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 9, 2025

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

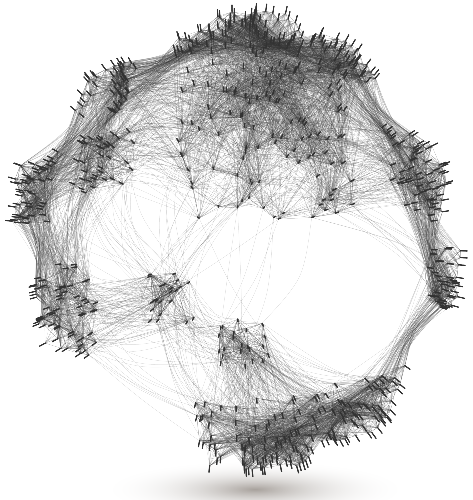
Mark S. Kim

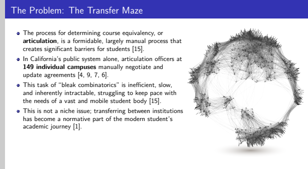
San Francisco State University
Department of Data Science and Artificial Intelligence

July 9, 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].



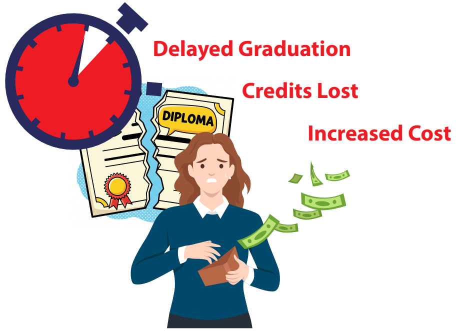


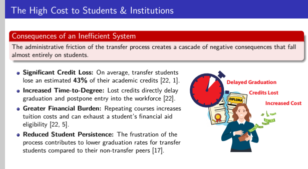
The High Cost to Students & Institutions

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 academic credits [22, 1].
- **Increased Time-to-Degree:** Lost credits directly delay graduation and postpone entry into the workforce [22].
- **Greater Financial Burden:** Repeating courses increases tuition costs and can exhaust a student's financial aid eligibility [22, 5].
- **Reduced Student Persistence:** The frustration of the process contributes to lower graduation rates for transfer students compared to their non-transfer peers [17].





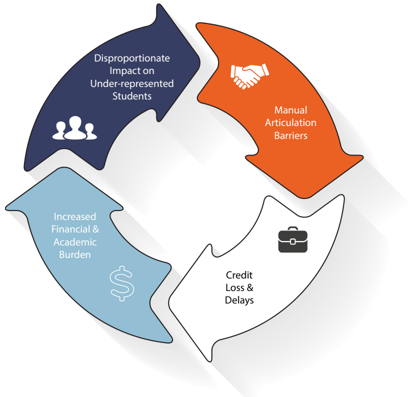
1. This high rate of credit loss often forces students to repeat courses for which they have already received a passing grade.
2. This also increases their overall time in the educational system.
3. This means a process often undertaken to save money can paradoxically result in a greater overall financial commitment.
4. The frustration also has a measurable impact on student morale.

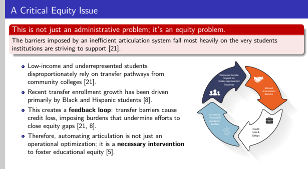
A Critical Equity Issue

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 and underrepresented students disproportionately rely on transfer pathways from community colleges [21].
- Recent transfer enrollment growth has been driven primarily by Black and Hispanic students [8].
- This creates a **feedback loop**: transfer barriers cause credit loss, imposing burdens that undermine efforts to close equity gaps [21, 8].
- Therefore, automating articulation is not just an operational optimization; it is a **necessary intervention** to foster educational equity [5].





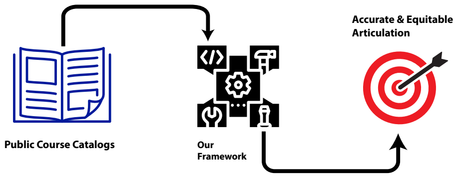
1. These are the very student populations that institutions are striving to support, making transfer efficiency a critical equity lever.
2. The National Student Clearinghouse reported this trend in Fall 2023, highlighting the increasing diversity of the transfer population.
3. This troubling loop means the manual system directly counteracts institutional goals of supporting underrepresented students.

The Goal & My Contribution

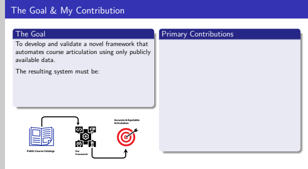
The Goal

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

The resulting system must be:



Primary Contributions



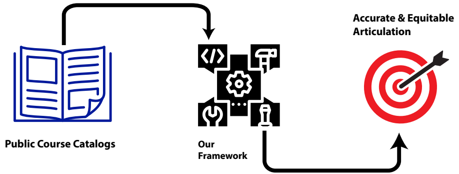
1. For our framework, we achieved F1-scores exceeding 0.97 on the held-out test set.
2. The composite vector combines the element-wise difference of course embeddings with their cosine similarity, which was shown to be a superior feature set in ablation studies.
3. Specifically, our approach avoids using sensitive student enrollment data and the high computational cost and opacity of direct LLM classification.

The Goal

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

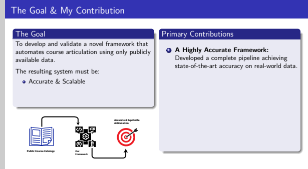
The resulting system must be:

- Accurate & Scalable



Primary Contributions

1. **A Highly Accurate Framework:** Developed a complete pipeline achieving state-of-the-art accuracy on real-world data.



1. For our framework, we achieved F1-scores exceeding 0.97 on the held-out test set.
2. The composite vector combines the element-wise difference of course embeddings with their cosine similarity, which was shown to be a superior feature set in ablation studies.
3. Specifically, our approach avoids using sensitive student enrollment data and the high computational cost and opacity of direct LLM classification.

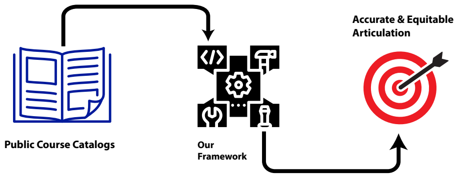
The Goal & My Contribution

The Goal

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

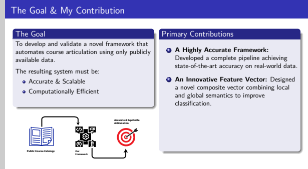
The resulting system must be:

- Accurate & Scalable
- Computationally Efficient



Primary Contributions

1. **A Highly Accurate Framework:** Developed a complete pipeline achieving state-of-the-art accuracy on real-world data.
2. **An Innovative Feature Vector:** Designed a novel composite vector combining local and global semantics to improve classification.



1. For our framework, we achieved F1-scores exceeding 0.97 on the held-out test set.
2. The composite vector combines the element-wise difference of course embeddings with their cosine similarity, which was shown to be a superior feature set in ablation studies.
3. Specifically, our approach avoids using sensitive student enrollment data and the high computational cost and opacity of direct LLM classification.

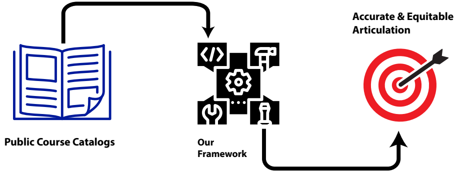
The Goal & My Contribution

The Goal

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

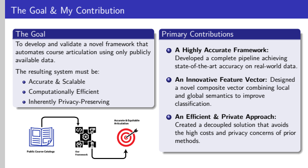
The resulting system must be:

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



Primary Contributions

1. **A Highly Accurate Framework:** Developed a complete pipeline achieving state-of-the-art accuracy on real-world data.
2. **An Innovative Feature Vector:** Designed a novel composite vector combining local and global semantics to improve classification.
3. **An Efficient & Private Approach:** Created a decoupled solution that avoids the high costs and privacy concerns of prior methods.



1. For our framework, we achieved F1-scores exceeding 0.97 on the held-out test set.
2. The composite vector combines the element-wise difference of course embeddings with their cosine similarity, which was shown to be a superior feature set in ablation studies.
3. Specifically, our approach avoids using sensitive student enrollment data and the high computational cost and opacity of direct LLM 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.

Approach	Key Characteristic	Core Limitation
Keyword & Statistical (TF-IDF)	Weight terms based on statistical importance [2].	No semantic understanding; cannot grasp synonyms or context.
Static Embeddings (word2vec, GloVe)	Represent words as averaged, pre-trained vectors.	Context-insensitive, and averaging vectors loses critical semantic information.
Enrollment-Based (course2vec)	Learn similarity from student co-enrollment patterns [16].	Requires sensitive student data, raising major privacy and generalizability issues [20].
Direct LLM Classification	Use a large language model as an end-to-end classifier.	High computational cost, opaque "black box" 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.
Approach | Key Characteristic | Core Limitation
Keyword & Statistical (TF-IDF) | Weight terms based on statistical importance [2]. | No semantic understanding; cannot grasp synonyms or context.
Static Embeddings (word2vec, GloVe) | Represent words as averaged, pre-trained vectors. | Context-insensitive, and averaging vectors loses critical semantic information.
Enrollment-Based (course2vec) | Learn similarity from student co-enrollment patterns [16]. | Requires sensitive student data, raising major privacy and generalizability issues [20].
Direct LLM Classification | Use a large language model as an end-to-end classifier. | High computational cost, opaque "black box" reasoning, and sensitive to prompt phrasing [11].

This slide provides the full context for our work. The key takeaway is that each prior method had a significant drawback, whether it was a lack of semantic understanding, a reliance on private data, or operational complexity. This creates a clear research gap that our framework is designed to fill.

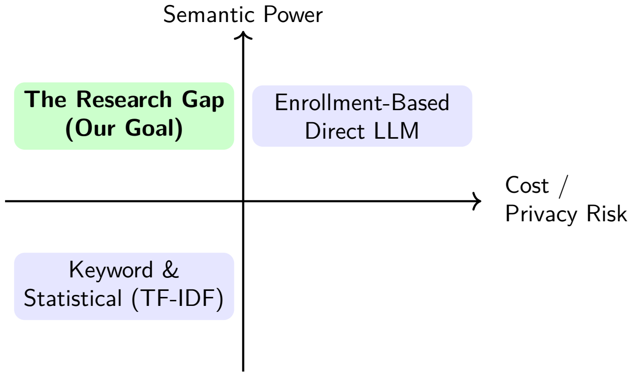
The Research Gap

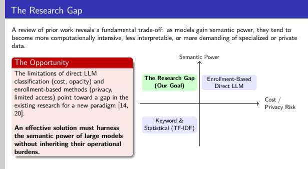
A review of prior work reveals a fundamental trade-off: as models gain semantic power, they tend to become more computationally intensive, less interpretable, or more demanding of specialized or private data.

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.





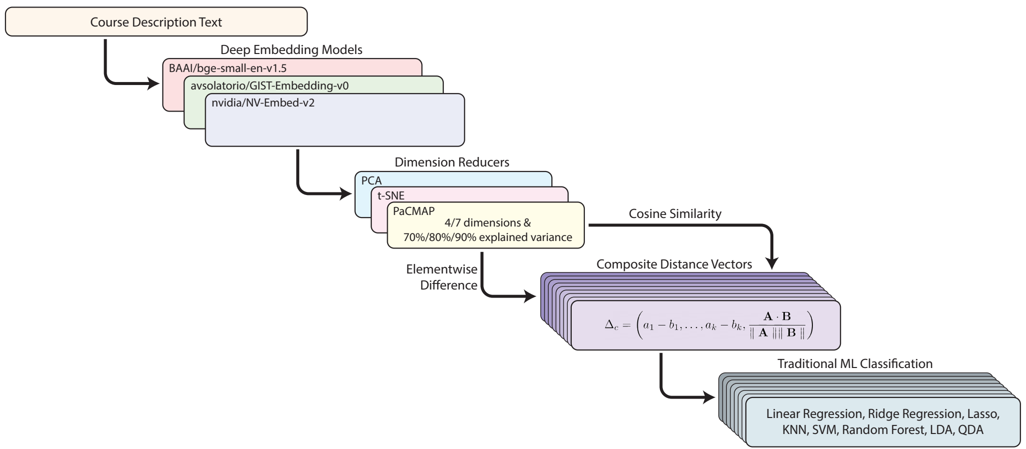
This diagram clearly shows the trade-offs. In the bottom-left, we have simple methods with low semantic power. In the top-right, we have powerful methods that are either expensive or raise privacy concerns. The top-left quadrant is where we want to be: high semantic power, but with low cost and no privacy risk. This is the gap our research addresses.

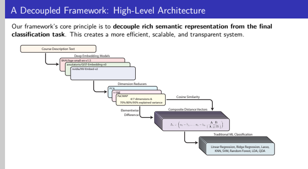
Agenda

Agenda
Introduction
Background & Related Work
A Decoupled Framework for Articulation
Experimental Setup & Results
Qualitative Analysis: Beyond the Metrics
Conclusion
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.





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].

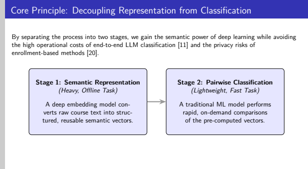
Stage 1: Semantic Representation
(Heavy, Offline Task)

A deep embedding model converts raw course text into structured, reusable semantic vectors.

→

Stage 2: Pairwise Classification
(Lightweight, Fast Task)

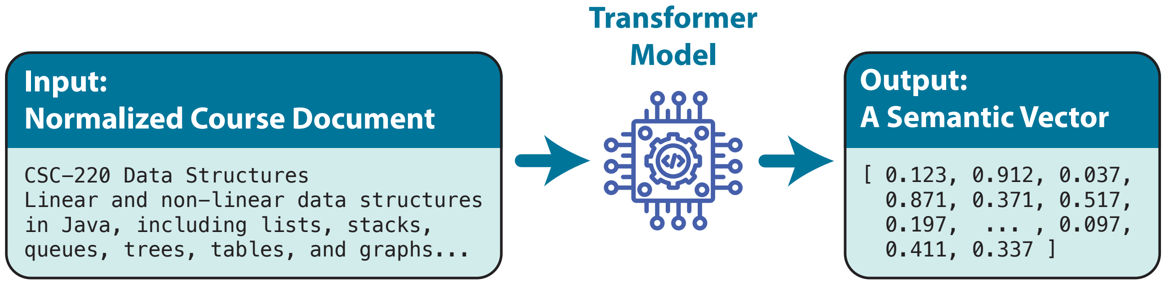
A traditional ML model performs rapid, on-demand comparisons of the pre-computed vectors.

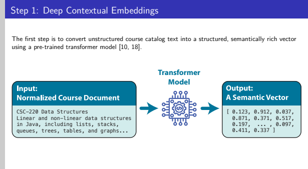


This is the most important concept in the framework. We do the heavy lifting of understanding language only once, offline. This lets us do the actual classification of pairs incredibly quickly and efficiently. This two-stage process is what gives us the best of both worlds: the semantic power of deep learning without the high operational costs for every comparison, which is what makes the system so scalable and practical.

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].





The key here is that we're turning unstructured text into a structured, mathematical object—a vector. We use a pre-trained transformer model for this, specifically models from the Sentence-BERT family [10, 18]. The resulting vector captures the actual meaning of the course, so similar courses will have vectors that are closer together in this high-dimensional space.

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

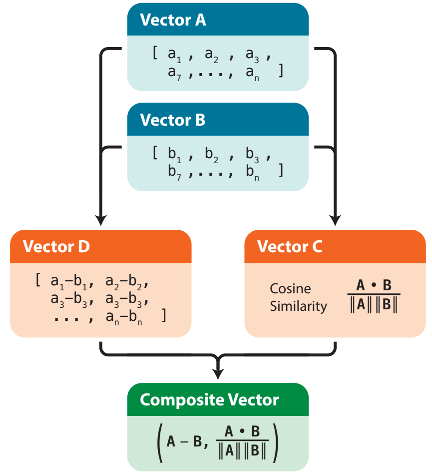
The vector combines two distinct types of information:

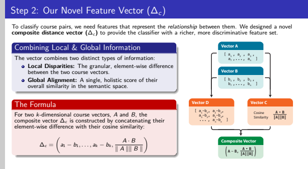
- **Local Disparities:** The granular, element-wise difference between the two course vectors.
- **Global Alignment:** A single, holistic score of their overall similarity in the semantic space.

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**.

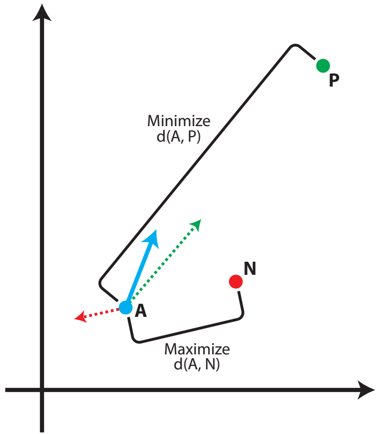
Learning Objective: The Triplet Loss

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):

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



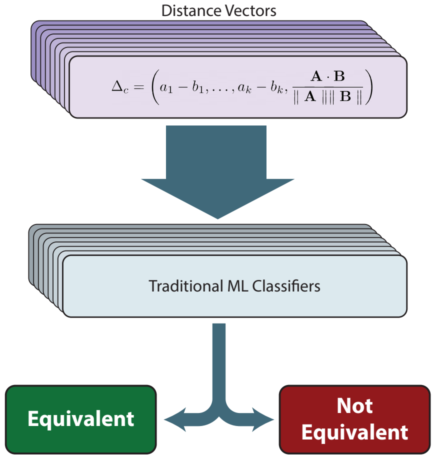
Step 4: Downstream Classification

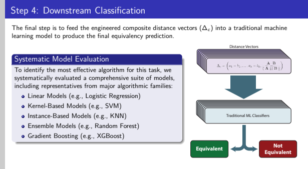
The final step is to feed the engineered composite distance vectors (Δ_c) into a traditional machine learning model to produce the final equivalency prediction.

Systematic Model Evaluation

To identify the most effective algorithm for this task, we systematically evaluated a comprehensive suite of models, including representatives from major algorithmic families:

- Linear Models (e.g., Logistic Regression)
- Kernel-Based Models (e.g., SVM)
- Instance-Based Models (e.g., KNN)
- Ensemble Models (e.g., Random Forest)
- Gradient Boosting (e.g., XGBoost)





1. The goal here was to find the best possible classifier for our specific feature vectors. By testing models from different families, we could probe the data for things like linear separability, complex non-linear decision boundaries, and feature interactions.
2. The graphic on the right shows the simple version of this process: our feature vector goes in, the classifier makes a judgment, and a binary prediction comes out.
3. The specific results of this competitive evaluation—that is, which models performed best—will be discussed in the upcoming Results section.

Agenda

Datasets & Evaluation

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

Characteristic	**Initial Dataset**	**PPM Corpus**
Source	Manually curated via ASSIST	Program Pathways Mapper (PPM)
Purpose	Preliminary screening, prototyping, and initial classifier evaluation	Definitive fine-tuning and final pipeline evaluation
Ground Truth	Established articulation agreements	Course Identification Number (C-ID)
Final Size	400 course pairs (for evaluation set)	2,157 courses (across 157 classes)
Partitioning	Stratified random sample	Stratified 50/50 train/test split

Datasets & Evaluation

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

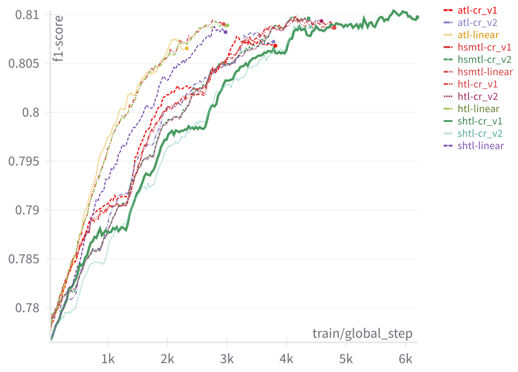
Characteristic	Initial Dataset	PPM Corpus
Source	Manually curated via ASSIST	Program Pathways Mapper (PPM)
Purpose	Preliminary screening, prototyping, and initial classifier evaluation	Definitive fine-tuning and final pipeline evaluation
Ground Truth	Established articulation agreements	Course Identification Number (C-ID)
Final Size	400 course pairs (for evaluation set)	2,157 courses (across 157 classes)
Partitioning	Stratified random sample	Stratified 50/50 train/test split

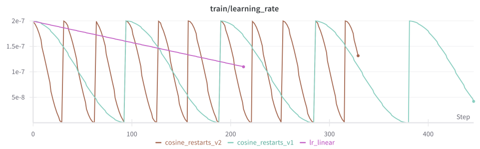
Adapting a Model with Domain-Specific Fine-Tuning

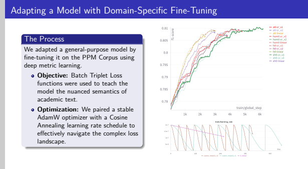
The Process

We adapted a general-purpose model by fine-tuning it on the PPM Corpus using deep metric learning.

- **Objective:** Batch Triplet Loss functions were used to teach the model the nuanced semantics of academic text.
- **Optimization:** We paired a stable AdamW optimizer with a Cosine Annealing learning rate schedule to effectively navigate the complex loss landscape.







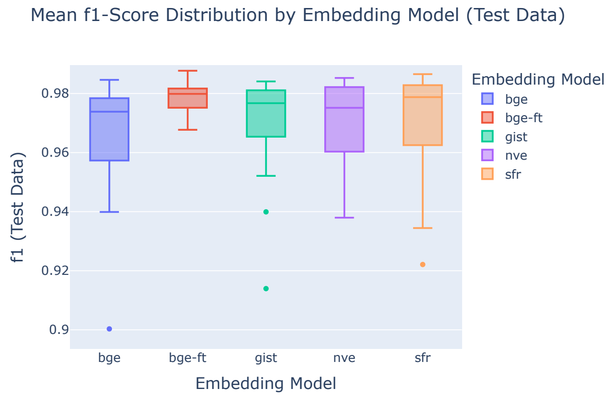
1. General-purpose models often miss the subtle but critical differences in academic language, so we needed to create a specialist.
2. We used a Triplet Loss objective, which teaches the model to produce more discriminative embeddings for our specific task.
3. The graph on the left shows the F1-score on a validation set improving as the model learns. The graph on the right shows the Cosine Annealing learning rate schedule we used, which helps the optimizer settle into a high-quality solution.

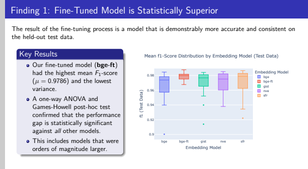
Finding 1: Fine-Tuned Model is Statistically Superior

The result of the fine-tuning process is a model that is demonstrably more accurate and consistent on the held-out test data.

Key Results

- Our fine-tuned model (**bge-ft**) had the highest mean F_1-score ($\mu = 0.9786$) and the lowest variance.
- A one-way ANOVA and Games-Howell post-hoc test confirmed that the performance gap is statistically significant against *all* other models.
- This includes models that were orders of magnitude larger.





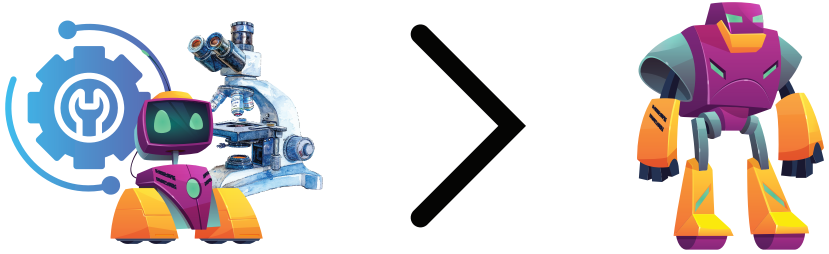
1. This bar chart is the punchline. It compares the final F1 scores of our fine-tuned model, ‘bge-ft‘, against its base version and other, much larger off-the-shelf models.
2. As you can see, our model is the clear winner in terms of performance and consistency.
3. This isn’t just a small improvement; formal statistical tests (a one-way ANOVA and a Games-Howell post-hoc test) confirmed that this performance gap is statistically significant.

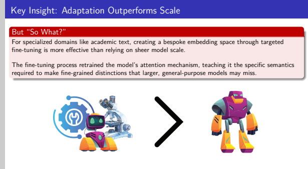
Key Insight: Adaptation Outperforms Scale

But “So What?”

For specialized domains like academic text, creating a bespoke embedding space through targeted fine-tuning is more effective than relying on sheer model scale.

The fine-tuning process retrained the model’s attention mechanism, teaching it the specific semantics required to make fine-grained distinctions that larger, general-purpose models may miss.





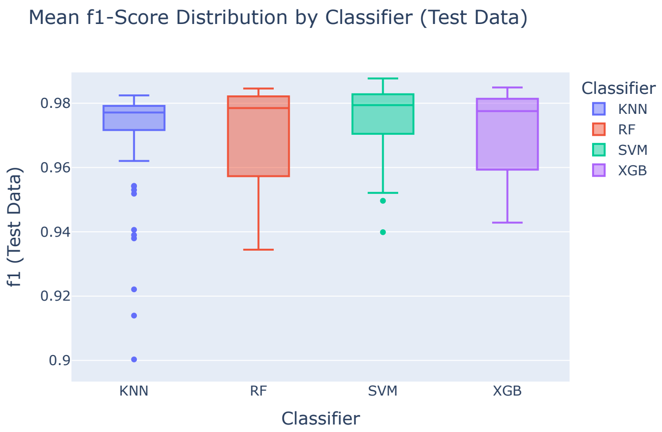
1. This is the main takeaway from this finding. It answers the "so what?" question.
2. A general-purpose model, even a huge one, might not understand the subtle but critical difference between a "survey" course, an "introductory" course, and a "foundations" course. Our model does.
3. This is a key insight not just for this project, but for anyone working on NLP problems in specialized domains: targeted adaptation is crucial.

Finding 2: All Finalists Perform Exceptionally Well

The first key result from the classifier evaluation is that our feature engineering and fine-tuning were highly effective, leading to exceptional performance across all finalist models.

High & Stable Performance

- All four finalist classifiers (SVM, RF, XGBoost, KNN) achieved high and stable F1-scores.
- As the boxplot shows, the distributions are tightly clustered with mean scores approaching or exceeding **0.97**.
- This demonstrates the robustness of the upstream feature representation.



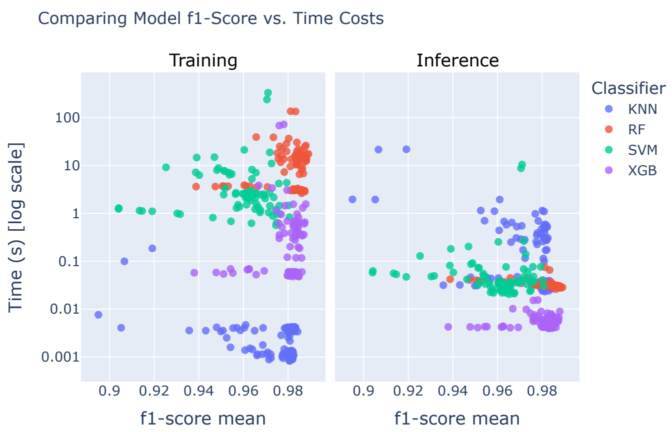
1. The main message here is that our framework is successful. The features we created are so discriminative that any of the top-tier classifiers we chose for the final stage did an excellent job.
2. This boxplot visualizes the F1-score distributions from our cross-validation on the test set. You can see how tightly grouped they are, all centered well above 0.97.
3. This high baseline performance is a testament to the quality of the fine-tuned embeddings and the composite distance vector.

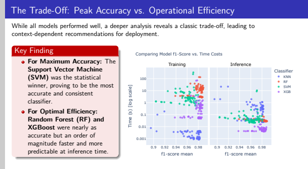
The Trade-Off: Peak Accuracy vs. Operational Efficiency

While all models performed well, a deeper analysis reveals a classic trade-off, leading to context-dependent recommendations for deployment.

Key Finding

- **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.





1. So, if all the models are good, which one should we choose? This scatterplot answers that question by showing accuracy versus inference speed.
2. On the right, you can see that RF and XGBoost are incredibly fast, which is critical for a scalable, real-world system with many users.
3. However, SVM, while slower, is statistically the most accurate. The choice depends on the deployment context: if you need the absolute highest accuracy, you choose SVM. If you need speed and scalability, you choose Random Forest or XGBoost.

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

Automating Course Articulation
└─Qualitative Analysis: Beyond the Metrics
└─Agenda

2025-07-09

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

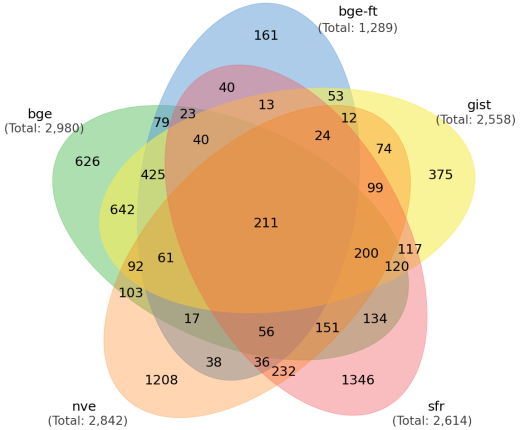
Qualitative Analysis: Why Do Errors Still Occur?

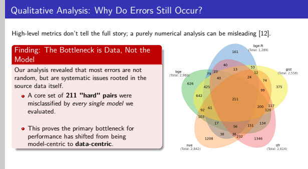
High-level metrics don't tell the full story; a purely numerical analysis can be misleading [12].

Finding: The Bottleneck is Data, Not the Model

Our analysis revealed that most errors are not random, but are systematic issues rooted in the source data itself.

- A core set of **211 "hard" pairs** were misclassified by *every single model* we evaluated.
- This proves the primary bottleneck for performance has shifted from being model-centric to **data-centric**.





1. The high degree of error overlap shown in the Venn diagram is critical. It shows that diverse models—from small specialists to large generalists—all failed on the exact same set of "hard" examples.
2. This tells us the problem isn't idiosyncratic model weaknesses. The errors are systematic, pointing to inherent challenges in the source data itself.
3. Therefore, the limiting factor is no longer model sophistication. The model often fails because the source data is ambiguous, inconsistent, or lacks a clear textual signal to support the ground-truth label.

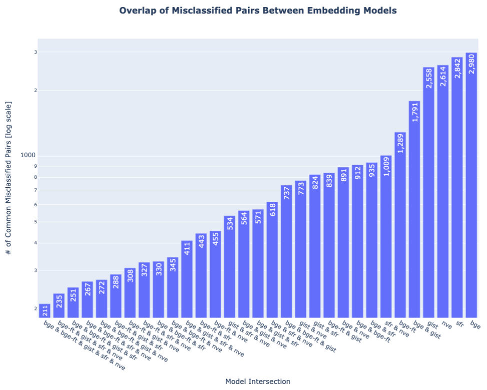
Shared Misclassifications: A Data-Centric Problem

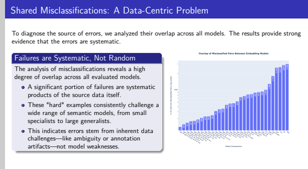
To diagnose the source of errors, we analyzed their overlap across all models. The results provide strong evidence that the errors are systematic.

Failures are Systematic, Not Random

The analysis of misclassifications reveals a high degree of overlap across all evaluated models.

- A significant portion of failures are systematic products of the source data itself.
- These "hard" examples consistently challenge a wide range of semantic models, from small specialists to large generalists.
- This indicates errors stem from inherent data challenges—like ambiguity or annotation artifacts—not model weaknesses.





1. This slide digs deeper into the "why" behind the errors. The key takeaway is that the models are largely agreeing with each other about which course pairs are difficult to classify.
2. The bar chart on the right reinforces this, showing high counts of shared misclassified pairs between all combinations of the models we tested.
3. This is strong evidence that the problem isn't the model's fault. In many cases, the models are correctly reporting that two course descriptions are not semantically similar. The issue is that the ground-truth label says they *should* be equivalent, pointing to an inconsistency in the source data itself.

Root Cause Analysis: False Negatives (Missed Equivalencies)

A False Negative occurs when the system fails to identify a true, existing equivalence. This represents a missed opportunity for a student.

Causes

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

Example: Semantic Divergence

Course A	*"... examines... developmental milestones from middle childhood through adolescence... "*
Course B	*"... examines... diversity and inclusion... anti-bias curriculum... promote inclusive... classroom... "*

Result: False Negative

(Model correctly sees texts as different)

1. The first major error class is the False Negative, where our system misses a real equivalency. This could cause a student to unnecessarily retake a course.
2. The example on the right is the most interesting cause. These two courses are both officially labeled with the same C-ID, making them equivalent. However, one is described using traditional developmental psychology language, while the other uses the language of social justice pedagogy.
3. Our model correctly concludes that these two texts are not semantically similar. The failure here isn't in the model's understanding; it's in the inconsistency of the source data provided by the institutions.

Root Cause Analysis: False Positives (Incorrect Matches)

A False Positive occurs when the system incorrectly classifies two non-equivalent courses as equivalent. This is a harmful error that could mislead a student.

Causes

- **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). The model correctly identifies high topical similarity but can't infer the sequence.
- **Vague Descriptions:** Descriptions use generic language, lacking the specific detail needed for differentiation. This is a known challenge in short-text semantic similarity [3].

Example: Topical Overlap

Course A	PHYS-4A: *"...a systematic introduction to the principles of classical mechanics..."*
Course B	PHYS-4D: *"...a systematic introduction to the principles of modern physics..."*

Result: False Positive

(Model sees high topical overlap)

Root Cause Analysis: False Positives (Incorrect Matches)

1. The other major error class is the False Positive, where we incorrectly approve a pair. This is dangerous because it could lead a student to take a non-transferable course.
2. The most common cause is topical overlap. For example, two physics courses at the same college. They are sequential, not equivalent.
3. Because their descriptions both contain phrases like "systematic introduction to the principles of," the model sees high semantic overlap and makes an incorrect match. It cannot infer the sequential relationship from the text alone.

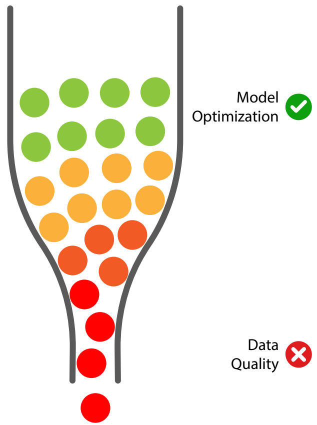
Conclusion of Analysis: The Primary Bottleneck

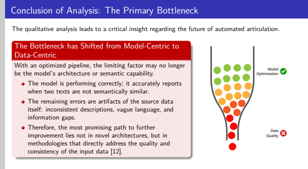
The qualitative analysis leads to a critical insight regarding the future of automated articulation.

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

With an optimized pipeline, the limiting factor may no longer be 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].





1. This slide presents the single most important conclusion from the entire qualitative analysis. After digging into why the model still makes mistakes, we arrived at this critical insight.
2. To be clear on what this means: when the model is given two course descriptions that are written very differently, it correctly reports that they are not a good semantic match. The 'error' isn't in the model's logic, but in the ground-truth label that says they *should* be equivalent despite the textual evidence.
3. Therefore, the problem has shifted. We've pushed the model's performance about as far as it can go with the current data. Simply using a bigger or more complex model is unlikely to resolve these data-inherent issues.
4. This insight directly informs the future work for this project. The most promising path to further improvement lies not in novel model architectures, but in data-centric AI methodologies that focus on the quality and consistency of the input data.

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

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].

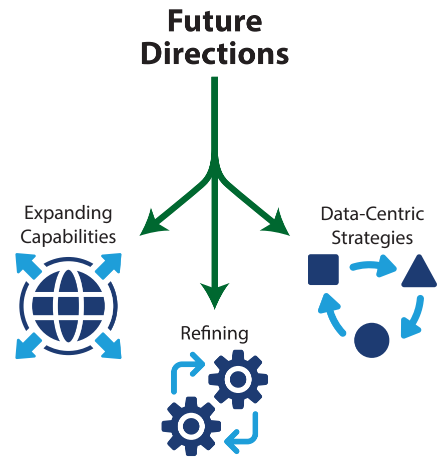
Limitations
While the proposed framework represents a significant advance, it is essential to acknowledge the boundaries of the current study.
Key Limitations
Performance is Capped by Data Quality: The system's performance is fundamentally limited by the quality and content of the public course descriptions. 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].

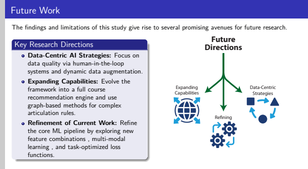
Future Work

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

Key Research Directions

- **Data-Centric AI Strategies:** Focus on data quality via human-in-the-loop systems and dynamic data augmentation.
- **Expanding Capabilities:** Evolve the framework into a full course recommendation engine and use graph-based methods for complex articulation rules.
- **Refinement of Current Work:** Refine the core ML pipeline by exploring new feature combinations , multi-modal learning , and task-optimized loss functions.





1. Given that data quality is the primary bottleneck , the most critical future work involves an interactive, human-in-the-loop system where an expert can review ambiguous pairs flagged by the model. We could also explore automatically requesting a more detailed syllabus if confidence is low.
2. Active development is already underway to expand the framework into a full-scale course recommendation engine , which will eventually have a conversational interface. To handle complex one-to-many rules, another path is to model curricula as graphs and apply graph neural networks.
3. Finally, there are opportunities to refine the core pipeline by investigating alternative composite distance measures , extending the model to analyze not just catalog descriptions but also the full text of syllabi or textbook lists , and exploring more sophisticated training methods like instruction-tuning.

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 reduce administrative burden and help mitigate the systemic inequities faced by transfer students.

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 reduce administrative burden and help mitigate the systemic inequities faced by transfer students.

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 III

[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 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.