Automating Course Articulation:
A Deep Metric Learning Framework Using Public Data

Mark S. Kim

San Francisco State University
Department of Data Science and Artificial Intelligence

July 8, 2025

Automating Course Articulation:
A Deep Metric Learning Framework Using Public Data

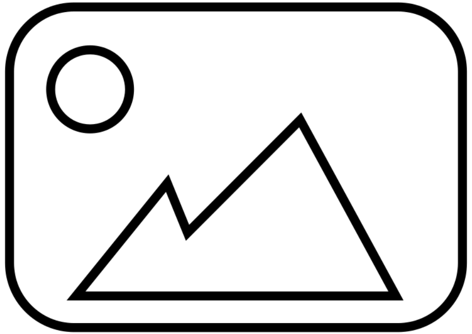
Mark S. Kim

San Francisco State University
Department of Data Science and Artificial Intelligence

July 8, 2025

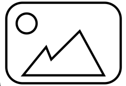
The Problem: The Transfer Maze

- The process for determining course equivalency, or **articulation**, is a formidable, largely manual process that creates significant barriers for students [15].
- In California's public system alone, articulation officers at **149 individual campuses** manually negotiate and update agreements [4, 9, 7, 6].
- This task of "bleak combinatorics" is inefficient, slow, and inherently intractable, struggling to keep pace with the needs of a vast and mobile student body [15].
- This is not a niche issue; transferring between institutions has become a normative part of the modern student's academic journey [1].



The Problem: The Transfer Maze

- The process for determining course equivalency, or **articulation**, is a formidable, largely manual process that creates significant barriers for students [15].
- In California's public system alone, articulation officers at **149 individual campuses** manually negotiate and update agreements [4, 9, 7, 6].
- This task of "bleak combinatorics" is inefficient, slow, and inherently intractable, struggling to keep pace with the needs of a vast and mobile student body [15].
- This is not a niche issue; transferring between institutions has become a normative part of the modern student's academic journey [1].

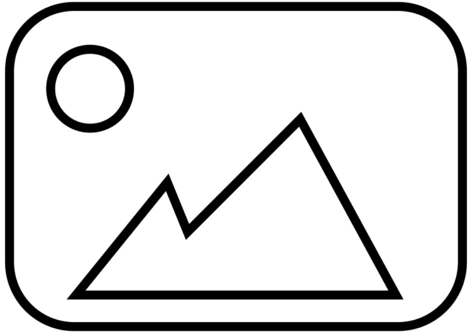


The High Cost to Students & Institutions

Consequences of an Inefficient System

The administrative friction of the transfer process creates a cascade of negative consequences that fall almost entirely on students.

- **Significant Credit Loss:** On average, transfer students lose an estimated **43%** of their earned academic credits, often forcing them to repeat courses [22, 1].
- **Increased Time-to-Degree:** Lost credits invariably delay graduation, postponing entry into the workforce and increasing overall educational time [22].
- **Greater Financial Burden:** Repeating courses increases total tuition costs and can exhaust a student's eligibility for financial aid, paradoxically making a cost-saving measure more expensive [22, 5].
- **Reduced Student Persistence:** The frustration of the process impacts student morale and contributes to lower



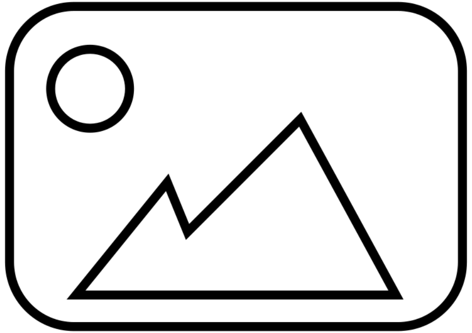
The High Cost to Students & Institutions
Consequences of an Inefficient System
The administrative friction of the transfer process creates a cascade of negative consequences that fall almost entirely on students.
- Significant Credit Loss: On average, transfer students lose an estimated 43% of their earned academic credits, often forcing them to repeat courses [22, 1].
- Increased Time-to-Degree: Lost credits invariably delay graduation, postponing entry into the workforce and increasing overall educational time [22].
- Greater Financial Burden: Repeating courses increases total tuition costs and can exhaust a student's eligibility for financial aid, paradoxically making a cost-saving measure more expensive [22, 5].
- Reduced Student Persistence: The frustration of the process impacts student morale and contributes to lower graduation rates for transfer students compared to their

A Critical Equity Issue

This is not just an administrative problem; it's an equity problem.

The barriers imposed by an inefficient articulation system fall most heavily on the very students institutions are striving to support [21].

- Low-income students and students from historically underrepresented racial and ethnic groups are more likely to begin their journey at community colleges and rely on transfer pathways [21].
- Recent growth in transfer enrollment has been driven disproportionately by Black and Hispanic students [8].
- This creates a **troubling feedback loop**: manual articulation causes credit loss, which imposes burdens that undermine institutional goals of closing equity gaps [21, 8].
- Therefore, automating course articulation is not merely an operational optimization; it is a **necessary intervention** to foster a more equitable educational ecosystem [5].



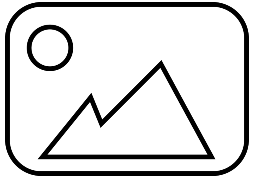
The Goal & My Contribution

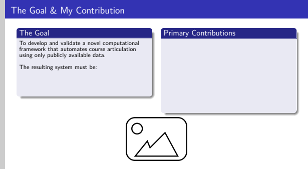
The Goal

To develop and validate a novel computational framework that automates course articulation using only publicly available data.

The resulting system must be:

Primary Contributions





The Goal & My Contribution

The Goal

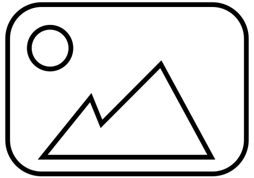
To develop and validate a novel computational framework that automates course articulation using only publicly available data.

The resulting system must be:

- Accurate & Scalable

Primary Contributions

1. **A Highly Accurate Framework:** Develop a complete, end-to-end pipeline that achieves state-of-the-art accuracy on real-world data.



The Goal & My Contribution
The Goal
To develop and validate a novel computational framework that automates course articulation using only publicly available data.
The resulting system must be:
Accurate & Scalable
Primary Contributions
A Highly Accurate Framework: Develop a complete, end-to-end pipeline that achieves state-of-the-art accuracy on real-world data.

The Goal & My Contribution

The Goal

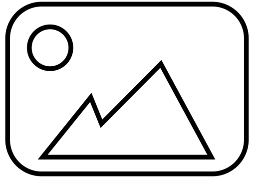
To develop and validate a novel computational framework that automates course articulation using only publicly available data.

The resulting system must be:

- Accurate & Scalable
- Computationally Efficient

Primary Contributions

1. **A Highly Accurate Framework:** Develop a complete, end-to-end pipeline that achieves state-of-the-art accuracy on real-world data.
2. **An Innovative Feature Vector:** Design a novel composite distance vector that improves classification by combining local and global semantic information.



The Goal & My Contribution
The Goal
To develop and validate a novel computational framework that automates course articulation using only publicly available data.
The resulting system must be:
- Accurate & Scalable
- Computationally Efficient
Primary Contributions
1. A Highly Accurate Framework: Develop a complete, end-to-end pipeline that achieves state-of-the-art accuracy on real-world data.
2. An Innovative Feature Vector: Design a novel composite distance vector that improves classification by combining local and global semantic information.

The Goal & My Contribution

The Goal

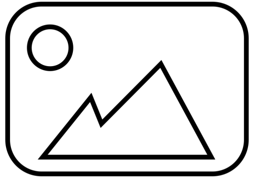
To develop and validate a novel computational framework that automates course articulation using only publicly available data.

The resulting system must be:

- Accurate & Scalable
- Computationally Efficient
- Inherently Privacy-Preserving

Primary Contributions

1. **A Highly Accurate Framework:** Develop a complete, end-to-end pipeline that achieves state-of-the-art accuracy on real-world data.
2. **An Innovative Feature Vector:** Design a novel composite distance vector that improves classification by combining local and global semantic information.
3. **An Efficient & Private Approach:** Create a solution that avoids the high costs and privacy concerns of prior methods by decoupling representation from classification.



2025-07-08

The Goal & My Contribution
The Goal
To develop and validate a novel computational framework that automates course articulation using only publicly available data.
The resulting system must be:
Accurate & Scalable
Computationally Efficient
Inherently Privacy-Preserving
Primary Contributions
1. A Highly Accurate Framework: Develop a complete, end-to-end pipeline that achieves state-of-the-art accuracy on real-world data.
2. An Innovative Feature Vector: Design a novel composite distance vector that improves classification by combining local and global semantic information.
3. An Efficient & Private Approach: Create a solution that avoids the high costs and privacy concerns of prior methods by decoupling representation from classification.

Agenda

Agenda
1. Introduction
2. Background & Related Work
3. A Decoupled Framework for Articulation
4. Experimental Setup & Results
5. Qualitative Analysis: Beyond the Metrics
6. Conclusion
7. Wrap Up

The Landscape of Automation

Prior attempts at automation have evolved, with each generation introducing new capabilities while also exposing new limitations.

Keyword & Statistical (TF-IDF)

- **Idea:** Weight terms based on statistical importance [2].
- **Limitation:** No semantic understanding; cannot grasp synonyms or context.

Contextual Embeddings (BERT)

- **Idea:** Generate vector representations of text that understand context [10, 18].
- **Limitation:** When used for direct classification, can be a "black box".

Enrollment-Based (course2vec)

- **Idea:** Learn course similarity from student co-enrollment patterns [16].
- **Limitation:** Requires sensitive student data, raising major privacy and generalizability issues [20].

Direct LLM Classification

- **Idea:** Use a large language model (e.g., GPT-4) as an end-to-end classifier.
- **Limitation:** High computational cost, opaque reasoning, and sensitive to prompt phrasing [11].

The Landscape of Automation
Prior attempts at automation have evolved, with each generation introducing new capabilities while also exposing new limitations.
Keyword & Statistical (TF-IDF)
Idea: Weight terms based on statistical importance [2].
Limitation: No semantic understanding; cannot grasp synonyms or context.
Contextual Embeddings (BERT)
Idea: Generate vector representations of text that understand context [10, 18].
Limitation: When used for direct classification, can be a "black box".
Enrollment-Based (course2vec)
Idea: Learn course similarity from student co-enrollment patterns [16].
Limitation: Requires sensitive student data, raising major privacy and generalizability issues [20].
Direct LLM Classification
Idea: Use a large language model (e.g., GPT-4) as an end-to-end classifier.
Limitation: High computational cost, opaque reasoning, and sensitive to prompt phrasing [11].

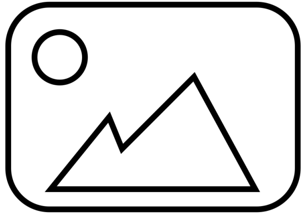
The Research Gap

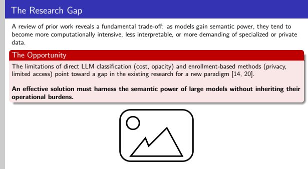
A review of prior work reveals a fundamental trade-off: as models gain semantic power, they tend to become more computationally intensive, less interpretable, or more demanding of specialized or private data.

The Opportunity

The limitations of direct LLM classification (cost, opacity) and enrollment-based methods (privacy, limited access) point toward a gap in the existing research for a new paradigm [14, 20].

An effective solution must harness the semantic power of large models without inheriting their operational burdens.





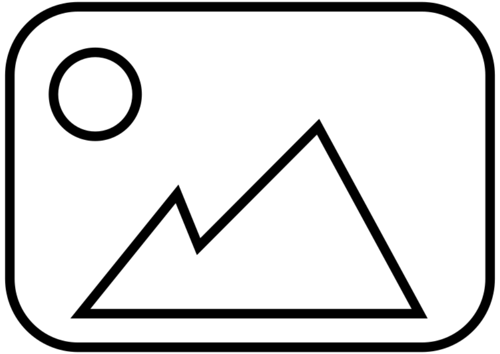
Agenda

Agenda

1 Introduction
2 Background & Related Work
3 A Decoupled Framework for Articulation
4 Experimental Setup & Results
5 Qualitative Analysis: Beyond the Metrics
6 Conclusion
7 Wrap Up

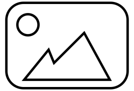
A Decoupled Framework: High-Level Architecture

Our framework's core principle is to **decouple rich semantic representation from the final classification task**. This creates a more efficient, scalable, and transparent system.



A Decoupled Framework: High-Level Architecture

Our framework's core principle is to **decouple rich semantic representation from the final classification task**. This creates a more efficient, scalable, and transparent system.



Core Principle: Decoupling Representation from Classification

By separating the process into two stages, we gain the semantic power of deep learning while avoiding the high operational costs of end-to-end LLM classification [11] and the privacy risks of enrollment-based methods [20].

Stage 1: Semantic Representation

The computationally intensive work of understanding language is done **once, offline**.

- A deep embedding model converts raw course text into a structured, reusable semantic vector.
- This captures the nuanced meaning and context of the course description.

Stage 2: Pairwise Classification

The classification of course pairs becomes **computationally cheap and fast**.

- A traditional machine learning model simply compares the pre-computed vectors.
- This allows for rapid, on-demand comparison of any two courses in the database.

The Benefit: The Best of Both Worlds

We leverage the power of transformers for deep semantic understanding without incurring their high inference costs for every comparison, creating a highly scalable system.

Step 1: Deep Contextual Embeddings

The first step is to convert unstructured course catalog text into a structured, semantically rich vector using a pre-trained transformer model [10, 18].

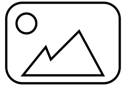
Input: Normalized Course Document

For each course, we create a single, consistent input string by concatenating four key fields:

- Department Name & Course Number
- Course Title
- Full Course Description

Output: A Semantic Vector

The model produces a high-dimensional vector (e.g., 384 dimensions) for each course. This vector represents the course's location in a "semantic space," where similar courses are positioned closer together.



Step 1: Deep Contextual Embeddings
The first step is to convert unstructured course catalog text into a structured, semantically rich vector using a pre-trained transformer model [10, 18].
Input: Normalized Course Document
For each course, we create a single, consistent input string by concatenating four key fields:
Department Name & Course Number
Course Title
Full Course Description
Output: A Semantic Vector
The model produces a high-dimensional vector (e.g., 384 dimensions) for each course. This vector represents the course's location in a "semantic space," where similar courses are positioned closer together.

Step 2: Our Novel Feature Vector (Δ_c)

To classify course pairs, we need features that represent the *relationship* between them. We designed a novel **composite distance vector (Δ_c)** to provide the classifier with a richer, more discriminative feature set.

Combining Local & Global Information

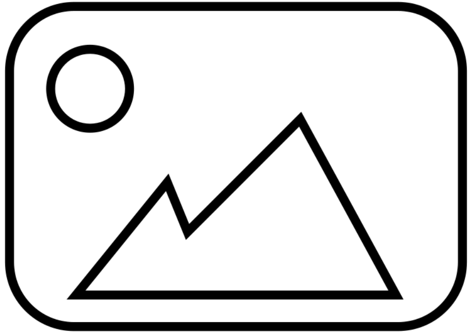
The vector combines two distinct types of information:

- **Local Disparities:** The granular, dimension-by-dimension difference between the two course vectors.
- **Global Alignment:** A single, holistic score of their overall similarity in the semantic space.

The Formula

For two k-dimensional course vectors, A and B, the composite vector Δ_c is constructed by concatenating their element-wise difference with their cosine similarity:

$$\Delta_c = \left(a_1 - b_1, \ldots, a_k - b_k, \frac{A \cdot B}{\| A \| \| B \|} \right)$$



Step 2: Our Novel Feature Vector (Δ_c)
To classify course pairs, we need features that represent the *relationship* between them. We designed a novel **composite distance vector (Δ_c)** to provide the classifier with a richer, more discriminative feature set.
Combining Local & Global Information
The vector combines two distinct types of information:
- **Local Disparities:** The granular, dimension-by-dimension difference between the two course vectors.
- **Global Alignment:** A single, holistic score of their overall similarity in the semantic space.
The Formula
For two k-dimensional course vectors, A and B, the composite vector Δ_c is constructed by concatenating their element-wise difference with their cosine similarity:
$$\Delta_c = \left(a_1 - b_1, \ldots, a_k - b_k, \frac{A \cdot B}{\| A \| \| B \|} \right)$$

Step 3: Domain-Specific Fine-Tuning

General-purpose models lack the specialized "vocabulary" for academic text. To create a more discriminative embedding space, we fine-tune a pre-trained model on our course data using **deep metric learning**.

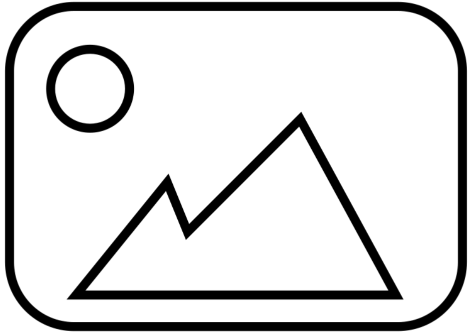
Learning Objective: The Triplet Loss

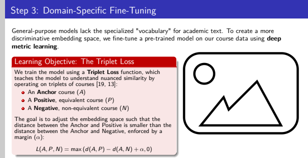
We train the model using a **Triplet Loss** function, which teaches the model to understand nuanced similarity by operating on triplets of courses [19, 13]:

- An **Anchor** course (A)
- A **Positive**, equivalent course (P)
- A **Negative**, non-equivalent course (N)

The goal is to adjust the embedding space such that the distance between the Anchor and Positive is smaller than the distance between the Anchor and Negative, enforced by a margin (α):

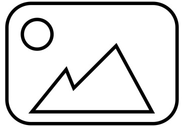
$$L(A, P, N) = \max\left(d(A, P) - d(A, N) + \alpha, 0\right)$$





Step 4: Downstream Classification

The final step is to feed the engineered composite distance vectors (Δ_c) into efficient, traditional machine learning classifiers for the final equivalency prediction. This decoupled approach allows for rapid, on-demand comparisons.



Classifier Evaluation

We evaluated a comprehensive suite of models. The top finalists were:

- Support Vector Machine (SVM)
- Random Forest (RF)
- XGBoost
- K-Nearest Neighbors (KNN)

Key Finding: Accuracy vs. Efficiency

A clear trade-off emerged from our analysis:

- **Maximum Accuracy:** SVM proved to be the most accurate and consistent model statistically.
- **Optimal Efficiency:** Random Forest and XGBoost were an order of magnitude faster at inference time, making them ideal for scalable, real-world deployment.

Step 4: Downstream Classification
The final step is to feed the engineered composite distance vectors (Δ_c) into efficient, traditional machine learning classifiers for the final equivalency prediction. This decoupled approach allows for rapid, on-demand comparisons.
Classifier Evaluation
We evaluated a comprehensive suite of models. The top finalists were:
Support Vector Machine (SVM)
Random Forest (RF)
XGBoost
K-Nearest Neighbors (KNN)
Key Finding: Accuracy vs. Efficiency
A clear trade-off emerged from our analysis:
Maximum Accuracy: SVM proved to be the most accurate and consistent model statistically.
Optimal Efficiency: Random Forest and XGBoost were an order of magnitude faster at inference time, making them ideal for scalable, real-world deployment.

Data & Evaluation

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

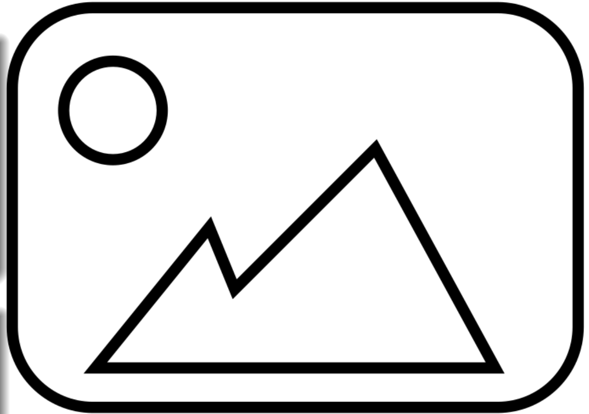
The PPM Corpus

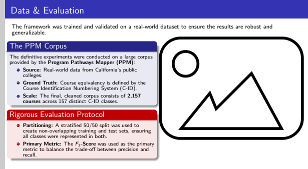
The definitive experiments were conducted on a large corpus provided by the **Program Pathways Mapper (PPM)**:

- **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.



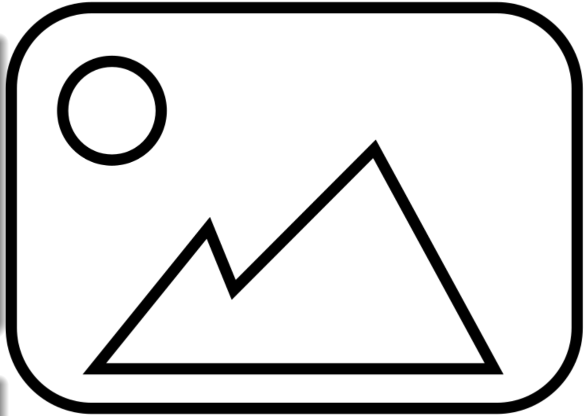


Finding 1: Domain-Specific Fine-Tuning is Critical

Our experiments show that adapting a generic model to the specific language of academia is more effective than relying on sheer model scale.

Superior Performance

- The fine-tuned model (**bge-ft**) achieved the highest mean F_1-score and the lowest variance on the held-out test data.
- A one-way ANOVA and subsequent Games-Howell post-hoc test confirmed that our fine-tuned model was **statistically superior** to all off-the-shelf models evaluated.
- This includes models that were orders of magnitude larger, demonstrating that targeted adaptation is more effective than scale for this specialized task.



Key Insight

For specialized domains, creating a bespoke embedding space through fine-tuning is crucial. It teaches the model the specific semantics and nuances required to make fine-grained distinctions that general-purpose

Finding 2: Final Classifier Performance

The final evaluation, conducted on the held-out test data, measured the efficacy and efficiency of the top-performing classifiers.

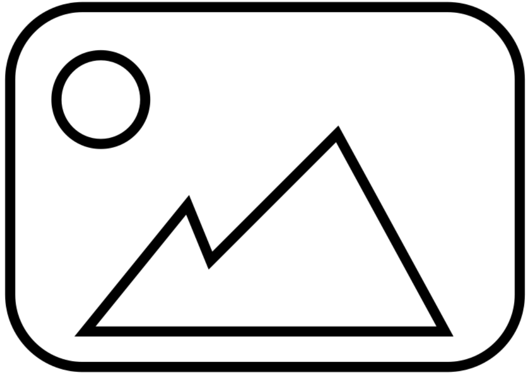
Exceptional Performance Across the Board

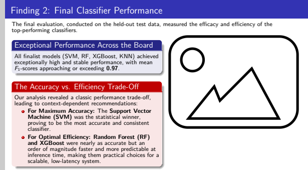
All finalist models (SVM, RF, XGBoost, KNN) achieved exceptionally high and stable performance, with mean F_1-scores approaching or exceeding **0.97**.

The Accuracy vs. Efficiency Trade-Off

Our analysis revealed a classic performance trade-off, leading to context-dependent recommendations:

- **For Maximum Accuracy:** The **Support Vector Machine (SVM)** was the statistical winner, proving to be the most accurate and consistent classifier.
- **For Optimal Efficiency: Random Forest (RF) and XGBoost** were nearly as accurate but an order of magnitude faster and more predictable at inference time, making them practical choices for a scalable, low-latency system.





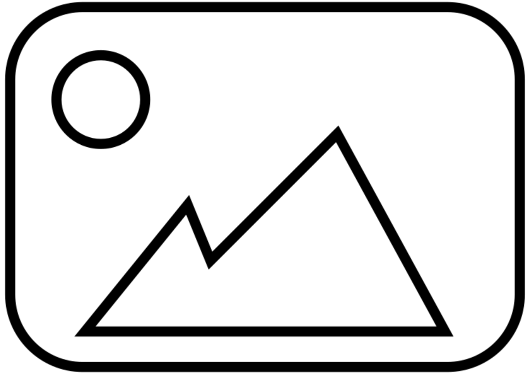
Qualitative Analysis: Why Do Errors Still Occur?

With such high accuracy, it's crucial to ask why the system isn't perfect. A purely numerical analysis can be misleading, as aggregate scores can hide systematic failure modes [12].

Shared Misclassifications Across Models

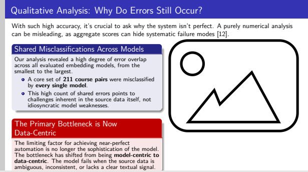
Our analysis revealed a high degree of error overlap across all evaluated embedding models, from the smallest to the largest.

- A core set of **211 course pairs** were misclassified by **every single model**.
- This high count of shared errors points to challenges inherent in the source data itself, not idiosyncratic model weaknesses.



The Primary Bottleneck is Now Data-Centric

The limiting factor for achieving near-perfect automation is no longer the sophistication of the model. The bottleneck has shifted from being **model-centric to data-centric**. The model fails when the source data is



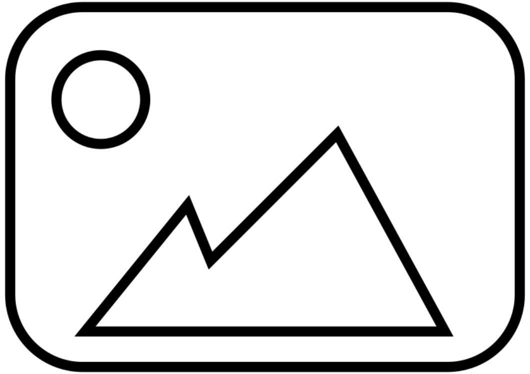
Shared Misclassifications: A Data-Centric Problem

To diagnose the source of errors, we analyzed their overlap across all embedding models—from the large, general-purpose models to our small, fine-tuned specialist. The results provide strong evidence that the errors are systematic.

The Finding: Models Agree on What's Hard

A significant portion of failures are not random but are systematic products of the course catalog data itself.

- A large number of misclassified course pairs were common to all model combinations.
- This indicates that these "hard" examples consistently challenge a wide range of semantic models.
- Such errors often arise from annotation artifacts, inherent ambiguity in the source text, or insufficient information to support a clear classification.



Interpretation

These shared failures strongly suggest that the errors

Shared Misclassifications: A Data-Centric Problem
To diagnose the source of errors, we analyzed their overlap across all embedding models—from the large, general-purpose models to our small, fine-tuned specialist. The results provide strong evidence that the errors are systematic.
The Finding: Models Agree on What's Hard
A significant portion of failures are not random but are systematic products of the course catalog data itself.
A large number of misclassified course pairs were common to all model combinations.
This indicates that these "hard" examples consistently challenge a wide range of semantic models.
Such errors often arise from annotation artifacts, inherent ambiguity in the source text, or insufficient information to support a clear classification.
Interpretation
These shared failures strongly suggest that the errors stem from the data itself, not the models. The models

Root Cause Analysis: Common Error Patterns

A manual case-based review of the "hard" examples reveals that most errors fall into predictable categories caused by issues in the source data.

False Negatives (Missed Equivalencies)

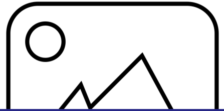
A False Negative occurs when the system fails to identify a true equivalence. These are primarily caused by:

- **Semantic Divergence:** Officially equivalent courses are described with vastly different terminology or pedagogical focus. The model correctly assesses the texts as dissimilar; the failure lies in the inconsistent source data.
- **Minimalist Descriptions:** One or both course descriptions in a pair are too sparse or incomplete to provide enough textual signal for the model to establish a match.

False Positives (Incorrect Equivalencies)

A False Positive occurs when two non-equivalent courses are incorrectly approved. These are primarily caused by:

- **Topical Overlap:** Courses cover the same broad subject but differ critically in academic level or their position in a sequence (e.g., Physics I vs. Physics II).
- **Vague Descriptions:** Descriptions use generic language, lacking the specific detail needed for differentiation. This is a known challenge in short-text semantic similarity [3].



Root Cause Analysis: Common Error Patterns
A manual case-based review of the "hard" examples reveals that most errors fall into predictable categories caused by issues in the source data.
False Negatives (Missed Equivalencies)
A False Negative occurs when the system fails to identify a true equivalence. These are primarily caused by:
Semantic Divergence: Officially equivalent courses are described with vastly different terminology or pedagogical focus. The model correctly assesses the texts as dissimilar; the failure lies in the inconsistent source data.
Minimalist Descriptions: One or both course descriptions in a pair are too sparse or incomplete to provide enough textual signal for the model to establish a match.
False Positives (Incorrect Equivalencies)
A False Positive occurs when two non-equivalent courses are incorrectly approved. These are primarily caused by:
Topical Overlap: Courses cover the same broad subject but differ critically in academic level or their position in a sequence (e.g., Physics I vs. Physics II).
Vague Descriptions: Descriptions use generic language, lacking the specific detail needed for differentiation. This is a known challenge in short-text semantic similarity [3].

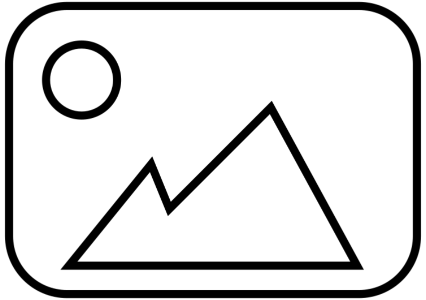
Conclusion of Analysis: The Primary Bottleneck

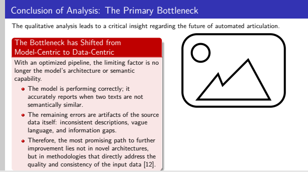
The qualitative analysis leads to a critical insight regarding the future of automated articulation.

The Bottleneck has Shifted from Model-Centric to Data-Centric

With an optimized pipeline, the limiting factor is no longer the model's architecture or semantic capability.

- The model is performing correctly; it accurately reports when two texts are not semantically similar.
- The remaining errors are artifacts of the source data itself: inconsistent descriptions, vague language, and information gaps.
- Therefore, the most promising path to further improvement lies not in novel architectures, but in methodologies that directly address the quality and consistency of the input data [12].





Limitations

While the proposed framework represents a significant advance, it is essential to acknowledge the boundaries of the current study.

Key Limitations

- **Performance is Capped by Data Quality:** The system's performance is fundamentally limited by the quality and content of the public course descriptions. It cannot infer information that is absent from vague, minimalist, or inconsistent source texts.
- **Generalizability of the Fine-Tuned Model:** The specialized *bge-ft* model was tuned on data from California's public colleges. Its performance may not be as high "out-of-the-box" in other contexts (e.g., private or non-US institutions) without re-tuning on local data.
- **Handling of Complex Articulation Rules:** The framework simplifies articulation into a binary classification of course pairs and does not natively handle complex one-to-many or many-to-many agreements, a challenge that persists for many automated systems [14].

Limitations

While the proposed framework represents a significant advance, it is essential to acknowledge the boundaries of the current study.

Key Limitations

- Performance is Capped by Data Quality: The system's performance is fundamentally limited by the quality and content of the public course descriptions. It cannot infer information that is absent from vague, minimalist, or inconsistent source texts.
- Generalizability of the Fine-Tuned Model: The specialized bge-ft model was tuned on data from California's public colleges. Its performance may not be as high "out-of-the-box" in other contexts (e.g., private or non-US institutions) without re-tuning on local data.
- Handling of Complex Articulation Rules: The framework simplifies articulation into a binary classification of course pairs and does not natively handle complex one-to-many or many-to-many agreements, a challenge that persists for many automated systems [14].

The findings and limitations of this study give rise to several promising avenues for future research that build upon this work.

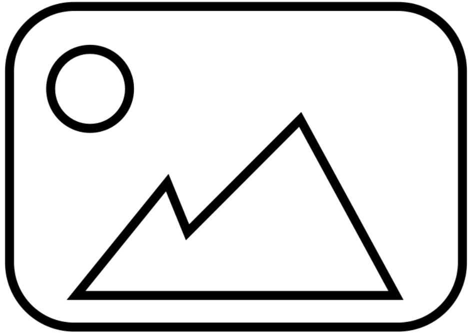
Data-Centric AI Strategies

Since data quality is the primary bottleneck, future work should focus on:

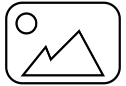
- Developing an interactive, **human-in-the-loop** system for expert review of ambiguous pairs.
- Exploring dynamic data augmentation, such as requesting a full syllabus when classification confidence is low.

Expanding Framework Capabilities

- Evolving the framework into a full-scale **course recommendation engine** with a conversational interface.
- Investigating graph-based methods (e.g., GNNs) to identify and model complex one-to-many articulation rules.



Future Work
The findings and limitations of this study give rise to several promising avenues for future research that build upon this work.
Data-Centric AI Strategies
Since data quality is the primary bottleneck, future work should focus on:
Developing an interactive, human-in-the-loop system for expert review of ambiguous pairs.
Exploring dynamic data augmentation, such as requesting a full syllabus when classification confidence is low.
Expanding Framework Capabilities
Evolving the framework into a full-scale course recommendation engine with a conversational interface.
Investigating graph-based methods (e.g., GNNs) to identify and model complex one-to-many articulation rules.

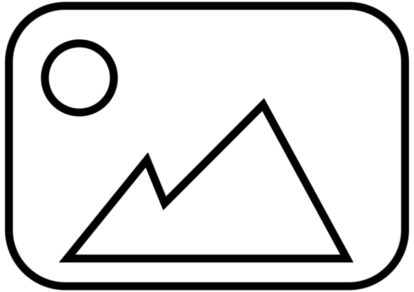


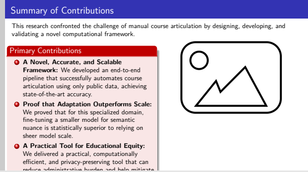
Summary of Contributions

This research confronted the challenge of manual course articulation by designing, developing, and validating a novel computational framework.

Primary Contributions

1. **A Novel, Accurate, and Scalable Framework:** We developed an end-to-end pipeline that successfully automates course articulation using only public data, achieving state-of-the-art accuracy.
2. **Proof that Adaptation Outperforms Scale:** We proved that for this specialized domain, fine-tuning a smaller model for semantic nuance is statistically superior to relying on sheer model scale.
3. **A Practical Tool for Educational Equity:** We delivered a practical, computationally efficient, and privacy-preserving tool that can





Thank You!!!

Questions?

Thank You!!!

Questions?

Contact & Acknowledgments

Contact Information

Mark S. Kim

- mkim22@mail.sfsu.edu

Acknowledgments

I would like to express my deepest appreciation to:

- My advisors, Professors **Hui Yang**, **Arno Puder**, and **Anagha Kulkarni**, for their invaluable guidance and support.
- Professor **Tao He**, for the crucial suggestion to incorporate a global similarity metric into the feature vector.
- The **Program Pathways Mapper (PPM)** team for providing the foundational data for this work.
- The **SFSU Academic Technology Systems Team** for their support and the use of the POLARIS High-Performance Computing cluster.

Contact & Acknowledgments
Contact Information
Mark S. Kim
mkim22@mail.sfsu.edu
Acknowledgments
I would like to express my deepest appreciation to:
My advisors, Professors Hui Yang, Arno Puder, and Anagha Kulkarni, for their invaluable guidance and support.
Professor Tao He, for the crucial suggestion to incorporate a global similarity metric into the feature vector.
The Program Pathways Mapper (PPM) team for providing the foundational data for this work.
The SFSU Academic Technology Systems Team for their support and the use of the POLARIS High-Performance Computing cluster.

References I

[1] Public Agenda. *Beyond Transfer: Insights from a Survey of American Adults.* URL: https://publicagenda.org/resource/beyond-transfer/ (visited on 06/30/2025).

[2] Akiko Aizawa. "An information-theoretic perspective of tf-idf measures". In: *Information Processing & Management* 39.1 (2003), pp. 45–65. ISSN: 0306-4573. DOI: https://doi.org/10.1016/S0306-4573(02)00021-3. URL: https://www.sciencedirect.com/science/article/pii/S0306457302000213.

[3] Zaira Hassan Amur et al. "Short-Text Semantic Similarity (STSS): Techniques, Challenges and Future Perspectives". In: *Applied Sciences* 13.6 (2023). ISSN: 2076-3417. DOI: 10.3390/app13063911. URL: https://www.mdpi.com/2076-3417/13/6/3911.

[4] ASSIST. *Frequently Asked Questions.* URL: https://resource.assist.org/FAQ (visited on 06/30/2025).

[5] Leticia Tomas Bustillos et al. *The Transfer Maze: The High Cost to Students and the State of California.* The Campaign for College Opportunity, Sept. 17, 2017.

References I

[1] Public Agenda. *Beyond Transfer: Insights from a Survey of American Adults.* URL: https://publicagenda.org/resource/beyond-transfer/ (visited on 06/30/2025).

[2] Akiko Aizawa. "An information-theoretic perspective of tf-idf measures". In: *Information Processing & Management* 39.1 (2003), pp. 45–65. ISSN: 0306-4573. DOI: https://doi.org/10.1016/S0306-4573(02)00021-3. URL: https://www.sciencedirect.com/science/article/pii/S0306457302000213.

[3] Zaira Hassan Amur et al. "Short-Text Semantic Similarity (STSS): Techniques, Challenges and Future Perspectives". In: *Applied Sciences* 13.6 (2023). ISSN: 2076-3417. DOI: 10.3390/app13063911. URL: https://www.mdpi.com/2076-3417/13/6/3911.

[4] ASSIST. *Frequently Asked Questions.* URL: https://resource.assist.org/FAQ (visited on 06/30/2025).

[5] Leticia Tomas Bustillos et al. *The Transfer Maze: The High Cost to Students and the State of California.* The Campaign for College Opportunity, Sept. 17, 2017.

References II

[6] California Community Colleges Chancellor's Office. *Management Information Systems Data Mart.* 2024. URL: https://datamart.cccco.edu/Students/Student%5C_Headcount%5C_Term%5C_Annual.aspx (visited on 09/13/2024).

[7] California State University Office of the Chancellor. *Enrollment.* 2024. URL: https://www.calstate.edu/csu-system/about-the-csu/facts-about-the-csu/enrollment (visited on 09/13/2024).

[8] National Student Clearinghouse. *College Transfer Enrollment Grew by 5.3% in the Fall of 2023.* URL: https://www.studentclearinghouse.org/news/college-transfer-enrollment-grew-by-5-3-in-the-fall-of-2023/ (visited on 06/30/2025).

[9] Kevin Cook. *California's Higher Education System.* 2024. URL: https://www.ppic.org/publication/californias-higher-education-system/ (visited on 09/06/2024).

References II

[6] California Community Colleges Chancellor's Office. *Management Information Systems Data Mart.* 2024. URL: https://datamart.cccco.edu/Students/Student%5C_Headcount%5C_Term%5C_Annual.aspx (visited on 09/13/2024).

[7] California State University Office of the Chancellor. *Enrollment.* 2024. URL: https://www.calstate.edu/csu-system/about-the-csu/facts-about-the-csu/enrollment (visited on 09/13/2024).

[8] National Student Clearinghouse. *College Transfer Enrollment Grew by 5.3% in the Fall of 2023.* URL: https://www.studentclearinghouse.org/news/college-transfer-enrollment-grew-by-5-3-in-the-fall-of-2023/ (visited on 06/30/2025).

[9] Kevin Cook. *California's Higher Education System.* 2024. URL: https://www.ppic.org/publication/californias-higher-education-system/ (visited on 09/06/2024).

[10] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805.

[11] Federico Errica et al. "What Did I Do Wrong? Quantifying LLMs' Sensitivity and Consistency to Prompt Engineering". In: *ArXiv* abs/2406.12334 (2024). URL: https://api.semanticscholar.org/CorpusID:270562829.

[12] Gabrielle Gauthier-melancon et al. "Azimuth: Systematic Error Analysis for Text Classification". In: Jan. 2022, pp. 298–310. DOI: 10.18653/v1/2022.emnlp-demos.30.

[13] Alexander Hermans, Lucas Beyer, and Bastian Leibe. *In Defense of the Triplet Loss for Person Re-Identification*. 2017. arXiv: 1703.07737 [cs.CV]. URL: https://arxiv.org/abs/1703.07737.

[14] Z. A Pardos, H Chau, and H Zhao. "Data-assistive course-to-course articulation using machine translation". In: *Proceedings of the Sixth Conference on Learning@ Scale*. 2019, pp. 1–10.

[15] Zachary Pardos, Hung Chau, and Haocheng Zhao. "Data-Assistive Course-to-Course Articulation Using Machine Translation". In: (July 2019). DOI: 10.1145/3330430.3333622.

References IV

[16] Zachary A. Pardos, Zihao Fan, and Weijie Jiang. "Connectionist recommendation in the wild: on the utility and scrutability of neural networks for personalized course guidance". In: *User Modeling and User-Adapted Interaction* 29.2 (Apr. 2019), pp. 487–525. ISSN: 0924-1868. DOI: 10.1007/s11257-019-09218-7. URL: https://doi.org/10.1007/s11257-019-09218-7.

[17] Stephen Porter. "Assessing Transfer and Native Student Performance at Four-Year Institutions". In: *39th Annual Forum of the Association for Institutional Research*. June 1999.

[18] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019. URL: https://arxiv.org/abs/1908.10084.

[19] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "FaceNet: A Unified Embedding for Face Recognition and Clustering". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015.

References V

[20] Sharon Slade and Paul Prinsloo. “Learning Analytics: Ethical Issues and Dilemmas”. In: *American Behavioral Scientist* 57.10 (2013), pp. 1510–1529. DOI: 10.1177/0002764213479366. eprint: https://doi.org/10.1177/0002764213479366. URL: https://doi.org/10.1177/0002764213479366.

[21] The National Task Force on the Transfer and Award of Credit. *Reimagining Transfer for Student Success*. Report to Congressional Requesters. American Council on Education, Mar. 2020. URL: https://www.gao.gov/products/gao-17-574.

[22] United States Government Accountability Office. *Higher Education: Students Need More Information to Help Reduce Challenges in Transferring College Credits*. Report to Congressional Requesters GAO-17-574. United States Government Accountability Office, Aug. 14, 2017. URL: https://www.gao.gov/products/gao-17-574.

References V