**Data Analytics Capstone Topic Approval Form**

The purpose of this document is to help you clearly explain your capstone topic, project scope, and timeline. Identify each of the following areas so you will have a complete and realistic overview of your project. Your course instructor cannot approve your project topic without this information*.*

**Student Name:** Peter Krupa

**Student ID: 011629231**

**Capstone Project Name:** How stochastic is “stochastic”?

**Project Topic:** Measuring the reliability of large language models in mission-critical workflows

**Research Question:** Large language models (LLMs) are often referred to as “stochastic computing.” But how stochastic are they? That is, for a particular task, what is the variability of their responses to a particular query?

**Hypothesis:** In close-ended, repetitive tasks, the stochasticity of large language models can be measured and then mitigated, making them suitable for supervised use in data transformation tasks.

**Context:**A company wants to explore using large language models (LLMs) for data transformations in its data pipelines. This entails some risk, because generative artificial intelligence technology is "stochastic computing": that is, there is always a certain degree of randomness in its output. To mitigate this risk, the company could do a data analysis project to quantify it. The project would seek to measure the stochasticity of LLM output and identify the factors that increase or decrease it. The company would then be able to make an informed decision about whether to use LLMs for data transformation at the different points of its data pipelines based on its appetite for risk.

**Data:** Because this project will be working with LLM output, collecting the data will involve making calls to APIs and cataloging the responses.

I will be paying to query APIs offered to the public by companies like OpenAI and Anthropic that allow the data resulting from inference to be used for commercial and research purposes.

**Data Gathering:**  For each test case, I will run a Python script that sends a test query to an LLM asking it to use a sample text to populate a JSON schema. I will run the query with the same prompt a statistically-significant number of times to gather the necessary amount of data, saving each output to a table in a PostgreSQL database. A second script will then run validation tests of the LLM responses collected and store the results in the same database.

**Data Analytics Tools and Techniques:** I will use basic statistical measures of probability to analyze the output of each test run. Also, once I have one or more base cases for comparison, I will do further runs that adjust factors like prompt length, temperature, top k, and prompt wording to conduct A/B tests in order to gauge how best to narrow the randomness while maintaining successful parsing.

**Justification of Tools/Techniques:** With basic probability measures, I can get the answer to my data analysis question in the form of a percent. If *p*=0.95 that a given prompt will result in an output that passes validation, I can say that model stochasticity will cause it to fail 5% of the time. Likewise, the A/B testing will make it possible to measure the degree to which changes to the querying technique affect *p*, and therefore affect stochasticity. And once my process is built out, I can run these tests for a range of different models and present the results in a visualization.

**Application Type, if applicable (select one):**

mobile

web

stand-alone

**Programming/Development Language(s), if applicable:** Python

**Operating System(s)/Platform(s), if applicable:** Linux/Unix

**Database Management System, if applicable:** PostgreSQL

**Project Outcomes:** 1) Baseline stochasticity measure for at least two large language models; 2) Stochasticity impact of querying adjustment techniques for at least three techniques (result of A/B testing); 3) Presentation with visualizations and summary of findings; 4) Recommendation: use LLMs for data transformation, or not?

**Projected Project End Date:** 8/31/2025

**Sources:** Click here to enter text.

**Human Subjects or Proprietary Information**

Does your project involve the potential use of human subjects? (Y/N): N

Does your project involve the potential use of proprietary company information? (Y/N): N

**STUDENT SIGNATURE**

**­­­­­­­­­­­­­­­­­­­­­­­­­­­­­\_\_PETER KRUPA\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**By signing and submitting this form, you acknowledge** that any cost associated with the development and execution of your data analytics solution will be your (the student) responsibility.

**TO BE COMPLETED BY AN INSTRUCTOR**

**The capstone topic is approved by an instructor.**

**INSTRUCTOR’S NAME AND SIGNATURE:**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**INSTRUCTOR APPROVAL DATE:**

**Project Compliance with IRB (Y/N):**