: *Null Generation Results*

LLM	Retrieval	Precision	Recall	F1	Meteor	Rouge- L F1	Bert Precision	Bert Recall
Gemma	**B-B**	0.06	0.03	0.04	0.03	0.04	0.80	0.86
Gemma	**C-C**	0.10	0.07	0.08	0.04	0.08	0.81	0.87
Llama2	**B-B**	0.35	0.02	0.02	0.03	0.02	0.78	0.84
Llama2	**C-C**	0.31	0.02	0.03	0.03	0.03	0.78	0.84
Minstral	**B-B**	0.01	0.00	0.00	0.00	0.00	0.79	0.86
Minstral	**C-C**	0.01	0.00	0.00	0.00	0.00	0.79	0.86
Qwen	**B-B**	0.07	0.07	0.07	0.04	0.07	0.81	0.87
Qwen	**C-C**	0.06	0.06	0.06	0.03	0.06	0.80	0.87

Table 7: *Temporal Generation Results*

7. APPENDIX TWO – QWEN + BEST (LEFT) / WORST (RIGHT) RETRIEVAL PER QUERY RESULTS

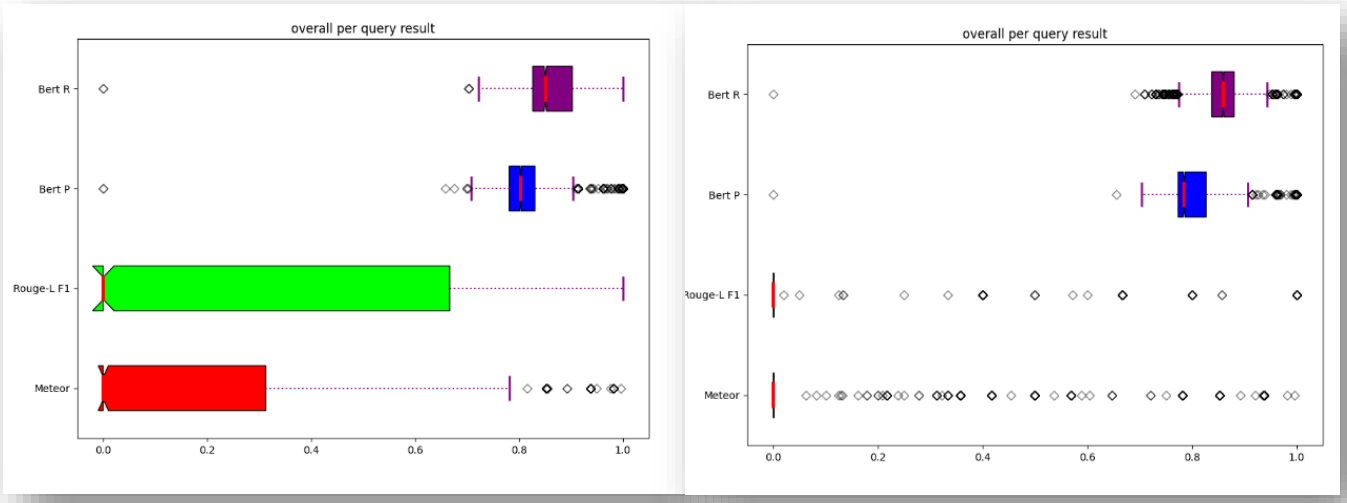

Figure 4: *Overall per query results*

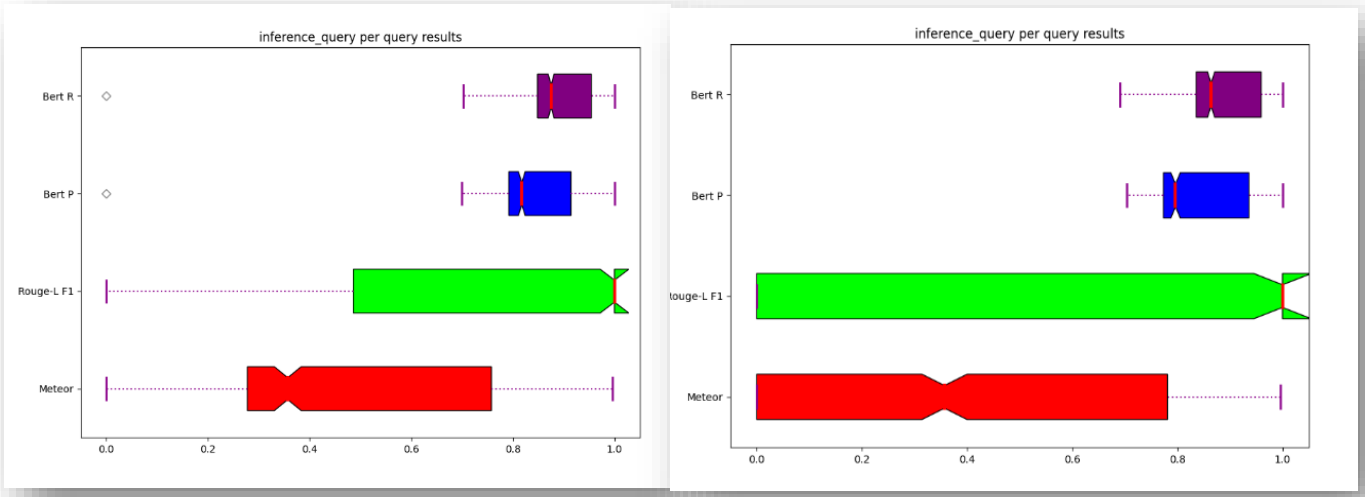

Figure 5: *Inference per query results*

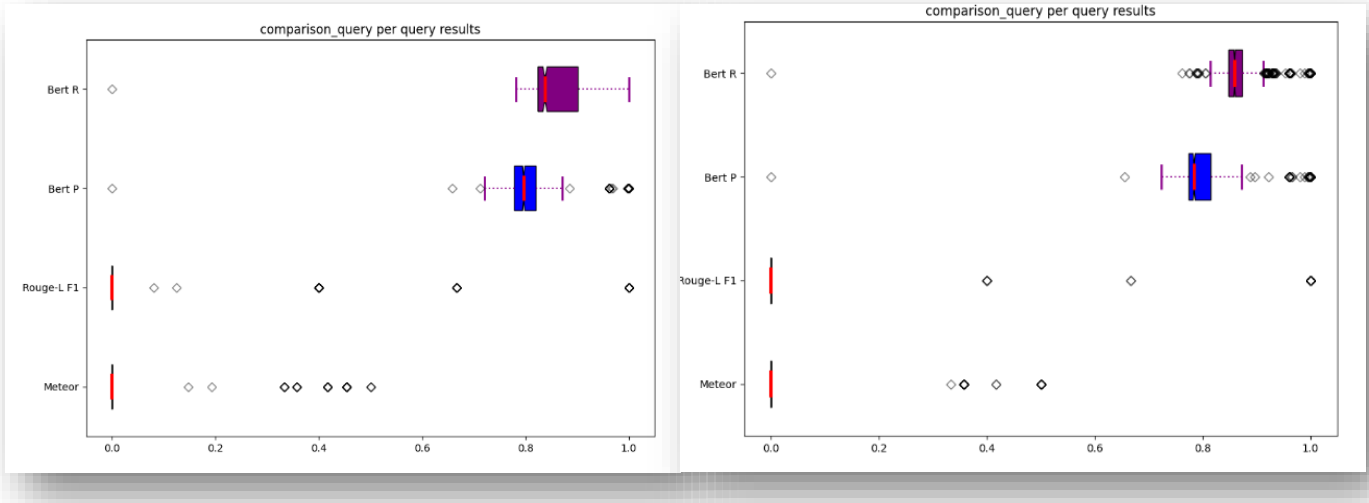

Figure 6: *Comparison per query results*

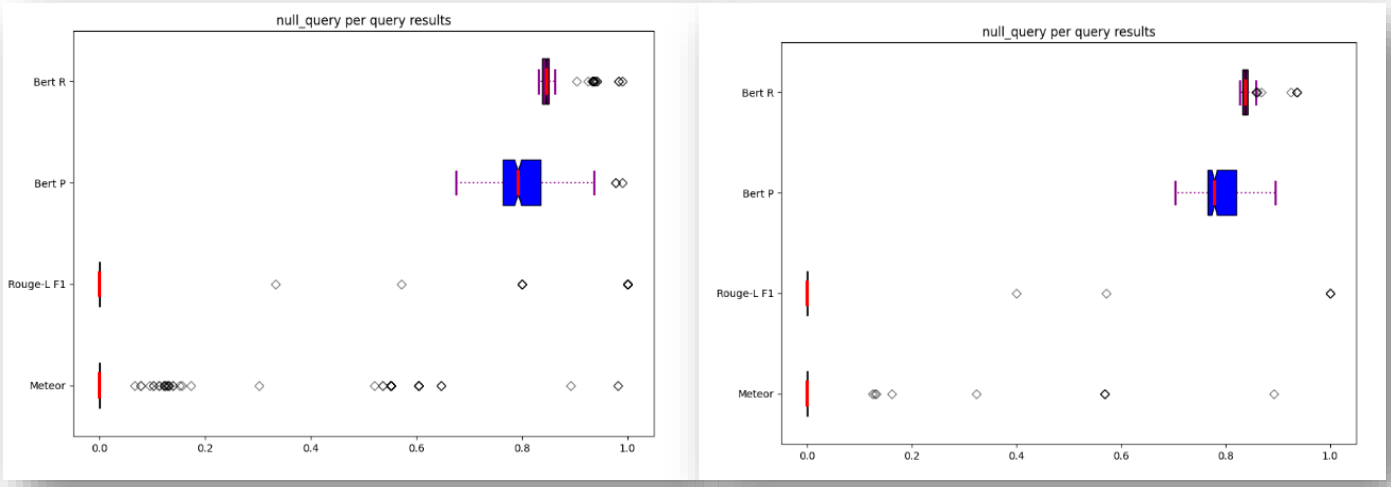

Figure 7: *Null per query results*

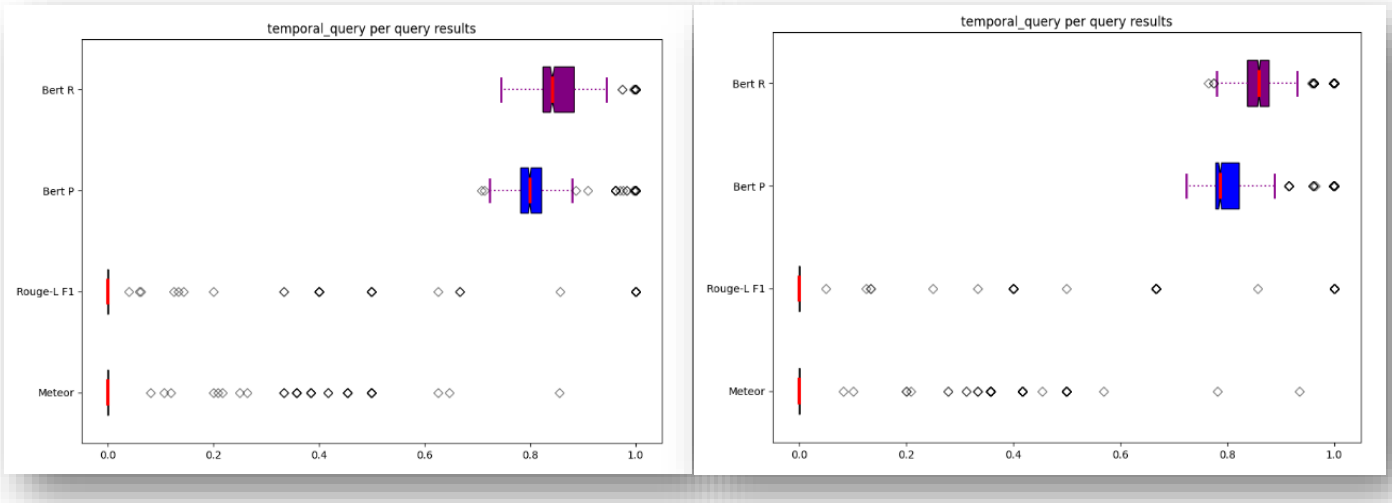

Figure 8: *Temporal per query results*