Evaluation Results Summary

METEOR Scores

- LLM: 0.765 ± 0.211
- Google: 0.304 ± 0.226

BLEU Scores

- LLM: 0.540 ± 0.355
- Google: 0.106 ± 0.103

The results clearly demonstrate that the LLM system significantly outperforms the Google system across both evaluation metrics. The LLM achieved METEOR scores more than twice as high as Google's system (0.765 vs 0.304) and BLEU scores approximately five times higher (0.540 vs 0.106).

Notable observations:

- The standard deviations indicate some variability across test cases for both systems
- The LLM performance is consistently superior in both precision-focused (BLEU) and recall-oriented (METEOR) metrics
- This suggests the LLM delivers better overall translation quality with significantly improved semantic accuracy
- There are other metrics that could be used to gauge similarity, especially fully semantic ones like BERTScore, COMET, BLEURT, but I didn’t use them because of time constraints

Insights

The LLM is definitely better than google translate. There is just one area where it struggles, but even there it is far better than google translate. That area is parametrizing. It sometimes believes there parameters in places where they are not and this was consistent even after multiple iterations of prompt engineering.