Fine-Tuning Llama 2: Evaluating Text Generation with Cosine Similarity and Human Assessment

Alina Lozovskaya
Innova
alinailozovskaya@gmail.com

Abstract

In the rapidly evolving field of Natural Language Processing (NLP), fine-tuning large language models (LLMs) has become pivotal for addressing specific tasks without complex feature engineering. However, the absence of standardized evaluation metrics for LLMs presents a challenge. The purpose of this small to explore the viability of using a cosine similarity metric, computed with context-aware embeddings, as a pragmatic alternative to intricate evaluation methods, validated against human assessments. Besides, we also use the GPT-4 Turbo API to additionally evaluate its text generation evaluation capabilities.

1 Introduction

In a dynamic NLP environment, fine-tuning open-source LLMs adapts models to specific tasks without the need for complex feature engineering. Despite practical benefits, the absence of standardized evaluation metrics poses a challenge. This study assesses the efficacy of a cosine similarity metric, computed with context-aware embeddings, as a pragmatic alternative to intricate evaluation methods, validated against human assessments.

2 Methodology

For the analysis, we use the open Llama-2-7b-hf model, fine-tuned on the USA News Dataset using the combined "title" and "abstract" columns. We evaluate the resulting generation using cosine similarity with SentenceTransformers "all-mpnet-base-v2" pre-trained embedding model, GPT-4 Turbo, and human evaluation as well.

3 Human Evaluation

Human assessment, employing a five-point scale, acts as ground truth for evaluating computed metrics. Assessors grade 100 lines, assessing coherence and relevance, providing nuanced insights into text generation quality.

4 Analysis and Results

Cosine similarity scores reveal a range of results, setting the stage for human evaluation. The distribution between the reference and the generated text using SentenceTransformer embeddings shows that the computed score tends to a normal distribution. Ongoing correlation calculations and extended practical findings will be presented in the form of a poster.

5 Conclusion

In this preliminary abstract, we introduce our study in text generation, employing a cosine similarity metric and human assessments mainly. While the comprehensive analysis is ongoing, initial observations suggest the potential of accessible metrics for evaluating LLMs. The integration of GPT-4 Turbo and ongoing correlation calculations will provide deeper insights. Since we investigate the interaction of human evaluations and computational metrics, this research provides a basis for designing methodologies for language model evaluation, contributing to the evolving discourse in the field of natural language processing.