Overview

The project involves annotating and evaluating distractors for LLM-generated datasets. The goal is to create high-quality distractors that challenge LLMs while maintaining semantic relevance to the ongoing conversation. Supervisors will provide domain-specific instructions via email.

Key Definitions

- **Distractor:** An off-topic term or phrase that confuses an LLM and breaks part of the system instruction, while still being close enough to the conversation domain to appear natural.

- **Dataset Format:**

    - **Conversation + Distractors = Conversation with Distractors**

    - Distractors will be inserted into conversations programmatically if needed (e.g., using a Python script).

Guidelines for Creating Distractors

1. **Quality**

    - Distractors must be high quality, semantically relevant, and domain-appropriate.

    - They should break the system instruction clearly, not vaguely.

    - Distractors should be slightly out-of-topic, not completely irrelevant.

2. **Characteristics of Good Distractors**

    - Related/familiar to the conversation.

    - Appears natural in context.

    - Intentionally confuse the LLM without being obviously irrelevant.

3. **Characteristics of Bad Distractors**

    - Overly simple.

    - Too similar semantically (minimal difference).
    - Generated in a formulaic or low-effort way.

Annotation Process

- **Duration:** ~2 weeks focused on annotating high-quality examples.
- **Team Division:** Split tasks between annotation and evaluation across team members.
- **Time Tracking:** Measure how long it takes to create annotations.
- **Target:** 5 distractors per conversation.

Evaluation Process

1. **Initial Evaluation**
    - Assess whether distractors truly break the system instruction.
    - Ensure group agreement on distractor quality.
2. **Methods**
    - Use topic control models or semantic similarity measures.
    - Compare LLM-generated distractors with human-generated ones.
3. **Findings**
    - Performance on **human-generated distractors** is lower when evaluated by LLMs compared to LLM-generated ones.

Dataset Setup

- **Annotation Format:**
    - Add a dedicated column identifying which part of the system instruction is broken by each distractor.

    - Reduces the need to analyze the entire instruction repeatedly.

- **Processing:**

    - Plan the dataset structure and develop scripts to automate insertion, formatting, and processing.

Next Steps

1. Annotate high-quality distractors (focus for 2 weeks).

2. Begin evaluation phase: test distractors on the LLMs mentioned in the paper and among each other.

3. Finalize dataset format, including columns for:

    - Conversation

    - Distractor

    - Broken system instruction component

4. Continue refining annotation guidelines and ensure consistent quality across team members.

5. Evaluate final distractors and present results