CAPSTONE PROJECT PROPOSAL | Data Science Institute at Columbia University

A project proposal submission consists of a project description and a dataset. It can be about a novel dataset or a publicly available dataset with a novel task.

Email address:
lilliann.rowan@fmr.com; indraneel.biswas@fmr.com; michael.threlfall@fmr.com

Please read BEFORE you propose a project

Unless otherwise noted (e.g., public data sets), students are instructed not to share the capstone dataset outside the team or use the data beyond the scope of the capstone project. Please propose a project with the expectation that its results will be shared publicly. Capstone projects are an important educational experience for our students and also provides them with a showcase to share in interviews for potential employers or for various other audiences inside and outside Columbia.

Columbia views capstone course work as work owned by the students and shared with the mentors. As a general practice to simplify the capstone project program, we do not sign NDAs for the capstone projects.

1. DSI INDUSTRY AFFILIATE
   Fidelity Investments

Project Description:

Create a LLM evaluation framework to quantify the following from the context of the financial industry. The group may create their own evaluation framework or choose one or more of the following.

1. Correctness: Measure of factual correctness of the generated responses, without any interpolation or extrapolation.
   a. Generate a set of questions and answers from a given context. Given the context, the generated responses will be compared with the ideal answers.
   b. Additionally, a set of questions to test the degree of hallucination, can also be created.
   c. Create a metric to appropriately capture the degree of factual correctness and/or hallucination.
2. Sensitivity: Variation in result due to minor variations in prompt
   a. Create a set of diverse prompts for testing responses. Ideally there should be two types, (1) Prompt pairs with minor change in language, but ideal response will be significantly different; (2) Prompt pairs with major change in language, but without expected change in ideal responses.

     b. Create a metric to jointly capture (1) difference between ideal and generated responses, and (2) change is responses within prompt pairs.

3. Reasoning: Evaluation of chain-of-thought against ground-truth
     a. Create a set of prompts, with ideal responses, which require chain of thought reasoning. Generated responses for these prompts should be compared with ideal responses.
     b. Create a metric that will quantify the accuracy of the ideal chain of thought with the generated.
     c. The prompts should be related to the finance and regulatory jargon.
     d. Capturing the length of chain-of-thought should be included in the framework.

2. MOTIVATION, BACKGROUND AND OVERVIEW

Please state briefly what is the problem that the project tackles. The projects need to be focused on a data science problem that is engaging, relevant, clearly defined and of the right scope for a semester. When assessing the proposals we will be looking for a diverse set of problems that address different topics and technical requirements that our students can address. The evaluation criteria will include: Is this a data-science project? Can our students learn about a data science application in the real world? Is the proposed research problem important and can potentially have a big impact? Will our students be excited about it? Please provide your project description having these criteria in mind.

Generative AI models have started creating significant impact in multiple industries and are providing innovative solutions. Solutions for tasks like summarization, question - answers, content generation, translations are creating new opportunities for industry. With these new opportunities, new challenges are emerging in getting these models adopted due to lack of standard, comprehensive benchmarks suitable for generative tasks. This project attempts to create a framework and specific benchmarks for generative AI tasks relevant for financial service industry.

The framework attempts to define key facets for generative AI model output like truthfulness, repeatability, referenceable, creativity/hallucination etc.

DATASETS

The dataset(s) can be public or private. Please keep in mind that the students will need to list the project on their CV and the report will be public. All datasets must be submitted by Friday, July 19, 2019 for Fall 2019.

3. DATASET

Please provide a detailed description of the type of data that is required to address the problem. For example, is this social media data, medical data, financial data, etc? What is

the size of the data. Will the organization provide the majority of the data or is the data accessible via other avenues/ sources? How much of the data is available? Do the students need to gather data? In assessing the projects, the availability and type of data will play an important role. Please consider these evaluation criteria for data requirements when submitting the proposal: Is the data set clearly defined? Is the data set complex and big enough for creating learning opportunities? Is the data set ready? (availability, need for processing) Does the data require extensive computing resources (if yes, can the affiliates provide resource/funding?)

The students may use any publicly available financial documents. Below is a list of possible public data sources for documents on financial topics. The students may choose any of these sources or use one that they have easy access to.

Google News RSS Feed: https://rss.app/rss-feed/google-news-rss-feed

SEC – New Rules: https://www.sec.gov/rules/final.htm

The links provided on this page will redirect the students to PDF documents with varying number of pages. These PDF documents can be downloaded either manually or by using any programming language.

SEC – Litigations: https://www.sec.gov/litigation/litreleases.htm

The links provided on this page will send the student to a webpage that will require web scraping to extract content.

The capstone group may then use any LLM to generate content, create summaries, or produce questions-answers. The LLM generated content will then need to be assessed using the evaluation framework formulated by the capstone group.

The amount of data available on both these websites is more than enough for the capstone project and requires little to no computing resources to extract.

4. Data Type
Public data is data made available by a third party and is available to the general public. Novel data is data that has been recently published by the proposer or will be made public as part of this project. Private data is data that can not be made available after the project ended. Please check all that apply.
-Uses Public Data

5. How will the dataset be made available?

For example: CSV/XLS file, remote database, raw images or documents, REST endpoint, etc.

Google News RSS Feed

There are multiple ways to gain access to the Google News RSS feed. Please see links below for some possibilities.

https://pypi.org/project/gnews/

https://github.com/kotartemiy/pygooglenews

https://www.aakashweb.com/articles/google-news-rss-feed-url/

Given the URL to the article, the students will need to perform web scraping to extract the text.

SEC – New Rules

The student group will need to either manually or programmatically download the PDF documents. After which, conversion from PDF to text will need to be done.

SEC – Litigations

The student group will need to perform web scraping to extract the text.

GOALS, OUTCOME and SKILLS

5. RESEARCH GOALS
What is the goal of this project? What questions do you want answered? What has been done already to achieve this goal?
What are the ethical considerations?
	Are there any ethical concerns about the proposed project such as privacy, transparency, and bias that we should pay special attention to?
What is the relevant background needed for the project?
	In order to make sure we build the right team of students for each project, please provide information on the relevant background information that someone working on the project should have. What technical skills they should have and/or relevant literature (please provide citations) or tools (please provide links) they will need to know or able to learn.

The goal of this project is to create an LLM produced content evaluation methodology and an automated way of measuring.

Relevant literature:
https://arxiv.org/abs/2305.12295#:~:text=Large%20Language%20Models%20(LLMs)%20have,to%20improve%20logical%20problem%2Dsolving.

OUTCOME
Model/Report/Paper/Software/Other

1. Paper on the LLM Evaluation Framework: Needs to describe the LLM generated content evaluation metric(s), the motivation and or justification for the metric(s), how the metric is measured, and the model or algorithm used for measuring.
2. Model(s) and/or process(es) to measure the quality of the LLM output.

SKILLS

What skills should students expect to learn through their project? Check all that apply.

Project planning and scoping
Data acquisition and scraping
Data versioning and management
Data cleaning
Combining data sources
Exploratory data analysis and visualization
Supervised modeling
Unsupervised modeling
Establishing evaluation metrics
Working with text data
Working with image data
Working with time series data
Working with tabular data
Working with geospatial data

In addition, the student will get exposure to large language models.

What are the quantitative and/or qualitative metrics that can be used to judge the successful completion of the capstone project?

To demonstrate success, the capstone project group will need to:

1. Define one or more evaluation metrics that can be used to measure LLM generated text quality (i.e., Degree of hallucination, reasoning)
2. Demonstrate the use of this evaluation metric on LLM generated text.

Are international students on a F1 or J1 student visa eligible to work on this project? Yes

PROJECT MENTORS

An important aspect of the capstone project is the opportunity for students to work with professionals across different industries or academic research labs. Thus each organization must provide mentorship to the students so that they can receive constant feedback and guidance (while each team will also have a faculty advisor the organization mentor will play a crucial role in guiding the team). Please specify who will work with the students and what are their qualifications or training? What amount of time per week do they intend to devote to working

with the project team? Each mentor will also ideally help the DSI faculty advisor assess the success of the project at the end which will translate to a grade for the students.

Each capstone project will be mentored by at least one industry mentor and one faculty mentor, with the industry project proposer(s)/mentor(s) as the primary mentor(s). Industry mentor(s), in addition to monitoring the project progress and provide timely guidance to the capstone team, are expected to

1. Meet with the team on a bi-weekly basis (teleconference is fine)

2. Review the midterm progress report and provide comments to both the team and the course instructor

3. Attend the final poster presentation session

4. Review and evaluate the team's final report

5. Provide comments on each team member's participation

Mentor #1

Name: Lilli Ann Rowan

> Title: Lead
> Dept./Division: Legal Risk & Compliance AI Center of Excellence
> Email address: lilliann.rowan@fmr.com
> Phone Number: +1 201 915 8099 / +1 201 920 8470
> LinkedIn Profile Link: https://www.linkedin.com/in/indraneelb/
> Current Resume (optional)
> What amount of time per week do you intend to devote to working with the project team? 1 hour each week.

Mentor #2

Name: Indraneel Biswas

> Title: Data Scientist
> Dept./Division: LRC DS
> Email: indraneel.biswas@fmr.com
> Phone: +91 99100 73612
> LinkedIn: https://www.linkedin.com/in/indraneelb/

16. Are you willing to work with two teams of students?

> No. If needed, the scope of the project may be reduced so that one team can manage it.

17. If yes, please indicate how the project may be appropriate to engage two teams.