Anote

A CASE STUDY

Preprocessing Structured Data with Gen AI to Enhance Longevity Research

CUSTOMER Harvard Medical School
TASK TYPE Programmatic Labeling

BACKGROUND

Harvard University is at the forefront of exploring longevity, with an aim to extend human lifespan. Aging remains the leading risk factor for chronic diseases and mortality. To understand it comprehensively, there's a need to measure biological age accurately. Several machine learning algorithms, termed aging clocks, have been developed that can predict the age of biological samples based on omics data. However, a systematic resource for profiling biological age is still lacking. Enter, ClockBase.

CLOCKBASE

ClockBase, a comprehensive platform that integrates multiple aging clock models, curates the 11 top-performing aging clock models and applies them to over 2,000 publicly available DNA methylation datasets from resources like the Gene Expression Omnibus (GEO). The platform offers an interactive analysis tool for statistical analyses and visualization of biological age data. By leveraging ClockBase, researchers can explore biological age in different samples, discover new longevity interventions, and identify age-accelerating conditions.

THE PROBLEM

During the course of the research, a dataset was encountered which needed conversion from one structured format to another. Manually carrying out this conversion was not feasible due to the large volume of data and the intensive labor and time required.

The original dataset encompassed **2,424 unique identifiers,** each containing the following types of files:

Desc file: This file included a 'description' and a 'summary' of the unique identifier.

Meta file: This was a CSV file containing specific medical data under different columns.

Target file: This file specified the desired column structure for the output CSV file.

GPT-3 INTEGRATION

We integrated GPT-3, a powerful language model, to generate predictions for specific columns in the dataset. By prompting GPT-3 with relevant information, we obtained accurate predictions for columns such as perturbation category and tissue. These GPT-3 integrated functions played a crucial role in automating the data conversion process and ensuring accurate predictions for specific columns.

GPT-4 INTEGRATION

We leveraged the capabilities of GPT-4, a powerful language model, to generate predictions based on the training data. The integration with GPT-4 enabled us to automate the data type conversion process. Through this automated method, the conversion of the large dataset became feasible and efficient, driving the longevity research further and enabling the scientists to focus on core research tasks.

RESULTS

Through the combined use of programmatic labels, GPT-3 integration, and GPT-4 integration, we were able to automate the conversion of the large dataset. This automation not only saved valuable time and cost but also ensured high accuracy in the generated predictions.

The successful outcome of the solution allowed the team to leverage the algorithm for training a model via clock-based to explore the possibility of extending human life. This research has the potential to bring significant benefits to society.