Gated Cross-Attention Matcher for Aligning Course With Relevant KUs

Aashish Pandey, Ram Dantu
Department of Computer Science and Engineering,
University of North Texas, Denton, Texas, 76207, USA

Introduction

We propose a Gated Cross-Attention Module (GCAM) for curriculum–exam alignment by learning robust semantic matching between course knowledge units (KUs) and exam micro-assessments. Raw curriculum artifacts (syllabi, outcome tables) are normalized into KUs containing titles, topics, and learning outcomes. GCAM uses a frozen LLM encoder to generate contextual token states and trains only lightweight LoRA projections plus two interpretable gates: a Token Gate to suppress irrelevant memory tokens and a Head Gate to activate the most useful attention heads per query. The model is trained with contrastive InfoNCE loss on supervised text pairs and supports retrieval-based alignment using a KU memory bank.

Problem Definition

There is a significant lack of work that needs to be done on the alignment of CAE Knowledge units (KU) and the courses taught at university level. Currently, the matching is done by instructor which is a lot of manual task from matching the KU topics and sub-topics based on the syllabus and assignments . This tool is designed to automate this manual job.

We approach this problems based on a custom module built using token gate and head gate, which we collectively call GCAM module.

Gates Workflow

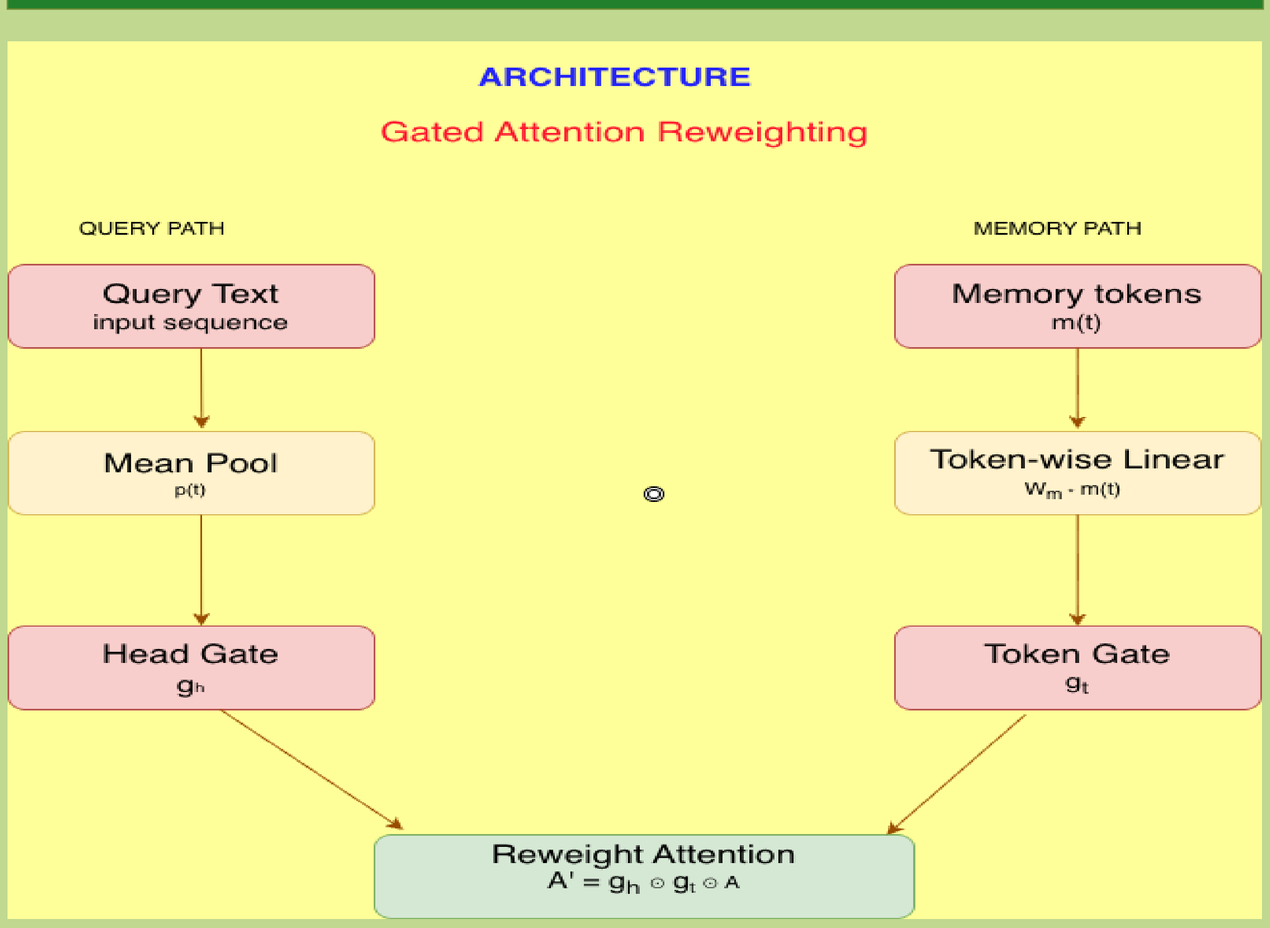

Fig 1: High-level Working of Two Gates; Token Gate and Head Gate

Course syllabi and outcome tables describe broad instructional intent, while exams evaluate narrow concepts through micro-assessments such as short-answer prompts, sub-topics, or competency checks. Traditional keyword matching and generic embedding retrieval struggle in this setting due to cross-granularity mismatch, inconsistent phrasing, and the presence of boilerplate educational language. To address these issues, this project introduces a retrieval-based alignment framework built around a lightweight Gated Cross-Attention Module (GCAM) that learns to match courses to the most relevant curriculum knowledge units (KUs) with improved robustness and interpretability.

At the model level, the system uses a frozen large language model encoder as a semantic backbone to produce contextual token representations for both queries and memory items. Rather than fine-tuning the full model, GCAM trains only lightweight parameters. LoRA-augmented projection layers and two gating mechanisms, making the approach computationally efficient and modular.

System Design

GCAM performs cross-attention between query tokens and memory tokens to produce a query-aware memory summary, then applies two gates that explicitly control information flow. The Token Gate learns to suppress irrelevant memory tokens (such as generic instructional phrasing) and emphasize concept-bearing terms, which is critical when aligning to verbose KU descriptions. The Head Gate learns to activate only the most useful attention heads for a given query, allowing the model to select the appropriate alignment behavior (lexical overlap, semantic similarity, or skill-oriented matching) depending on the query's intent. Together, these gates convert standard cross-attention into an intent-adaptive, noise-resistant matching mechanism that is well suited to educational text.

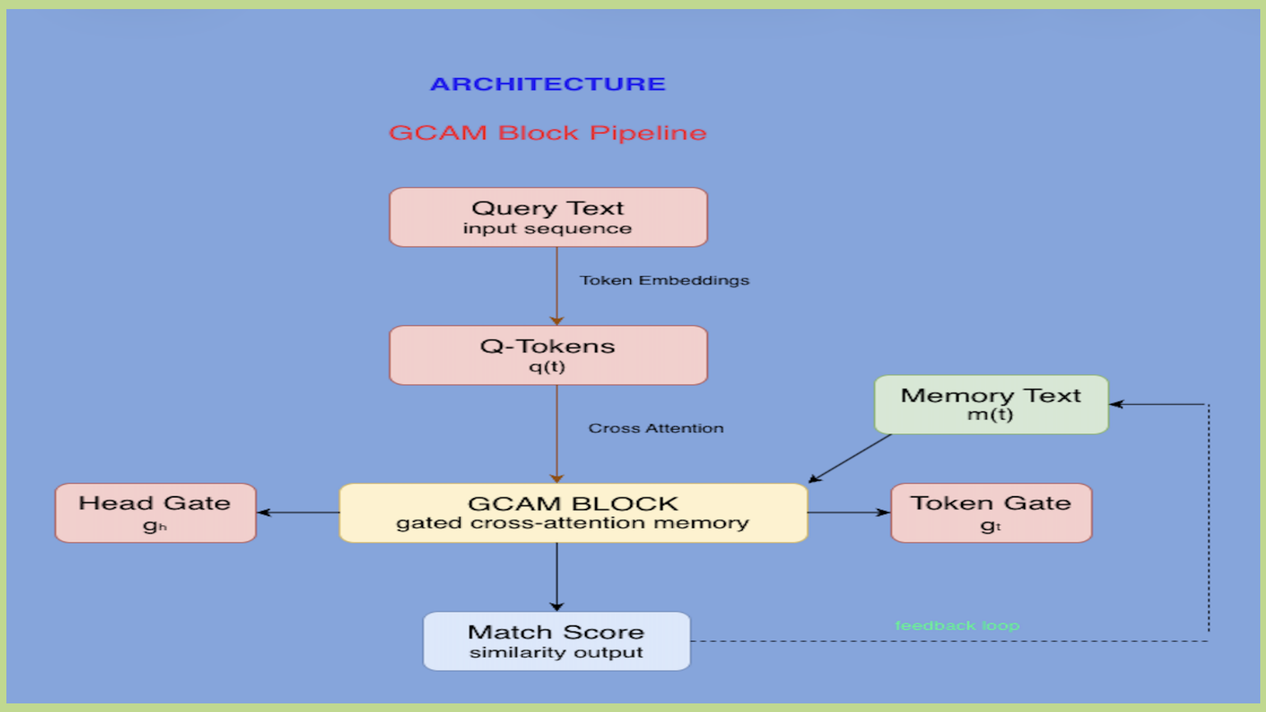

Fig 2: High-Level Working of GCAM with Two Gates

Work-Flow

The pipeline has three stages. (1) Data Layer: curriculum documents and exam specifications are parsed into normalized Knowledge Units (KUs) with structured fields: Title, Topics[], LearningOutcomes[]. (2) Training Construction: supervised text pairs are generated (e.g., Title↔Topics, Title↔LOs, Topic↔Title) and split into train/val/test for contrastive learning. In parallel, a memory bank is built from KU "views" (Title-only, Topics-only, Title+Topics, etc.) for retrieval. (3) Model Layer: a frozen base LLM encodes query and memory texts into token states. GCAM applies cross-attention with LoRA projections, then uses Token Gate to filter memory tokens and Head Gate to select useful attention heads. At inference, queries (course titles, topics, or micro-assessments) are scored against the memory bank to retrieve the most aligned KUs and outcomes.

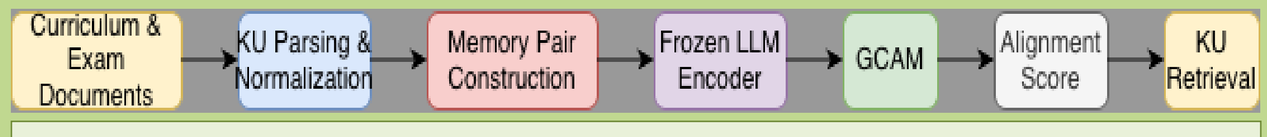

Fig 3: End-to-End Pipeline

Conclusion

Our aim is to provide students with an extra evaluation method to assess their skills based on the guidelines of CAE- Knowledge Units. This way they can focus more on the skills they need and become more prepared for industry. Similarly, industry can benefit from this process to find workforce that is more suitable for specific work role by assessing skills of candidates.

Acknowledgement

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
College of Engineering
Network Security Lab (NSL)
UNT
IEEE UNT Advancing Technology for Humanity
COLLEGE OF ENGINEERING
IEEE North Tech-SAS (Students Advancing Society)