**Title: Leveraging Language Models to Curate Skin Care Research Findings**

**Background:**

P&G is committed to enhancing the well-being of people's skin through continuous research and development (R&D) efforts. We strive to develop skincare treatments with improved performance and novel functionality. As part of our endeavor, we are interested in exploring the vast landscape of skin care research conducted by other researchers. However, comprehensively reviewing the extensive literature and clinical research available can be laborious and costly for humans. Considering the significant advancements made by Language Models (LMs) in natural language processing, we aim to harness the power of LMs to explore and understand this clinical research more efficiently.

**The Challenge:**

**Structuring Meta Data on Skin Care Clinical Trials to Tell a Story and Answer Key Business Questions:**

Since LLMs are difficult to train and evaluate, we have a different challenge for the students. We want participants to demonstrate their skills as data scientists to leverage LLMs in telling compelling stories. They will be tasked with analyzing the collected metadata and presenting insights about the skin care research landscape. Participants should highlight what has been done in the field, summarize the key findings, and identify actionable insights that can drive further advancements in skincare treatments.

**Skills Gained for the Students:**

* How to write effective prompts for LLMs to answer questions of interest
* Learn how to structure LLMs efficiently when there are token and or hardware memory limits
* Learn how to fuse LLMs with conventional data-science methods in python
* Learn how to become proficient data scientists telling substantiated stories and make actionable claims

**Format of the Hackathon:**

The hackathon challenge will be organized into three stages (or checkpoints). Two earlier stages early on Friday afternoon, And one final stage (biggest portion of the project) throughout Friday evening and Saturday morning. While the earlier stages will be used more for you all to become acquainted with the framework. The latter stages will be worth more and will be more open-ended. Before each stage you will be given a prompt (a word document) on the challenge.

After each checkpoint (or stages) we will self-critique to improve our approaches and insights. After each stage the questions will be more challenging and more open-ended.

We will try to coach you all in assessing the correctness of your claims, substantiating your claims with evidence, and improving your overall approaches.

**Data: We have provided a set of data scraped from** <https://clinicaltrials.gov/>

Your work will be based on a subset of this data scraped on a condition of interest though you will be allowed to explore data outside this set if needed. We have provided some preliminary methods for reading the data.

**Methods:** We have set up some environments with UC to leverage python and LLMs to help perform this analysis. More Details on this will follow. We encourage you to use your personal (local resources) or the resources provided. Since this project is mainly for your edification, we ask that you not leverage LLMs hosted outside of this challenge (leveraging a subscription to gpt4 turbo for example will not be permitted. All other tools at your disposal are permitted.