VINCAE

Performance Analysis of Inference Pipelines

Matheus Vaz Manzke

Introduction

In the following pages, I will be analyzing data from two machine learning LLM pipelines. These datasets pose several challenges, as some metrics cannot be meaningfully compared without additional information about their distribution. A naive analysis could lead to unsupported conclusions and potentially have negative repercussions in a business context.

I have set an ambitious goal: not only to analyze and critique the data at hand but also to propose a model for standardized, statistically sound comparisons between inference pipelines that could be used in production. This model will be successful if it can provide a meaningful and consistent measure of the distance between pipelines concerning the metrics of choice. This should allow for rigorous evaluation of metrics across multiple runs and different hyperparameter configurations. facilitating communication between the data team and non-technical stakeholders.
I expect that there might be errors and faulty reasoning, but I hope you will find the core idea to be solid.

TL;DR

- Since we are comparing multiple runs for each pipeline, we are in fact comparing distributions. In this context, simply comparing the average for each metric is not enough.
- We solve this problem by using a sequence of statistical tests.
- Kruskal-Willis is a non-parametric test for group difference that allow us to know if our four pipelines are really differing when it comes to the relevant metrics.
- Dunn's test is a post hoc pairwise difference test that allow us to identify which pairs inside a group are actually differing.
- Then, we can measure effect size by using a custom non-parametric measure.
- For the dataset in question, there are statistically significant differences for some measures of interest between the pipelines.
- **But we have to be somewhat generous with our p-values to find statistically significant differences with respect to the baseline - except for the total cost. It may make sense to be generous in this business context, though.**

- Our goal is to compare the performance of four different inference pipelines. Each pipeline was run seven times, resulting in a dataset with 28 rows of measured results for the same input. **All statistical results should be considered in light of this somewhat limited amount of data.**

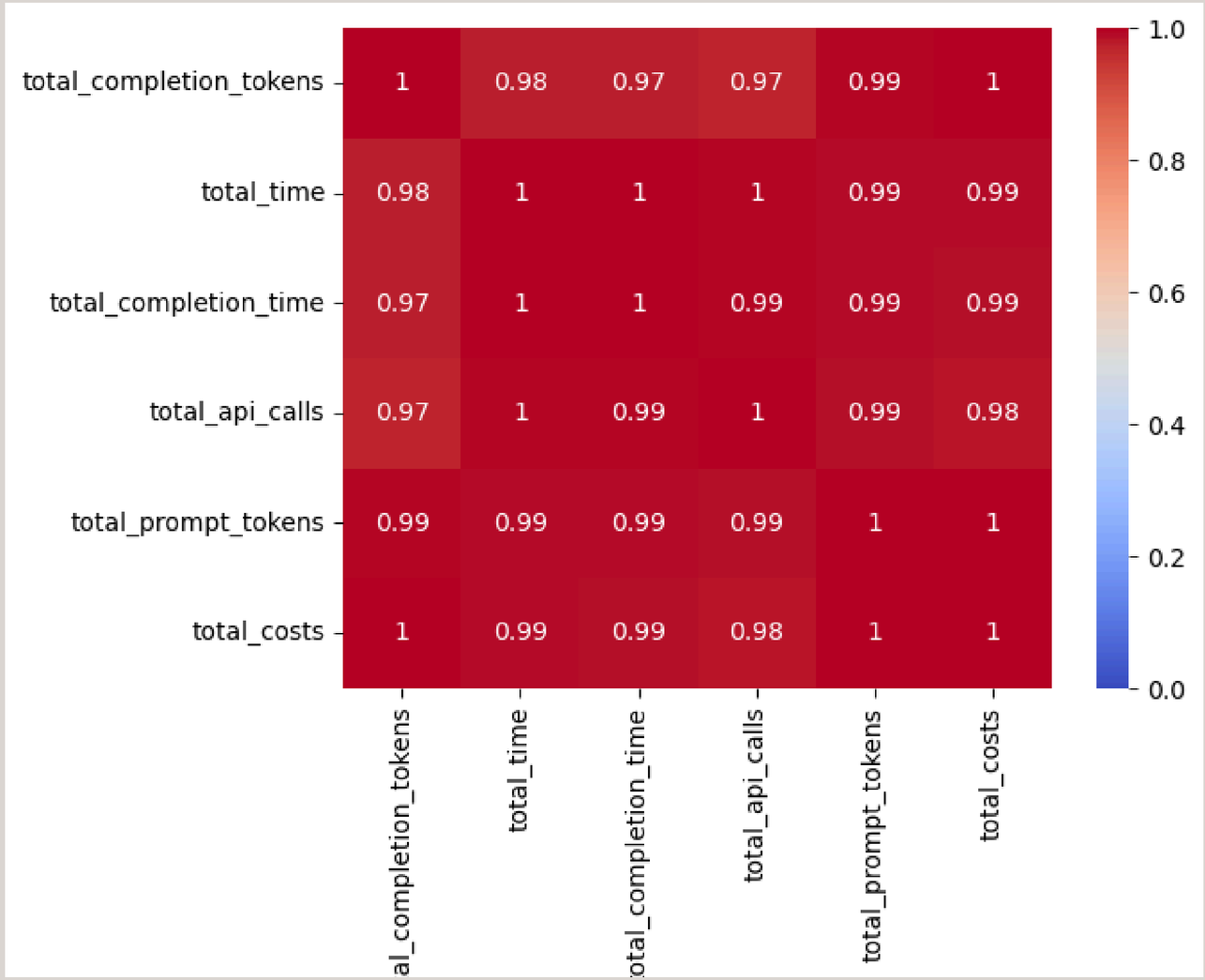
- **Recall seems to be the most critical metric for the business model** because failing to alert a client of a violation could significantly impact their evaluation of GetGenAI's product. Therefore, we will give special attention to this metric.

- For basic classification metrics, we have results for each run. However, for text-specific metrics like text overlap, we only have an average for each pipeline.

- We can't simply compare means without understanding the underlying distributions. Two variables with the same mean can behave very differently, leading to a product with unpredictable performance. Addressing this issue will be the core of my analysis.

- Given the uncertainty innate in machine learning, we want a new result to be considered 'better' only if we can be assured that it is at some statistically significant distance from our previous result.

- Our analysis will primarily focus on the metrics 'recall', 'precision', 'f1_score', and 'total_costs'. However, this model could be applied to all relevant metrics.
- Other columns in the 'statistical_inference.csv' file are excluded from the analysis because they are essentially collinear with 'total_costs'.
- That doesn’t mean that it is not relevant knowing the exact duration of a pipeline's completion or how many API calls were made, but passing effective judgment requires a deeper understanding of the business context. In this context, they are redundant.

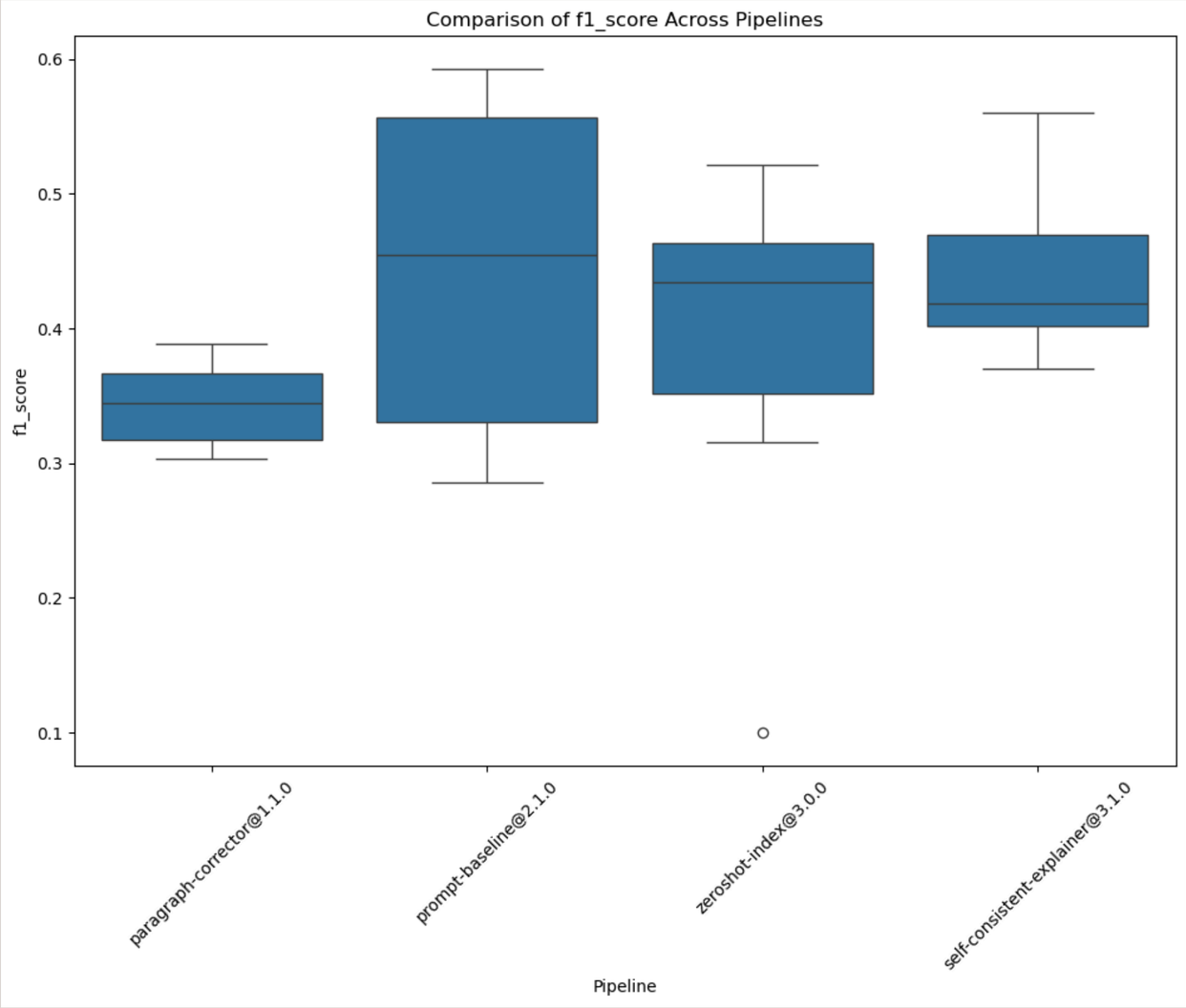


A correlation matrix for some of the columns in statistical_inference.csv

- After having all of the relevant runs in our hand, we might want to answer the following question: Are there actual, significant, differences in the performance across all pipelines being considered?

- If we take a look at the box plot for the F1 score across all four different pipelines in our dataset, we get a sense of how difficult this question can be.

- Baseline has the highest mean value, but it is also highly variable. Is it better than the more sophisticated 'self-consistent-explainer'?



Boxplot for the f1 score across all the pipelines

- We can start aswering the question of which model is performing better with a Kruskal-Wallis test. Contrary to ANOVA, Kruskal-Wallis is a non-parametric test that doesn't assume the data we are testing is normally distributed. We so few examples per pipeline, we can't really tell.
- **Our null hypothesis is that all pipelines perform the same. We test this hypothesis for each metric of interest.**
- In scientific practice, it's common to consider the null hypothesis disproven when $p < 0.05$. In our context, this threshold may be too rigid. Still, given that three of our measures fall below the 0.05 threshold, we could assert that our pipelines exhibit significant differences in performance.
- **We still don't know which pipelines are differing. Or by how much.**

Metric	Statistic	p-value
f1_score	5.377905	0.146126
recall	8.774624	0.032442
precision	9.263961	0.025980
total_costs	24.333333	0.000021

Kruskal-Wallis H-statistic and p-value for all four our metrics.We are trying to find significant difference among the four pipelines.

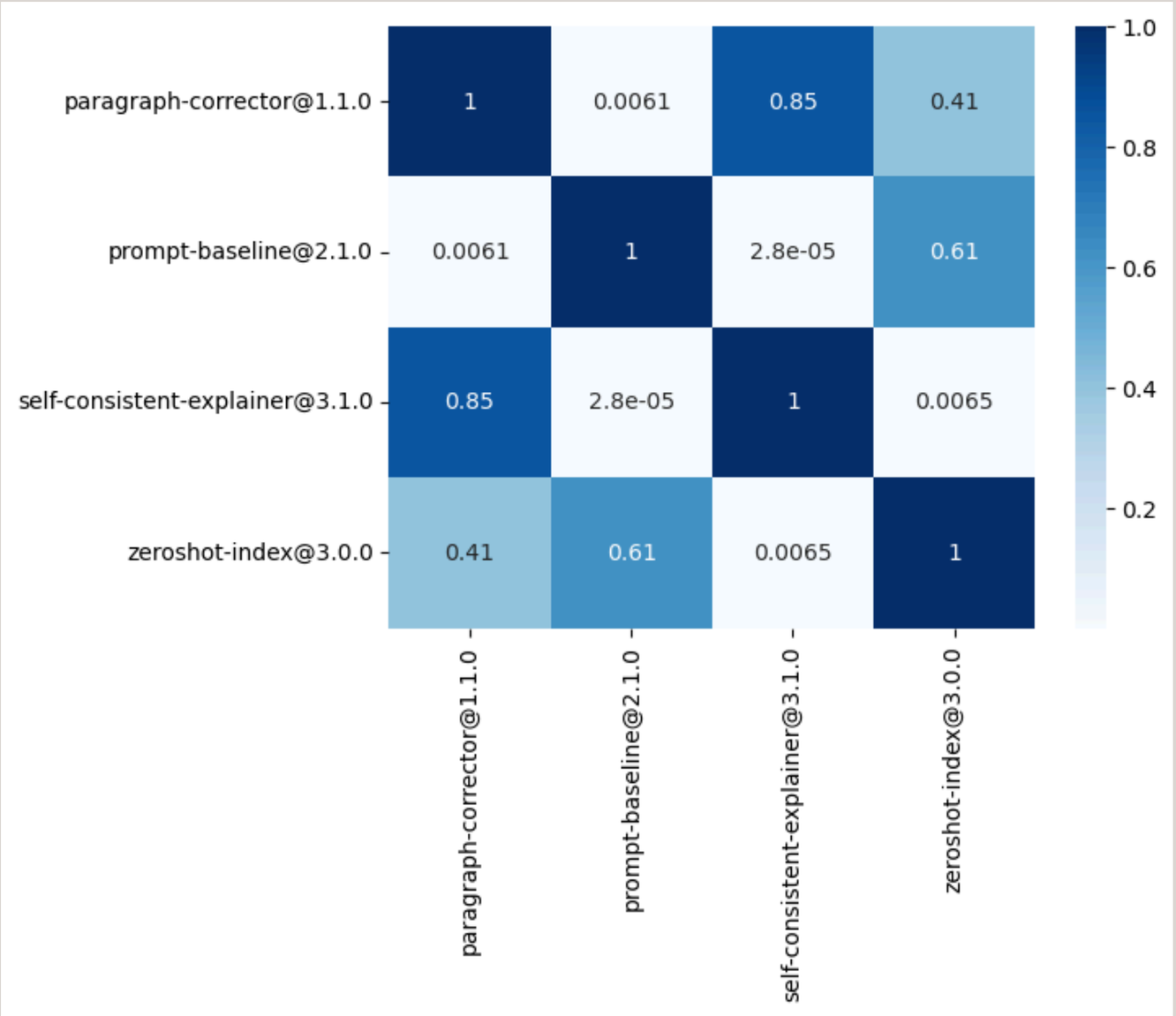
- To find out where out significant differences are, we apply a **post hoc Dunn's test**. This will do a pairwaise comparison between pipelines.

- We will end up with a matrix of p-values. Smaller values will tell us that the difference is significant.

- For example, when it comes to total_costs, 'prompt-baseline' and 'self-consistent-explainer' differ the most, with a p-value that is essentially 0.

- Extreme cases like this one are obvious to the eyes (check the notebook for the boxplot of total_costs). But we can't rely on our eyes all the time.

- We may have found our pairwise differences. **But we still don't know how big these differences are!**



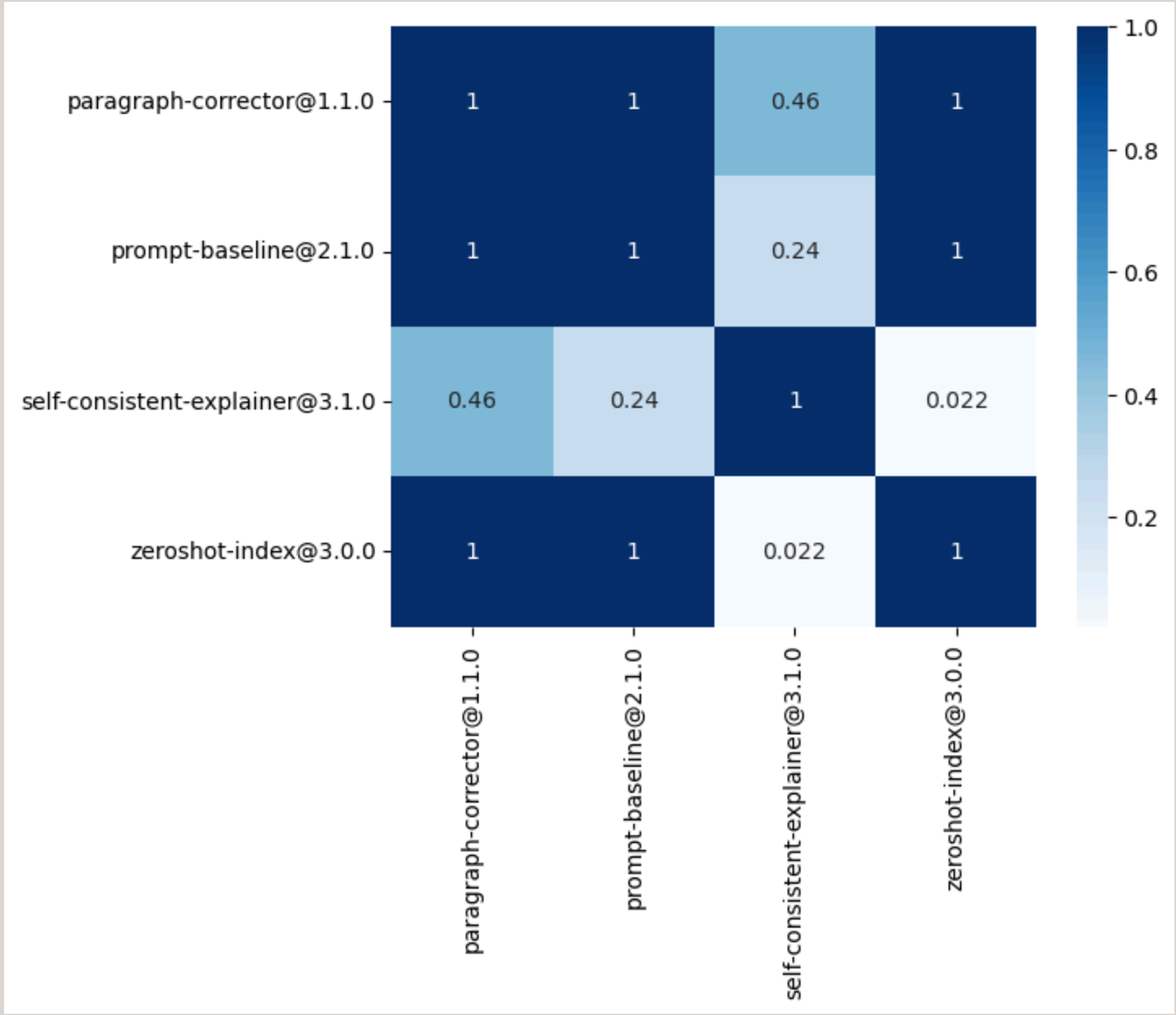
Dunn's test and pairwise difference between pipelines in respect to total_costs

- Trying to estimate the size of this difference (effect size, in jargon) is not trivial. The most common measure is Cohen's D, but it is not very adequate here, because our data is not normally distributed.
- We need to find a non-parametric alternative in order to avoid distorted values that could be misinterpreted later.
- The most common non-parametric methods to measure this difference won't help us. So I applied a custom measure created by another author - a non-parametric measure that is consistent with Cohen's D. Following the original author, we will call it **gamma effect size**, or **g**. (Source: https://aakinshin.net/posts/nonparametric-effect-size/)
- **W**ithout getting bogged down in the details, the gamma effect size function creates a 'pooled' measure of variance, taking into account the variance in both distributions and then measures the distance between the means *in terms of this pooled variance*. A G, then, can be interpreted as one 'pooled variance' of distance.
- If all is a bit too abstract, I believe the following pages will make everything clear.

- For example, take a look at the Dunn's test for 'recall'.

- If we are strict about what we define as significant ($p < 0.05$), the only significant difference between pipelines occur between 'self-consistent-explainer' and 'zeroshot-index'.

- Knowing that they differ, we can ask ourselves by how much.



Dunn's test and pairwise difference between pipelines in respect to total_costs

```
Baseline pipeline: zeroshot-index@3.0.0
Comparison pipeline: self-consistent-explainer@3.1.0
Total Cost Gamma: 9.679274015643973
Precision Gamma: -1.949166249132801
Recall Gamma: 2.023472278429786
F1 Score Gamma: -0.1676814342246375
```

Gamma effects accross different metrics when comparing 'self-consistent-explainer' and 'zeroshot-index' pipelines.

- When comparing recall, there is a difference of 2 Gs between the two groups.
- After Dunn's test, we determined that there is no significant difference between both pipelines concerning precision. **Consequently, the precision gamma should be ignored.**
- However, this is not the case for total cost; the pipelines exhibit significant differences, with a substantial effect size of 9 Gs separating them.

- A more interesting question is: **How much does the 'self-consistent-explainer' differs from the baseline pipeline, based on the the available runs?**
- If we are too strict ($p < 0.05$), after Dunn's test, the only significant difference is in cost.
- **We could allow for $p < 0.25$ (Which I believe could be justified in a business context where we are trying to find a direction to explore. We are not trying to prove once and for all that a result could not possibly come from a base distribution)**. If we adopt this threshold, Dunn's test will indicate a significant difference in recall.

Metric	self-consistent-explainer@3.1.0_gamme_value
Total Cost Gamma	8.927705
Precision Gamma	-1.313438
Recall Gamma	2.023472
F1 Score Gamma	-0.273350

Gamma effects accross different metrics when comparing 'self-consistent-explainer' to basseline.
Total Cost and Recall are the only gammas that should be taken into account.

- Let's do the same for the other pipelines: **How much does the 'zeroshot-index' differs from the baseline pipeline, based on the the available runs?**
- After Dunn's test, there's no significant difference (even if we allow for $p < 0.25$) between 'zeroshot-index' and the baseline.
- All the gamma values should be ignored.

Metric	'zeroshot-index@3.0.0'_gamma_value
Total Cost Gamma	2.088167
Precision Gamma	-0.059621
Recall Gamma	0.000000
F1 Score Gamma	-0.135727

Gamma effects accross different metrics when comparing ' 'zeroshot-index' to basseline. None of the gamma values should be considered.

- **How much does the 'paragraph-corrector' differs from the baseline pipeline, based on the the available runs?**
- After Dunn's test, there are significant differences ($p < 0.05$) in Precision and Total Cost.
- The significant Precision G value is negative! That means that performance has worsened, while costs have shot up.

Metric	paragraph-corrector@1.1.0_gamma_value
Total Cost Gamma	13.101380
Precision Gamma	-1.908833
Recall Gamma	0.000000
F1 Score Gamma	-0.881870

Gamma effects accross different metrics when comparing ' paragraph-corrector' to basseline. Only Precision and Total Cost should be taken into account.

Conclusion

As we have seen, despite the existence of significant differences among all four pipelines taken as a group, when compared to the baseline naive pipeline, it is much more challenging to identify a significant difference for the metrics under consideration here.

There are several reasons for this observation before we even consider the actual quality of the pipelines. The analysis being conducted here is quite limited; we are only using a small number of metrics from a very small sample. Additionally, the significance threshold may be too rigid. Extra care should be taken in how we approach tests and interpret measures like the gamma effect used here.

Once we establish an objective, statistically sound measure of performance, we can confidently assess model performance. Based on those game values and a baseline reference, we could develop a composite measure of the global performance of a model, so that our progress could be more easily understood by stakeholders who may not be comfortable with box plots and statistical terminology.