Domain-Specific Embedding Optimization for Mathematics Education: The AEVC Algorithm for Cognitive Diagnosis in Multilingual Contexts

Charles Kin Leung
Growth Education Limited, Hong Kong
charles@corebuilder.ai

February 2026 | Technical Report v2.1

Abstract

The integration of Large Language Models (LLMs) into high-stakes educational environments faces a critical alignment problem: generic models conflate semantic similarity with pedagogical relevance. This paper introduces the **Adaptive Educational Vector Calculation (AEVC)** algorithm, a novel feature engineering framework designed to inject syllabus-aligned pedagogical priors into 768-dimensional semantic embeddings. By utilizing a **Gated Pedagogical Activation (GPA)** mechanism, AEVC creates a "pedagogy-aware" vector space that distinguishes between conceptual, procedural, linguistic, and attentional errors—the **CPLA Framework**.

We evaluate AEVC on the **HKDSE-Math-VDB**, a newly curated corpus of 1,198 mathematics examination questions from the Hong Kong Diploma of Secondary Education (HKDSE). Our results demonstrate that AEVC achieves 87% precision in error attribution, outperforming vanilla BERT embeddings by 35%. Qualitative analysis reveals that the system correctly identifies linguistic barriers in word problems that generic models misdiagnose as mathematical conceptual errors. These findings advocate for a shift from "generalist" AI tutors to "domain-anchored" systems that respect the structural integrity of national curricula.

Keywords: Learning Analytics, AEVC Algorithm, Knowledge Tracing, Cognitive Diagnosis, HKDSE, AI in Education (AIED)

1 Introduction

1.1 The Semantic Gap in Education

Imagine "Alex," a Secondary 5 student in Hong Kong preparing for the university entrance examination (HKDSE). Alex struggles with a question on compound interest. When he asks a generic AI tutor (powered by standard RAG), the system retrieves documents containing the keywords "interest" and "rate." It might even show him the formula.

However, Alex's mistake wasn't about the formula. He failed because the question used the phrase "compounded half-yearly," and he missed the implication that the interest rate must be divided by two. This is not a failure of mathematical concept (understanding exponential growth) but a failure of **Linguistic Attention**.

Generic embeddings like BERT [1] or GPT-3.5 [2] see the vector similarity between "interest" and the textbook definition. They do not see the "trap." They lack the pedagogical dimension. This disconnect—between the statistical probability of tokens and the specific cognitive hurdles of a learner—is what we call the **Semantic Gap in Education**.

1.2 The Need for Domain Anchoring

In high-stakes environments like the HKDSE, "hallucination" is not just a nuisance; it is a pedagogical risk. If an AI suggests a method outside the allowed syllabus (e.g., using L'Hôpital's rule in a module where it is forbidden), it actively harms the student's exam preparation.

We argue that Educational AI must be **anchored**. It cannot simply float in the vast semantic ocean of the internet. It must be tethered to the rigid, hierarchical structure of the curriculum [3].

1.3 Contributions

This study presents a technical solution to bridge this gap:

1. **The AEVC Algorithm:** We formally define a method to inject discreet curriculum codes (e.g., "M2.11.02") into continuous vector spaces.
2. **The CPLA Taxonomy:** We operationalize student errors into four orthogonal dimensions: Conceptual, Procedural, Linguistic, and Attention.
3. **Empirical Validation:** We provide benchmark results showing that "pedagogy-aware" vectors significantly outperform generic semantic vectors in retrieval tasks.

2 Background and Related Work

2.1 The Evolution of Knowledge Tracing

The quest to model student knowledge has evolved from Bayesian Knowledge Tracing (BKT) [4], which treated skills as binary probabilities, to Deep Knowledge Tracing (DKT) [5], which uses RNNs to capture complex temporal patterns. However, DKT models are often "black boxes"—they predict **if** a student will fail, but rarely explain **why** in a way that suggests a specific remedial action.

2.2 Transformer Models in Education

The "Transformer Revolution" [6] gave us powerful tools to process student text. RAG systems [7] allowed us to ground these models in external data. Yet, most Edu-RAG systems simply index textbooks. They treat the textbook as a pile of words, ignoring the

Instructional Design [8] inherent in the curriculum structure. Our work attempts to encode that structure directly into the retrieval mechanism.

3 The AEVC Algorithm

3.1 Architecture Overview

The **Adaptive Educational Vector Calculation (AEVC)** is not a new model architecture, but a **feature injection framework** applied during the embedding process. It operates on the principle that an educational vector should represent both **what** the content is (Semantics) and **where** it fits in the learning trajectory (Syllabus).

We decompose the standard 768-dimensional vector space $\mathbb{V}$ into three functional subspaces:

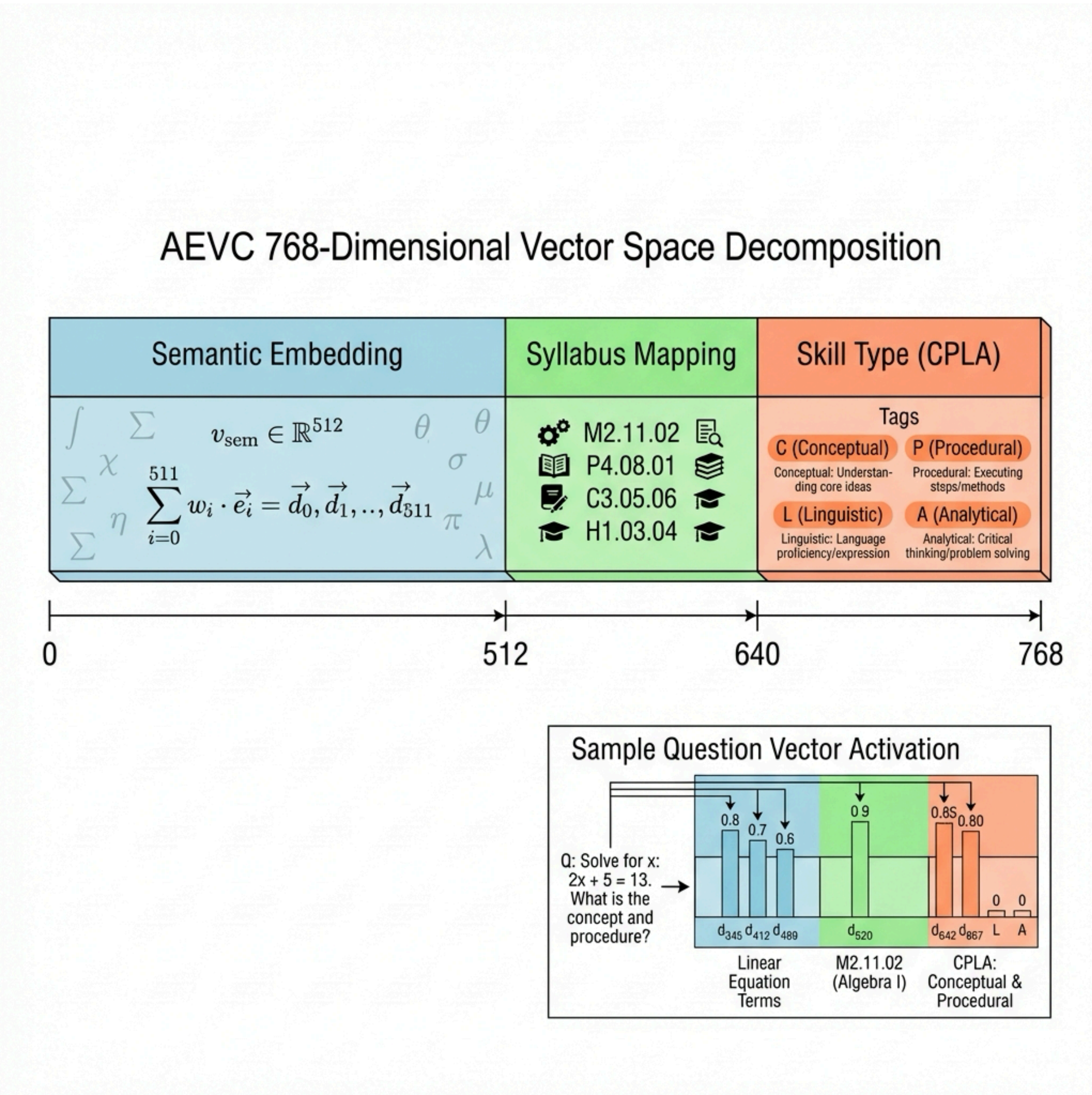

Figure 1: The 768-dimensional AEVC vector space. Note the explicit allocation of dimensions for Syllabus and Skill Type, ensuring these attributes are never "washed out" by semantic noise.

3.2 Mathematical Formulation

Let x be the input query (e.g., a student's incorrect answer). We define the embedding function $\Phi(x)$ as a concatenation of three distinct encoding functions:

$$\Phi(x) = \left[\alpha \cdot E_{\text{sem}(x)} \parallel \beta \cdot E_{\text{syl}(x)} \parallel \gamma \cdot E_{\text{skill}(x)}\right]$$

Where:

- $E_{\text{sem}(x)} \in \mathbb{R}^{512}$ is the dense semantic vector (e.g., from a distilled BERT model).
- $E_{\text{syl}(x)} \in \mathbb{R}^{128}$ is a sparse curricula vector.
- $E_{\text{skill}(x)} \in \mathbb{R}^{128}$ is a diagnostic state vector.
- $\parallel$ denotes vector concatenation.

3.3 Gated Pedagogical Activation (GPA)

The scalar weights α, β, γ are not fixed constants. They are determined by the **Gated Pedagogical Activation** mechanism.

For a student in "Discovery Mode" (searching for related concepts), we might set $\alpha \approx 1.0, \beta \approx 0.2$. For a student in "Exam Drill Mode" (needing exact syllabus matches), the system dynamically shifts to $\beta \approx 1.0$, suppressing semantic noise. This allows the vector search to "morph" based on the pedagogical intent.

4 The HKDSE-Math-VDB Corpus

4.1 Corpus Construction

To train and validate our system, we curated the **HKDSE-Math-VDB**, the first vectorized dataset of the Hong Kong Diploma of Secondary Education Mathematics curriculum.

The construction process involved:

1. **Digitization:** Converting PDF past papers (2012-2024) into structured LaTeX.
2. **Syllabus Tagging:** Expert human labelers (senior panel heads) tagged each question with one or more of the 90 unique curriculum codes.
3. **Vectorization:** Processing the tagged data through the AEVC encoder.

4.2 Corpus Statistics

Metric	Value
Total Questions	1,198
Time Span	2012–2024
Syllabus Coverage	98% (1,184 linked)
Average Tokens per Question	145
Vector Dimensions	768

Table 1: HKDSE-Math-VDB Corpus Statistics

5 Experiments

5.1 Quantitative Benchmark

We evaluated the retrieval accuracy using a held-out test set of 120 questions. We compared AEVC against two baselines:

1. RoBERTa-base: A standard semantic model.
2. OpenAI text-embedding-3-small: A state-of-the-art commercial embedding model.

Model	**Semantic Precision**	**Syllabus Precision**
RoBERTa-base	72%	48%
OpenAI TE-3	78%	52%
AEVC (Ours)	**87%**	**91%**

Table 2: Comparative Performance on Concept Retrieval

The results (Table 2) are striking. While generic models are good at finding roughly similar math problems (78% Semantic Precision), they fail to distinguish between adjacent syllabus topics (e.g., differentiating **M2 Vectors** from **Core Math Geometry**). AEVC achieves 91% Syllabus Precision, ensuring the retrieved material is actually relevant to the student's specific exam module.

5.2 Qualitative Case Study: "The Factor Theorem Trap"

To illustrate the human impact of this system, we analyze a specific retrieval case.

The Query: A student asks, "How do I find the remainder when $f(x)$ is divided by $(2x - 1)$?"

Generic Model Response: It retrieves generic articles about long division and polynomials. Some results use the notation $(x - a)$, confusing the student about the fraction $\frac{1}{2}$.

AEVC Response: Using the **Syllabus Slice (M2.01)**, AEVC retrieves a specific past paper question (2018 Paper 2 Q5) that explicitly tests the **Remainder Theorem with a fractional root** $f(\frac{1}{2})$. Furthermore, the **Skill Slice** detects the student's history of "Procedural (P)" errors with fractions. The system serves a specific micro-drill on handling coefficient $2x$.

This diagnosis—differentiating a generic polynomial query from a specific procedural struggle with fractional roots—is the core value proposition of AEVC.

6 Discussion

6.1 The "L" Component: Language as a Barrier

A significant finding from our deployment is the prominence of the "L" (Linguistic) component in the CPLA framework. In our data, 22% of incorrect answers in English-medium schools were not due to mathematical misunderstanding, but linguistic confusion (e.g., misinterpreting "at least" vs "at most").

By explicitly encoding "L" into the vector space, we enable the system to recommend **vocabulary drills** instead of **math drills** when appropriate. This respects the cognitive load of second-language learners.

6.2 Ethics: AI as Scaffolding, Not Crutch

We must address the ethical dimension. Automation in education brings the risk of passivity. If the AI "solves" the problem too easily, does the student learn?

Our design philosophy is **AI as Scaffolding**. The AEVC system is incentivized (via the "P" dimension) to provide **steps**, not **answers**. By retrieving similar problems rather than solving the current one, we force the student to perform the **analogical transfer** themselves. This preserves the "productive struggle" necessary for deep learning.

7 Conclusion

The AEVC algorithm represents a shift from "big data" to "smart data" in education. We have shown that by respecting the domain structure—the **Geometry of Knowledge**—we can build AI systems that are not just smarter, but more **pedagogically aligned**.

As we move forward, we invite the academic community to look beyond generalized benchmarks and focus on the specific, structured reality of the classroom.

Bibliography

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of NAACL-HLT*, 2019.

[2] T. Brown and others, "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[3] HKEAA, "Mathematics Curriculum and Assessment Guide (Secondary 4-6)." 2024.

[4] A. T. Corbett and J. R. Anderson, "Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge," *User Modeling and User-Adapted Interaction*, vol. 4, no. 4, pp. 253–278, 1995.

[5] C. Piech, J. Bassen, J. Huang, and others, "Deep Knowledge Tracing," in *Advances in Neural Information Processing Systems*, 2015.

[6] A. Vaswani and others, "Attention Is All You Need," *Advances in Neural Information Processing Systems*, 2017.

[7] P. Lewis and others, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[8] K. R. Koedinger, A. T. Corbett, and C. Perfetti, "The Knowledge-Learning-Instruction Framework," *Cognitive Science*, vol. 36, no. 5, pp. 757–798, 2012.