Renaissance Philanthropy

Technical Overview: CareerNet

Summary

Renaissance Philanthropy is leading a project to develop three state-of-the-art benchmark datasets, leveraging CareerVillage.org, a platform that has crowdsourced career advice. Selecting data from over 60,000 questions and more than 3.5 million learners, the project aims to enhance AI's ability to guide users in navigating careers and accessing social benefits.

Renaissance will partner with The Learning Agency to implement this AI benchmarking effort and the annotated datasets will be designed for AI model development and targeted for applications, supporting the career trajectory and upward mobility of lower-income individuals with a particular focus on reskilling, computer science occupations, and allied health occupations.

Overview of Data

The dataset will consist of 6,000 questions with 1-3 answers each. It will comprise three domains:

- **General career or reskilling advice** (3,000 questions and accompanying answers)
- **Computer science professions** (1,500 questions and accompanying answers)
- **Allied health professions (e.g., nurses, medtechs)** (1,500 questions and accompanying answers)

The "general career advice" section offers guidance that's relevant to any profession, while the "computer science" and "allied health" categories dive deeper into field-specific advice. Each domain will have a set of shared subdomains like job search, professional development, and reskilling. This setup will keep the datasets compatible, allowing comparisons between general and specialized responses for each career question.

Annotation Process

The data will go through a 4-part labeling and annotation process:

1. To identify the subset of questions and answers that will comprise the dataset (out of the 60,000 questions in total), three steps will be taken.
    a. To distinguish question relevance in each domain, an LLM will be used to parse related metadata.
        i. General career advice will be identified by tags selected by the asker. An LLM will identify tags related to reskilling or general career navigation advice.

        ii. The field-specific domains, computer science and allied health professions, will be identified by using an LLM to categorize occupation metadata associated with each question.
    b. To identify the questions, an LLM will be fine-tuned to label questions with meaningful scenario topics. These scenarios include topics such as career exploration, industry trends, credentials, job search strategies, networking, and more. The LLM will be prompted to provide 0+ topic labels for each question, such that a topic label will not be forced if one does not fit for the question, but more than 1 (up to 6) can be provided if multiple topics are related to the question. A subset of questions (that will become the curated dataset) will be derived after the labeling process so as to include a proportionate amount of questions across all topic areas.
    c. Once questions are sampled based on scenarios, answers will be selected based on score (upvotes - downvotes) and recency, such that those with the greatest score and most recently provided will be selected.

2. Each answer will be rated by human annotators on three separate 4-point Likert scales:
    a. Correctness: To what extent is the answer factually accurate and appropriate?
    b. Completeness: How thoroughly does the answer address the question?
    c. Coherency: To what extent is the answer coherent and readable?

3. Each question will be labeled by human annotators according to the goal of the asker, such as whether the question aims to explore options, take action and make a decision, find resources, seek validation or support, etc..

Together, these labels and annotations are designed to help generative AI produce high-quality responses and improve the ability of agentic AI to understand the reasoning behind the questions asked, thus providing more responsive answers to users.

Availability of the Dataset

The dataset will be publicly available on Kaggle in early 2026 and will be made available for use under a CC BY license.

For more information on the dataset or the process, contact ulrich@renphil.org.