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

4 D > 4 D > 4 E > 4 E > E 990

Automating Course Articulation $\begin{tabular}{l} \begin{tabular}{l} \begin{tabular}{l}$

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

Mark S. Kim

San Frankon State University

Department of Data Science and Artificial Intelligent

July 8, 2025

Mark S. Kim (SFSU)

Automating Course Articulation

July 8, 2025

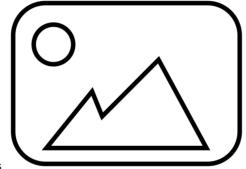
1/32

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].

Automating Course Articulation

Mark S. Kim (SESU)





July 8, 2025 2 / 32

Automating Course Articulation

-Introduction

☐ The Problem: The Transfer Maze

blem: The Transfer Maze

- a The process for determining course equivalency, or articulation, is a formidable, largely manual process that
- In California's public system alone, articulation officers at
 149 individual campuses manually negotiate and
 undate accounts 14 9 7 61
 - 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 institution has become a normative part of the modern student's student's part of the modern student's



The High Cost to Students & Institutions

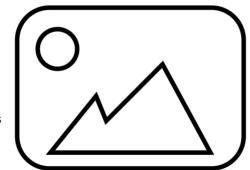
Consequences of an Inefficient System

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

Automating Course Articulation

- **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

Mark S. Kim (SESU)





July 8, 2025

Automating Course Articulation

Introduction

The High Cost to Students & Institutions

The High Cost to Students & Institutions

quences of an Inefficient System

istrative friction of the transfer process creates a rely on students.

- Significant Credit Loss: On average, transfer studen ose an estimated 43% of their earned academic credit
- 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 122; 51.
- noise and, paradoscany making a cost-saving e more expensive [22, 5] ed Student Persistence: The frustration of the impacts attribute morals and contributes to low

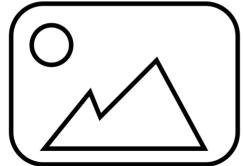


A Critical Equity Issue

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].





Mark S. Kim (SESU)

Automating Course Articulation

luly 8, 2025

4/32

Automating Course Articulation

-Introduction

A Critical Equity Issue

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 striving to succerd [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

- begin their journey at community colleges and rely on pathways [21].

 Recent growth in transfer enrollment has been driven discroportionately by Black and Hispanic students (III.
- a This creates a troubling feedback loop: manual articulation causes credit loss, which imposes bardons that undermine institutional goals of closing equity gaps [21, 8].
- 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 economiem [5].



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





Automating Course Articulation

Introduction

The Goal & My Contribution

The Goal & My Contribution

Mark S. Kim (SFSU) Automating Course Articulation July 8, 2025 5/32

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

Mark S. Kim (SFSU)

Primary Contributions

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





Automating Course Articulation July 8, 2025 5 / 32

Automating Course Articulation
Introduction

☐ 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

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





Automating Course Articulation

Introduction

The Goal & My Contribution

The Goal

The Horizontal and Automating Course Articulation

The Goal

The Goal & My Contribution

The Goal

The Horizontal and Automating Course

The Goal & My Contribution

The Goal

The Horizontal and Automating Course

The Goal & My Contribution

The Goal

The Horizontal and Automating Course

The Goal & My Contribution

The Goal

The Horizontal and Automating Course

The Goal

The Horizontal Automating Course

The Horizontal Automati

To develop and validate a novel computat framework that automates course articula using only publicly available data. The resulting system must be: A Accurate & Scalable a Computationally Efficient

citical
A Highly Accurate Framework: Devolop
strion
strion of policy and to end pipeline that achieve
state-of-the-art accuracy on sal-world da
An Inscoutible Feature Vector: Design
novel composite distance vector that
improves classification by combining local
and global semantic information.



☐ The Goal & My Contribution

Mark S. Kim (SESU)

Automating Course Articulation

July 8, 2025

/ 32

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

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





Mark S. Kim. (SESU) Automating Course Articulation July 8, 2025 5/32

Automating Course Articulation

-Introduction

The Goal & My Contribution

Primary Contributi

To develop and validate a novel computational framework that automates course articulation of the automates course articulation of the automates course articulation of the automates of the auto

The Goal & My Contribution

complete, and-to-end pipeline that achieves state-of-the-off accuracy on end-world data.

An inseconduce Feature Wecker: Design a improves classification by combining local and global semantic leformation.

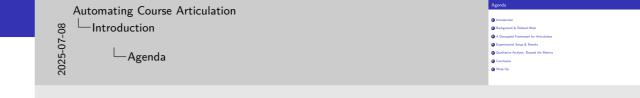
An Efficient B. Private Approach: Createsolution that sould the high costs and decopping representation from classification decopping representation from classification



Agenda

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





Mark S. Kim (SFSU) Automating Course Articulation July 8, 2025 6/32

The Landscape of Automation

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

Automating Course Articulation

Keyword & Statistical (TF-IDF)

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

Enrollment-Based (course2vec)

Mark S. Kim (SESU)

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

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".

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].



July 8, 2025

Automating Course Articulation

Background & Related Work

The Landscape of Automation



The Research Gap

Mark S. Kim (SESU)

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.



Automating Course Articulation



Automating Course Articulation

Background & Related Work

The Research Gap

The Research Gap

reverse of pin work works a fundamental trade-off in models gain summits power, they tend to
the contraction of the contraction of the contraction of supplied or protect
that the contraction of the contra

July 8, 2025

Agenda

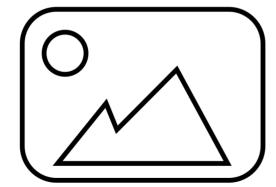
- Introduction
- 2 Background & Related Work
- A Decoupled Framework for Articulation
- 4 Experimental Setup & Results
- Qualitative Analysis: Beyond the Metrics
- 6 Conclusion
- Wrap Up



◆□▶◆□▶◆□▶◆□▶ ■ り<0

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.





Mark S. Kim (SFSU) Automating Course Articulation July 8, 2025 10 / 32

Automating Course Articulation

2025-07-08

-A Decoupled Framework for Articulation

A Decoupled Framework: High-Level Architecture

A Decoupled Framework: High-Level Architecture

ramework's core principle is to decouple rich semantic representation from the fin ification 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.



Mark S. Kim (SFSU) Automating Course Articulation July 8, 2025 11/32

Automating Course Articulation

A Decoupled Framework for Articulation

Core Principle: Decoupling Representation from Classification

Core Principle: Decoupling Representation from Classification

by spending the process into two stages, or goin the sensotic power of deep leaving while avoiding

by high operations close of end-to-end LLM classification [1] and the privacy risk of

continuous based methods [26].

Committee of the continuous closes of t

The constantionally intensive work of
mandermanding lengings in door ence, offline.

A deep enhedding model converts raw
course too little a structured, reseable
senseratic vector.

This capsume the reaccompanied vectors.

This allows for rigid, or-demand companie
of the dischalase.

Both Worlds emers for deep semantic understanding without incurring their high inference

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.





Automating Course Articulation

A Decoupled Framework for Articulation

Step 1: Deep Contextual Embeddings

Step 1: Deep Contentual Embeddings

The first kep is to remove unstructural came stating test into a disturbaned, sensorically risk vertice using a per install residence made [16, 18].

[Input: Neumalized Control Decorates

For less forces, we care a large, sensories to per single per sensories for key folds.

• Control Tale

• Control Tale

• Control Tale

• Control Tale

• Despit: A Sensories Vertice

Obligat: A Sensories Vertice

Despit: A Sensories Vertice



Mark S. Kim (SFSU)

Automating Course Articulation

July 8, 2025

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

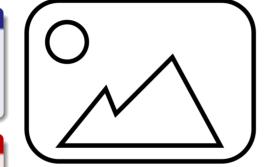
The vector combines two distinct types of information:

- 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:

$$\Delta_c = \left(extbf{a}_1 - extbf{b}_1, \ldots, extbf{a}_k - extbf{b}_k, rac{ extbf{A} \cdot extbf{B}}{\parallel extbf{A} \parallel \parallel extbf{B} \parallel}
ight)$$





Automating Course Articulation

A Decoupled Framework for Articulation

Step 2: Our Novel Feature Vector (Δ_c)



Mark S. Kim (SFSU) Automating Course Articulation July 8, 2025 13/32

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.

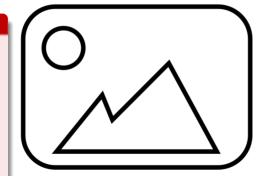
Learning Objective: The Triplet Loss

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):

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





Mark S. Kim (SESU) Automating Course Articulation July 8, 2025 Automating Course Articulation

A Decoupled Framework for Articulation

Step 3: Domain-Specific Fine-Tuning

· An Anchor course (A) A Position equivalent course (2) a & Mountless non-equipplent course (N)

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

The goal is to adjust the embedding space such that the distance between the Anchor and Positive is smaller than t

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.



Automating Course Articulation

A Decoupled Framework for Articulation

Step 4: Downstream Classification



Mark S. Kim (SESU)

Automating Course Articulation

July 8, 2025

15 / 32

Data & Evaluation

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

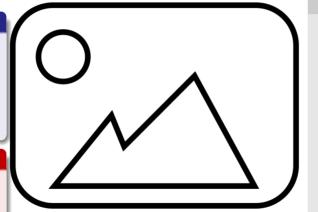
The PPM Corpus

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





Automating Course Articulation Experimental Setup & Results

☐Data & Evaluation

Data & Evaluation The framework was trained and walldated on a real-world dataset to ensure the results are robust and 4 Source Paul and dura from California's mildir . Ground Touth: Course equipplency is defined by the 4 Scale: The final, cleaned corpus consists of 2,157 operates across 157 distinct C-ID classes. ertitioning: A stratified 50/50 solit was used to Partitioning: A stratified 50/50 split was used to create non-overlapping training and test sets, end blickness were recovered in both

Primary Metric: The Fi-Score was used as the primare metric to balance the trade-off between precision and

Mark S. Kim (SESU) Automating Course Articulation July 8, 2025

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.

Automating Course Articulation

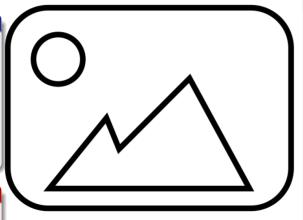
Superior Performance

- 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

Mark S. Kim (SFSU)





July 8, 2025

Automating Course Articulation Experimental Setup & Results

Finding 1: Domain-Specific Fine-Tuning is Critical

Finding 1: Domain-Specific Fine-Tuning is Critical post-noc test committed that our fine-tuned it was statistically superior to all off-the-shelf This includes models that were orders of magn

space through fine-tuning is crucial. It teaches the model the specific semantics and nuances required to make fine-grained distinctions that general-outspace

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

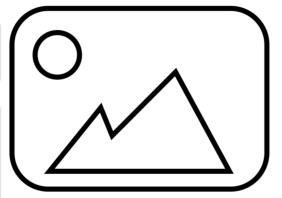
All finalist models (SVM, RF, XGBoost, KNN) achieved exceptionally high and stable performance, with mean F_1 -scores approaching or exceeding **0.97**.

The Accuracy vs. Efficiency Trade-Off

Mark S. Kim (SFSU)

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.





Automating Course Articulation Iuly 8, 2025 18 / 32

Automating Course Articulation

Experimental Setup & Results

Finding 2: Final Classifier Performance

The company of the co

Finding 2: Final Classifier Performance

4 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 enters.

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].

Automating Course Articulation

Shared Misclassifications Across Models

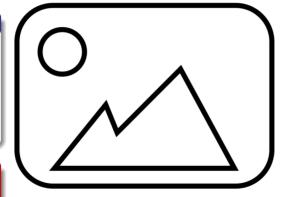
Our analysis revealed a high degree of error overlap across all evaluated embedding models, from the smallest to the largest.

- 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

Mark S. Kim (SESU)

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





July 8, 2025

Automating Course Articulation Qualitative Analysis: Beyond the Metrics Qualitative Analysis: Why Do Errors Still Occur? Qualitative Analysis: Why Do Errors Still Occur? A core set of 211 course pairs were miscle This high count of shared errors points to

automation is no longer the sophistication of the model.

Shared Misclassifications: A Data-Centric Problem

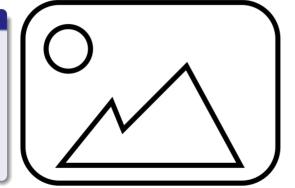
To diagnose the source of errors, we analyzed their overlap across all embedding models—from the large, general-purpose models to our small, fine-tuned specialist. The results provide strong evidence that the errors are systematic.

Automating Course Articulation

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

Mark S. Kim (SESU)

These shared failures strongly suggest that the errors

イロトイプトイミトイミト ヨーの90

July 8, 2025

Automating Course Articulation

Qualitative Analysis: Beyond the Metrics

Shared Misclassifications: A Data-Centric Problem

harad Misclassifications: A Data-Centric Problem

- A large number of misclassified course pairs were constion to all model combinations.
 This indicates that these "hard" examples.
- u Such errors often arise from annotation artifact sherent ambiguity in the source text, or neufficient information to support a clea



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.

Mark S. Kim (SESU)

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].





July 8, 2025

oot Cause Analysis: Common Error Patterns Automating Course Articulation Qualitative Analysis: Beyond the Metrics

Root Cause Analysis: Common Error Patterns

incomplete to provide enough testual signal for the model to establish a match.

A False Negative occurs when the system falls to

 Semantic Divergence: Officially equivalent courses are described with vasity different terminology or pedagogical focus. The model correctly assesses the texts as designing; the - Topical Owedow: Courses court the same but subject but differ critically in academic level of their position in a sequence (e.g., Physics I vs. Minimalist Descriptions: One or both course

 Vague Descriptions: Descriptions use generic language, lacking the specific detail needed for differentiation. This is a known challenge in about test semantic similarity [1]

Conclusion of Analysis: The Primary Bottleneck

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

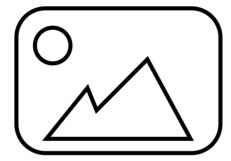
Automating Course Articulation

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

Mark S. Kim (SESU)

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]





4 D > 4 A > 4 B > 4 B > B 9 9 0

July 8, 2025

Automating Course Articulation

Qualitative Analysis: Beyond the Metrics

Conclusion of Analysis: The Primary Bottleneck

onclusion of Analysis: The Primary Bottleneck

longer the model's architecture or semantic 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, vanue language, and information gaps.
- Therefore, the most reomising path to further 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

Mark S. Kim (SESU)

- 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].

Automating Course Articulation



July 8, 2025

Automating Course Articulation

-Conclusion

Limitations

the quality and content of the public course descriptions. It cannot infer information that is also

Generalizability of the Fine-Tuned Model: The specialized headt 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-turing on local data.
- Handling of Complex Articulation Rules: The framework simplifies articulation into a binary ication of course pairs and does not natively handle complex one-to-many or many-to-man agreements, a challenge that negists for many sutremated systems [14]

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 Al Strategies

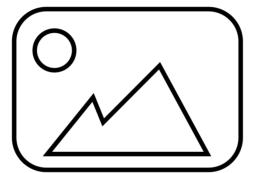
Since data quality is the primary bottleneck, future work should focus on:

- 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

Mark S. Kim (SFSU)

- 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.





Automating Course Articulation -Conclusion Future Work

The findings and limitations of this study give rise to several promising suggests for future research that 4 Developing as interactive human-in-the-loop

Investigating graph-based methods (e.g., GNNs) to identify and model complex one-to-many articulation rules.

Automating Course Articulation

July 8, 2025

Summary of Contributions

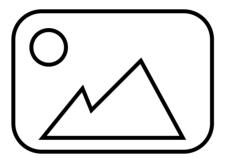
This research confronted the challenge of manual course articulation by designing, developing, and validating a novel computational framework.

Automating Course Articulation

Primary Contributions

- ◆ 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.
- 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.
- A Practical Tool for Educational Equity: We delivered a practical, computationally efficient, and privacy-preserving tool that can

Mark S. Kim (SFSU)





July 8, 2025

Automating Course Articulation

—Conclusion

Summary of Contributions

Summary of Contributions

This research confronted the challenge of manual course articulation by designing, developing, and unfoldning a noval commutational framework

calable
an end-to-end
comates course
data, achieving

Prantswark: We developed an end-to-end pipeline that successfully automates course articulation using only public data, achieving state-of-the-art accuracy.

Proof that Adaptation Outperforms Scale:
We cround this for this averaged domain.

We proved that for this specialized domain, fine-tuning a smaller model for semantic number is statistically superior to relying on sheer model scale.

A Practical Tool for Educational Equity:

We delivered a practical, computationally efficient, and privace preserving tool that can

Thank You!!!

Questions?

Automating Course Articulation $\overset{\text{\tiny{M}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}{\overset{\text{\tiny{L}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}{\overset{\text{\tiny{L}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}{\overset{\text{\tiny{L}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}{\overset{\text{\tiny{L}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}}}{\overset{\text{\tiny{L}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}{\overset{\text{\tiny{L}}}}}{\overset{\text{\tiny{L}}}{\overset{\text{\tiny{L}}}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}}}{\overset{\text{\tiny{L}}}}}{\overset{\text{\tiny{L}}}{\overset{\text{\tiny{L}}}}{\overset{\text{\tiny{L}}}}}}}}}}}}}}}}}}}}}}}}}}}}}$

Thank You!!!

Questions?

Mark S. Kim (SFSU) Automating Course Articulation

July 8, 2025

26 / 32

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 Automating Course Articulation -Wrap Up 4 mkim220mail.sfsu.edu Contact & Acknowledgments

My advisors, Professors Hui Yang, Arno Puder, an Anaetha Kulkarni, for their invaluable suidance and

- 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 SPSU Academic Technology Systems Team for their support and the use of the POLARIS. Hall-Performance Correction cluster.

Mark S. Kim (SESU)

Automating Course Articulation

July 8, 2025

References I

- Public Agenda. Beyond Transfer: Insights from a Survey of American Adults. URL: https://publicagenda.org/resource/bevond-transfer/ (visited on 06/30/2025).
- 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.
- 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.
- ASSIST. Frequently Asked Questions. URL: https://resource.assist.org/FAQ (visited on 06/30/2025).
- 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.



07.

References

Automating Course Articulation 1 Public Asserts Record Transfer Insights from a Survey of American Adults 11 -Wrap Up Akiko Aizawa. "An information-theoretic perspective of tf-idf measures". In: Information

Zaira Hassan Amur et al. "Short-Text Semantic Similarity (STSS): Techniques. Challenges an

ASSIST, Frequently Asked Questions, URL: https://resource.assist.org/FAG (visited

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

Mark S. Kim (SESU) July 8, 2025 Automating Course Articulation

References II

- 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).
- 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).
- National Student Clearinghouse. College Transfer Enrollment Grew by 5.3% in the Fall of 2023. URL: https://www.studentclearinghouse.org/news/college-transfer-enrollmentgrew-by-5-3-in-the-fall-of-2023/ (visited on 06/30/2025).
- Kevin Cook. California's Higher Education System. 2024. URL: https://www.ppic.org/publication/californias-higher-education-system/(visited on 09/06/2024).



References I Automating Course Articulation [6] California Community Colleges Chancellor's Office, Management Information Systems Data M. -Wrap Up

References

California State University Office of the Chancellor, Enrollment, 2024, URL: https://

National Student Clearinghouse. College Transfer Enrollment Grew by 5.3% in the Fall of 202

I Keein Cook California's Higher Education System 2024 1381-

Mark S. Kim (SESU) July 8, 2025 Automating Course Articulation

References III

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

Automating Course Articulation [10] Jacob Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019, arXiv: 1810.04805 Fcs.CLT, URL: -Wrap Up



M. Cabrielle Cauthier-malancon et al. "Azimuth: Sustamatic From Anabaia for Text Classification

31 Alexander Hermans, Lucas Bever, and Bastian Leibe. In Defense of the Triplet Loss for Person

14 Z. A Pardos, H Chau, and H Zhao, "Data-assistive course-to-course articulation using machine translation" In Proposedings of the Sixth Conference on Learning Scale 2019 on 1-10

SI Zachary Pardos, Hung Chau, and Haocheng Zhao, "Data-Assistive Course-to-Course Articulation Heine Machine Translation". In: (July 2010). p.cc. 10, 1145/3330430, 3333432

Mark S. Kim (SESU) Automating Course Articulation July 8, 2025

4 ロ ト 4 倒 ト 4 三 ト 4 三 ・ り Q ()・

References

References IV

Mark S. Kim (SESU)

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

Automating Course Articulation



Automating Course Articulation 61 Zarhani A. Pantos. Ziban Fan. and Weilie. East: "Connectionist recommendation in the wildthe utility and arrutability of neural naturals for paramalized course suidence" in There -Wrap Up

1 Stephen Porter, "Assessing Transfer and Native Student Performance at Four-Year Institution

II Nils Reimen, and Irvna Gurevych. "Sentence-BERT: Sentence Embeddings using Sigmese REPT Nationals, In Committee of the 2010 Conference on Empirical Mathematics In Mate

SI Florian Schroff, Dmitry Kalenichenko, and James Philbin. "FaceNet: A Unified Embedding for Face Recognition and Chatering". In: Proceedings of the IEEE Conference on Computer Vision

References

July 8, 2025

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.
- 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.



└─Wrap Up

Automating Course Articulation

References

 [21] The National Task Force on the Transfer and Award of Credit. Reimagning Transfer for Studen Success. Report to Compressional Requesters. American Council on Education, Mar. 2020. URL: https://www.goo.gov/products/goo-17-254.
 [22] United States Government Accountability Office. Higher Education: Students Need Move Information in Add Reduce Conference on Technologies (Alless Conference on Conference on Additional Conference on Conf

Sharen Slade and Paul Friedos. "Learning Analytics: Edical bases and Dilemena". In: Americal Biochemist Sci. 20 (2018).
 pp. 1550-1520. no. 10. 1177/000276-05.
 hatps://doi.org/10.1177/000276-03.13473-05.
 hatps://doi.org/10.1177/

[22] United States Government Accountability Office. Higher Education: Students Need More Information to Help Reduce Challenges in Transferring College Credits. Report to Congression. Requesters (AnI-T-SFA: United States Government Accountability Office, Aug. 14, 2017. URL https://www.pap.gor/products/gpa-17-574.

Mark S. Kim (SESU) Automating Course Articulation July 8, 2025 32/3