Data Scientist for GetGenAI - a venture-backed startup pioneering the automation of content review processes for highly regulated industries [remote]

Test assignment:
This assignment is designed to assess your analytical skills in a focused area of data science, specifically in the context of NLP and text analysis using AI-based inference pipelines. You are encouraged to choose the task that best aligns with your strengths and interests.

Objective:
Checker inference pipeline is given two inputs (a set of rules and context to be analyzed), and as output it returns violations found in a given context, supplemented with some additional details (violation categories, explanations and/or suggested corrections etc.). Your goal is to analyze exemplary data collected during evaluation of some of our checkers.

Task:
Choose one of the following tasks to complete:

Performance Analysis of Inference Pipelines (Option A):
Assess the accuracy of inference pipelines. Examine how effectively they identify and categorize violations in the text given the rules. Use the data from the 'Human Violation Metrics' files for this analysis. Optionally, you can evaluate checkers from the point of view of execution time, costs and other statistics provided in 'Inference Statistics' files.

Quality Analysis of Rules (Option B):
Analyze the rules used across the inference pipelines from the 'Human Rule Metrics' files.
Identify any rules that appear to be problematic due to high false positive rates, inconsistencies in application, or other factors. Suggest possible improvements or alternatives for these rules based on your analysis.

Data format:
Each directory is associated with a single experiment run (evaluation) and contains three files:

- Human Violation Metrics (CSV file): This file includes data collected during automatic evaluation of the violations detected by the inference pipelines given ground-truth. It contains details on the types of elementary metrics, found matches, false positives and false negatives.
- Human Rule Metrics (CSV file): This file contains metrics aggregated by rules across all the evaluated samples. It provides a different view on checker performance, enabling one

to abstract from implementation of this or that checker, and investigate and draw general conclusions on the impact of a particular rule on the detected violations,

- Inference Statistics (CSV file): This file provides additional metrics of the inference pipelines themselves. It includes data on execution time, number of failed API calls, number of input and output tokens and total execution costs given cost model.

Additionally, experiment results are supplemented with a presentation explaining our performance metrics.

Deliverables:

A detailed report including:

- Your chosen task and the methodology for the analysis.
- Findings from your analysis, including data visualizations and statistical comparisons (if applicable).
- Any scripts or code used for the analysis.
- We are interested to see your reasoning and critical thinking, don't hesitate to question and/or criticize the quality of the provided data.

Submission Guidelines:

Compile your findings and analysis in a concise yet comprehensive report in any form that you like (PDF, Microsoft or Google Doc file, etc.). Ensure all data visualizations are clear and appropriately labeled. Submit your report and any accompanying files.