

The framework was trained and validated on a real-world dataset to ensure the results are robust and generalizable.

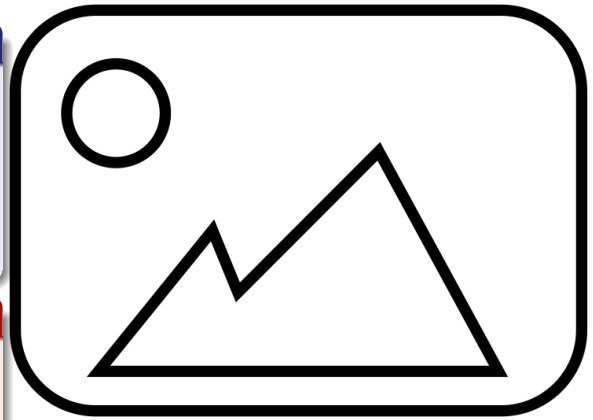
The PPM Corpus

The definitive experiments were conducted on a large corpus provided by the **Program Pathways Mapper (PPM)**:

- **Source:** Real-world data from California’s public colleges.
- **Ground Truth:** Course equivalency is defined by the Course Identification Numbering System (C-ID).
- **Scale:** The final, cleaned corpus consists of **2,157 courses** across 157 distinct C-ID classes.

Rigorous Evaluation Protocol

- **Partitioning:** A stratified 50/50 split was used to create non-overlapping training and test sets, ensuring all classes were represented in both.
- **Primary Metric:** The F_1 -Score was used as the primary metric to balance the trade-off between precision and recall.



2025-07-08

└ Introduction

└ Data & Evaluation

Data & Evaluation

The framework was trained and validated on a real-world dataset to ensure the results are robust and generalizable.

The PPM Corpus

The definitive experiments were conducted on a large corpus provided by the Program Pathways Mapper (PPM).

- **Source:** Real-world data from California’s public colleges.
- **Ground Truth:** Course equivalency is defined by the Course Identification Numbering System (C-ID).
- **Scale:** The final, cleaned corpus consists of 2,157 courses across 157 distinct C-ID classes.

Rigorous Evaluation Protocol

- **Partitioning:** A stratified 50/50 split was used to create non-overlapping training and test sets, ensuring all classes were represented in both.
- **Primary Metric:** The F_1 -Score was used as the primary metric to balance the trade-off between precision and recall.