Relatório Instruction Fine-Tuning e Avaliação via LLM-as-a-Judge – Jhonnathas Swerley

1. Introdução

Relatório que mostra o processo de adaptação de um modelo de linguagem base (GPT-2 Medio) para um modelo assistente capaz de seguir comandos e responder a perguntas. Modelos baseados na arquitetura original do GPT-2 são treinados estritamente para prever a próxima palavra em uma sequência, carecendo da habilidade de dialogar ou acatar instruções diretas. Para transpor essa limitação, aplicou-se a técnica de *Instruction Fine-Tuning* (Ajuste Fino por Instruções), seguida por uma avaliação automatizada e escalável utilizando a metodologia *LLM-as-a-Judge*.

2. Metodologia de Construção do Dataset

Para o processo de *Fine-Tuning*, o modelo necessitava de exemplos claros de como se comportar como um assistente.

- **Seleção dos Dados:** Inicialmente, o projeto utilizaria um conjunto reduzido de 1.100 instruções. Contudo, para obter um ganho semântico e estrutural mais robusto, o dataset foi substituído pelo **Stanford Alpaca**, composto por aproximadamente 52.000 exemplos.
- **Estrutura e Geração:** Este dataset foi construído através da técnica de *Knowledge Distillation* (Destilação de Conhecimento), onde um modelo de fronteira (estado da arte da OpenAI) gerou pares sintéticos de instrução e resposta. Os dados foram estruturados em formato JSON, compostos pelos eixos **`instruction`** (o comando), **`input`** (contexto opcional) e **`output`** (a resposta ideal).
- **Prevenção de Overfitting:** O conjunto de dados foi dividido em dados de treinamento (utilizados para a atualização dos pesos do modelo) e dados de teste (`test_data`). A avaliação foi conduzida estritamente sobre os dados de teste, garantindo que o modelo fosse avaliado apenas em instruções inéditas, atestando sua capacidade real de generalização.

3. Resultados Quantitativos e Qualitativos

3.1 Avaliação Quantitativa (LLM-as-a-Judge)

Para avaliar o desempenho do GPT-2 de forma escalável, implementou-se um pipeline de *LLM-as-a-Judge* baseado em referência. Utilizamos o modelo **Llama 3** atuando como juiz imparcial.

O script de avaliação iterou sobre o `test_data`, submetendo ao Llama 3 a instrução original, a resposta de referência (*Ground Truth*) e a resposta gerada pelo GPT-2 (`model_response`). O juiz foi instruído via *prompt* a emitir uma nota de 0 a 100. Para mitigar erros de formatação na saída do Llama 3, utilizou-se Expressões Regulares (Regex) para extrair o valor numérico com precisão..

4. Limitações do GPT-2 e do Juiz (Llama 3)

A implementação da arquitetura revelou limitações inerentes tanto ao modelo avaliado quanto ao modelo avaliador.

Limitações do GPT-2: Sendo um modelo de arquitetura mais antiga e com um número restrito de parâmetros, o GPT-2 apresenta um "teto de conhecimento". Embora o dataset Alpaca o ensine a se comportar como um assistente, o modelo possui uma janela de contexto curta e uma tendência notável à repetição de tokens ou à geração de alucinações quando confrontado com tarefas que exigem raciocínio lógico em múltiplas etapas.

Limitações do Llama 3 como Juiz: O uso de um LLM para avaliação introduz desafios de verbosidade. Apesar das instruções estritas para retornar apenas números inteiros, o Llama 3 frequentemente gerava textos periféricos (ex: "A nota é 85"). Isso exigiu a implementação de tratamento de erros no código (Regex e blocos `try/except`) para evitar a perda de dados avaliativos. Além disso, juízes algorítmicos podem apresentar um viés de favorecer respostas mais longas, independentemente da precisão absoluta.

5. Reflexão: O ganho veio mais do modelo ou dos dados?

Neste experimento, fica evidente que o salto em usabilidade e utilidade adveio majoritariamente dos **dados**.

O modelo GPT-2 original possuía o conhecimento sintático e estatístico do idioma, mas não a "postura" de um assistente de IA. A transformação de um mero preditor de texto para um sistema capaz de responder a um prompt foi inteiramente moldada pelo dataset Alpaca. Os dados funcionaram como as "regras do jogo".

Contudo, a qualidade bruta do conteúdo gerado (a ausência de alucinações e a profundidade factual) continua sendo limitada pela arquitetura do **modelo**. Em síntese: os dados ensinaram o GPT-2 *como* interagir, mas a capacidade computacional do modelo ditou o *limite* de sua inteligência nessa interação. O experimento reforça o paradigma da *Data-centric AI*: dados de alta qualidade são capazes de extrair o máximo potencial até mesmo de arquiteturas mais antigas.