## Metrics we use

To evaluate RAG systems effectively, we use a framework called Retrieval Augmented Generation Assessment (RAGAs). This framework aims to measure the two dimensions of a RAG system, the retriever and the generator. To achieve this, RAGAs utilizes an LLM agent to analyze the context, the question, the RAG system's response, and the ground truth labels to extract the components necessary for computing each metric. In the context of this experiment, with the resources available, the gpt-3.5-turbo-1106 model is employed as the LLM for RAGAs evaluation.

Additionally, the following metrics are chosen: Context Recall (CR), Context Precision (CP), Faithfulness (FA), and Answer Similarity (AS), Semantic Similarity(SS), Accuracy. The selected metrics all return a score ranging from 0 to 1, with higher values denoting better performance. In the experiment, the metrics are calculated for each of the 500 entries in the test set. The mean values of these scores are then computed to assess the average performance of the model over the entire test set.

**Context Recall (CR)** CR measures the extent to which the retrieved context aligns well with the ground truth answer. In the context of the experiment, this essentially computes the degree of relevance of the retrieved resumes to the ground truth resume. This metric can directly measure the retriever component's ability to search for candidates similar to the ideal resume.

$$\text{Context Recall} = \frac{|\text{Ground Truth sentences that can be found in context}|}{|\text{Number of sentences in Ground Truth}|}$$

**Context Precision (CP)** CP measures whether the ground-truth relevant details present in the contexts are ranked higher or not. In an ideal scenario, the relevant documents should be at the top ranks.

CP can evaluate whether the retriever component ranks resumes with more relevant details to the ground truth higher among all documents. This helps with the assessment of the precision of the retriever component's ranking ability.

$$\text{Context Precision}_K = \frac{\sum_{k=1}^{K}(\text{Precision}_k \times v_k)}{\text{Number of relevant documents in the top } K \text{ results}}$$

$$\text{Precision}_k = \frac{\text{True Positives}_k}{\text{True Positives}_k + \text{False Positives}_k}$$

**Faithfulness (FA)** FA computes the consistency of the answer to the provided context. According to RAGAs' definition, the generated answer is considered faithful if claims made in the answer can be inferred from the context.

FA measures the generator's faithfulness and consistency to the retrieved context, which helps to determine whether the LLM component is prone to hallucinations or not. This is especially important in a knowledge-intensive task like resume screening, where the processing of candidates' information is often required to be precise.

$$\text{Faithfulness} = \frac{|\text{Claims in the answer that can be inferred from context}|}{|\text{Total number of claims in the answer}|}$$

**Answer Similarity (AS)**

AS measures the semantic similarity between the generated answer and the ground truth. This can be utilized to assess the LLM's summarization quality of the selected resumes.

$$\text{Answer Similarity} = \text{Cos Similarity}(\text{Embedding}_{\text{Answer}}, \text{Embedding}_{\text{Ground Truth}})$$

### Semantic Similarity

Semantic similarity is determined by the cosine similarity score between the vector representations of the selected resume and the ground truth resume.

$$\text{Similarity} = \text{Cos Similarity}(\text{Embedding}_{\text{Selected Resume}}, \text{Embedding}_{\text{Ground Truth}})$$

In essence, the metric measures the relevance of the selected resume to the ground truth resume. This helps to assess the ability to retrieve resumes relevant to the ground truth of the RAG system.

### Accuracy

The accuracy measures the portion of correct selections of ground truth resumes, which can be useful in determining whether the system can locate the exact matching resume.

$$\text{Accuracy} = \frac{\text{Correctly Predictions}}{\text{All Predictions}} = \frac{\sum_{i=0}^{n} \mathbb{1}(\text{Selected}_i = \text{Ground Truth}_i)}{n}$$

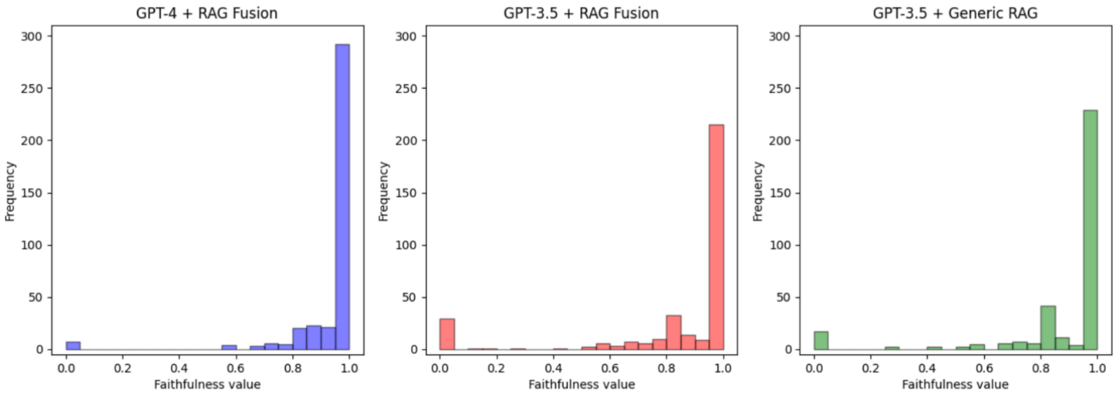
The first baseline is a RAG Fusion system utilizing gpt-3.5-turbo-1106 as the generator. It employs the same LLM agent for sub-questions generation tasks. Essentially, the baseline simply utilizes a less powerful model than GPT-4 as the LLM agent for multiple tasks, which can provide insights into how the LLM agent choice affects resume selection quality.

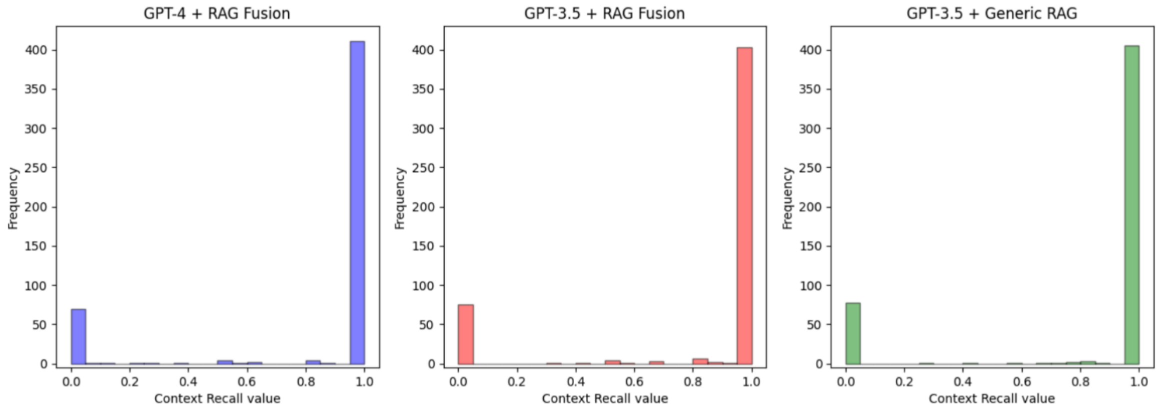
The second baseline system is a Naive RAG framework utilizing gpt-3.5- turbo-1106 as the generator. Besides the clear difference in the LLM agent choice, this RAG system utilizes just the generic similarity-based retrieval approach with no RRF re-ranking. It can be a meaningful baseline to understand whether a less powerful LLM agent with a simpler retriever approach can perform comparably with the more advanced models.

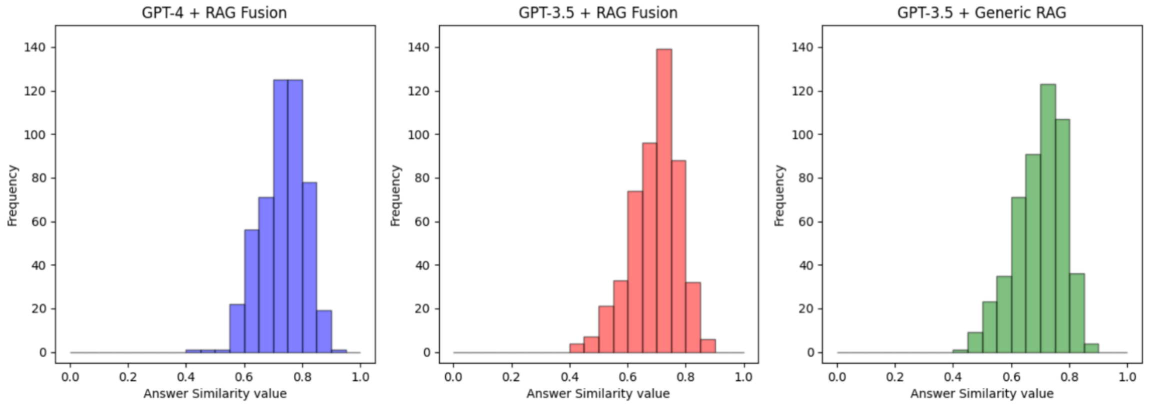
# Metrics calculate results

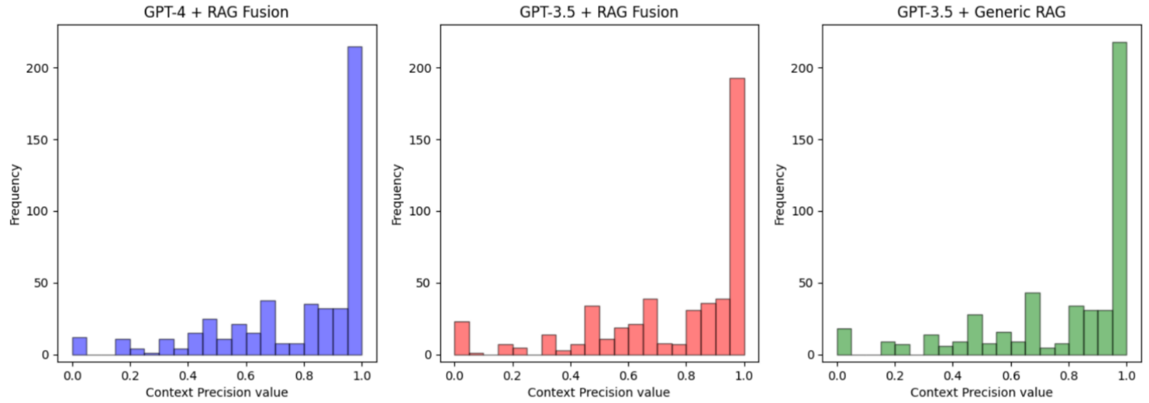
Calculate the scores of the 4 RAGAS metreics  and two other metrics for each model\framework and plot them in graphs.

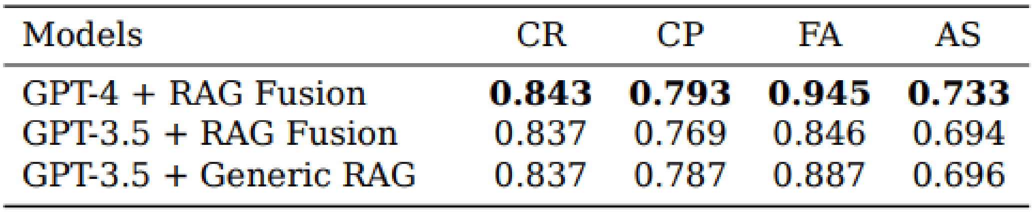
Extract the predicted resume ID from test_result_df and compare it with the target resume ID to calculate the accuracy.











| Models | CR | CP | FA | AS |
|---|---|---|---|---|
| GPT-4 + RAG Fusion | **0.843** | **0.793** | **0.945** | **0.733** |
| GPT-3.5 + RAG Fusion | 0.837 | 0.769 | 0.846 | 0.694 |
| GPT-3.5 + Generic RAG | 0.837 | 0.787 | 0.887 | 0.696 |

Overall, the distribution of the scores for the models for every metric is relatively similar. The similarity suggests a consistent performance of each model across all metrics, indicating that the evaluation process is relatively stable and balanced.

In addition, it is also clear to observe the same distribution trend, which is skewed towards higher values. For FA, CR, and CP particularly, a significantly large number of values is observed to be a perfect score of 1. In general, this result suggests that the performance of all models is mostly satisfactory so far.

For a better look at the performance of each component of each model over the whole test set, Table above presents the mean CR, CP, FA, and AS scores for the proposed model and the baseline models over the 500 job descriptions.

The CR of the proposed model to the baselines is slightly higher than the baselines. A higher CR means that the retriever is better at identifying and retrieving documents that are relevant to the ground truth resume. Even though this suggests that the GPT-4 + RAG Fusion retriever is more effective, it is clear that this difference is marginal. Despite a more advanced re-ranking approach, the proposed model does not provide a significant improvement to the recall of the context.

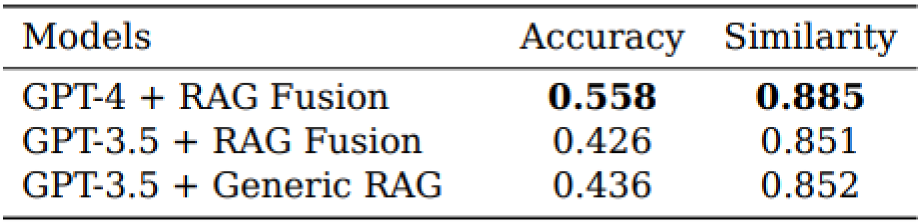
In terms of CP score, the GPT-4 + RAG Fusion model also performs relatively similarly to the Generic RAG model. Once again, this may reveal that the Fusion retriever model may not necessarily improve the final ranking results. Interestingly, the GPT-3.5 + RAG Fusion model exhibits a 2% decrease in performance compared to the other models. This can be due to various factors such as the GPT-3.5's less advanced language comprehension, which leads to suboptimal sub-queries generation.

On the other hand, the FA score of the proposed model significantly outperforms the baselines. This is expected, given that the GPT-4 model is substantially more powerful than GPT-3.5 in reading comprehension. Therefore, the generated answer can be considerably more consistent and faithful to the provided context. In addition, this result implies that GPT-4 tends to hallucinate far less than the LLM generators of the baseline models.

Likewise, the AS score of the proposed model is notably higher than the baselines. This outcome may be a result of a combination of the proposed model's more effective summarization of details and better identification of similar resumes. Once again, this may be because of the advanced comprehension and contextual understanding capabilities of the GPT-4 model, allowing it to generate more concise analyses and summaries of resumes.

To summary, even with a more advanced design, the retriever component of RAG Fusion models did not yield notable improvements in ranking and retrieving resumes. Moreover, in some cases GPT-3.5 is utilized for RAG Fusion's sub-queries generation, the outcome of document ranking is even slightly less effective than Naive RAG's generic retriever. On the other hand, the solid improvements in FA and AS scores of the proposed model can mostly be attributed to the GPT-4 LLM generator. Being a considerably more powerful LLM than GPT-3.5, this model can provide summaries and explanations of selected resumes that closely follow the details in the provided context.

To better assess the resume selection outcome for the system as a whole, it is necessary to investigate the rest of the metrics. Table below displays the average similarity and accuracy, which are obtained by comparing the selected resume to the ground truth resume.



| Models | Accuracy | Similarity |
|---|---|---|
| GPT-4 + RAG Fusion | **0.558** | **0.885** |
| GPT-3.5 + RAG Fusion | 0.426 | 0.851 |
| GPT-3.5 + Generic RAG | 0.436 | 0.852 |

Given that multiple candidates might match the ideal profile at the same time, the accuracy scores of the models can be considered satisfactory.

Among 500 job descriptions with up to 1000 resumes, the models can identify around 120 to 280 cases. Once again, it is visible that the selection accuracies of the two baseline models are relatively similar. Meanhwhile, the proposed model stands out with a significant improvement in accuracy, suggesting its higher effectiveness in accounting for more nuanced factors in the job requirements to identify the exact matching resume.

On the other hand, the selected resumes for all models exhibit a high degree of similarity to the ground truth. On average, the selected resumes by each model are 85% to 89% similar to the ground truth. This indicates that the RAG systems can effectively determine highly similar candidates among a large database of resumes. It is noteworthy that GPT-4 achieves the highest resume similarity score of up to around 89%. This suggests that the resumes selected by the proposed model are generally more suitable.