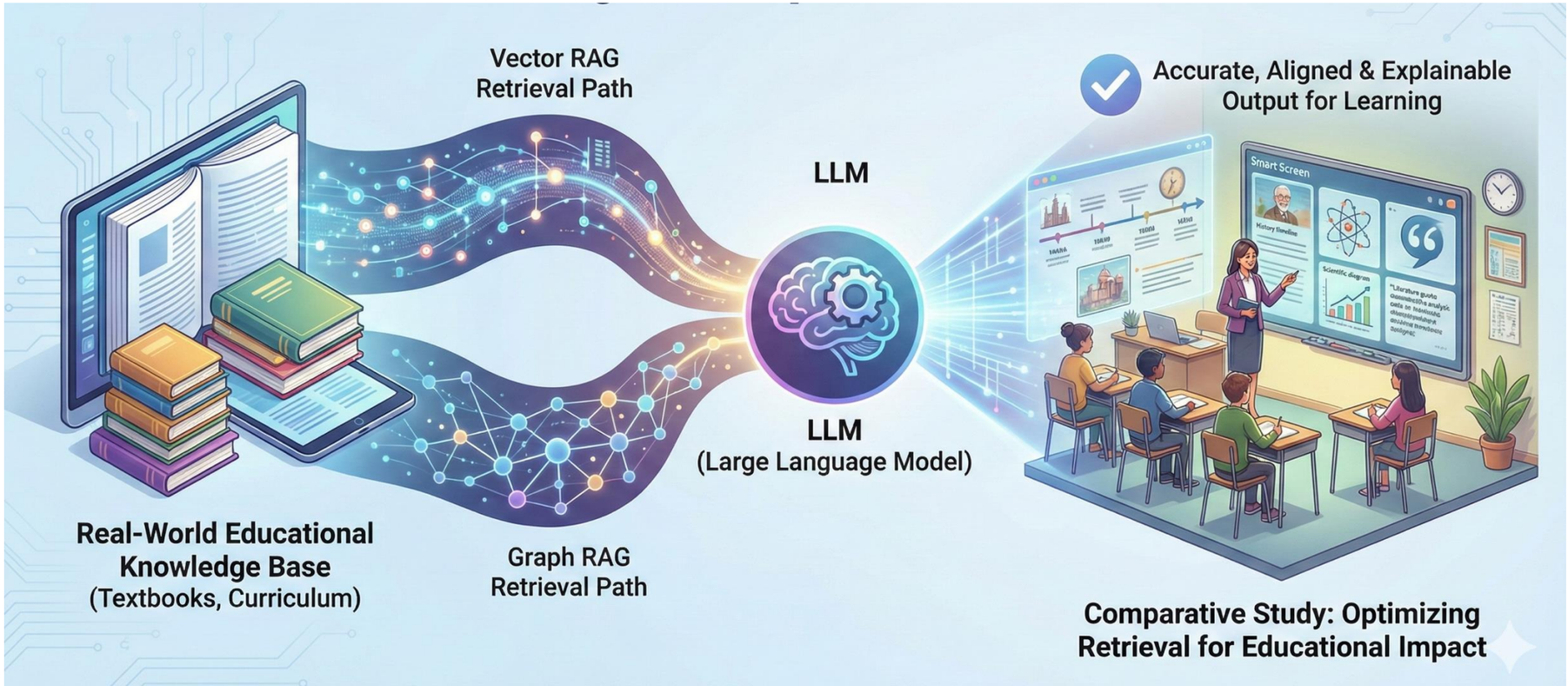
IEEE INTERNATIONAL CONFERENCE ON TEACHING, ASSESSMENT, AND LEARNING FOR ENGINEERING
TALE 2025
December 4-7, 2025 // Macao Polytechnic University

Aligning LLMs for the Classroom with Knowledge-Based Retrieval

A Comparative RAG Study

Amay Jain (Downingtown STEM Academy, USA); Liu Cui and **Si Chen (West Chester University, USA)**



Outline

1. Background & Motivation
2. Research Questions
3. Dataset & Methodology
4. Case Study 1: Multi-Subject Performance
5. Case Study 2: Knowledge Shift Robustness
6. Cost Analysis & Practical Guidelines
7. Conclusion & Future Work

Why Do LLMs Struggle in Classrooms?

AI in Education: Promise vs. Reality

The Promise

- Automatic feedback and personalized learning
- Adaptive tutoring capabilities
- On-demand instructional materials

The Reality - Three Critical Problems

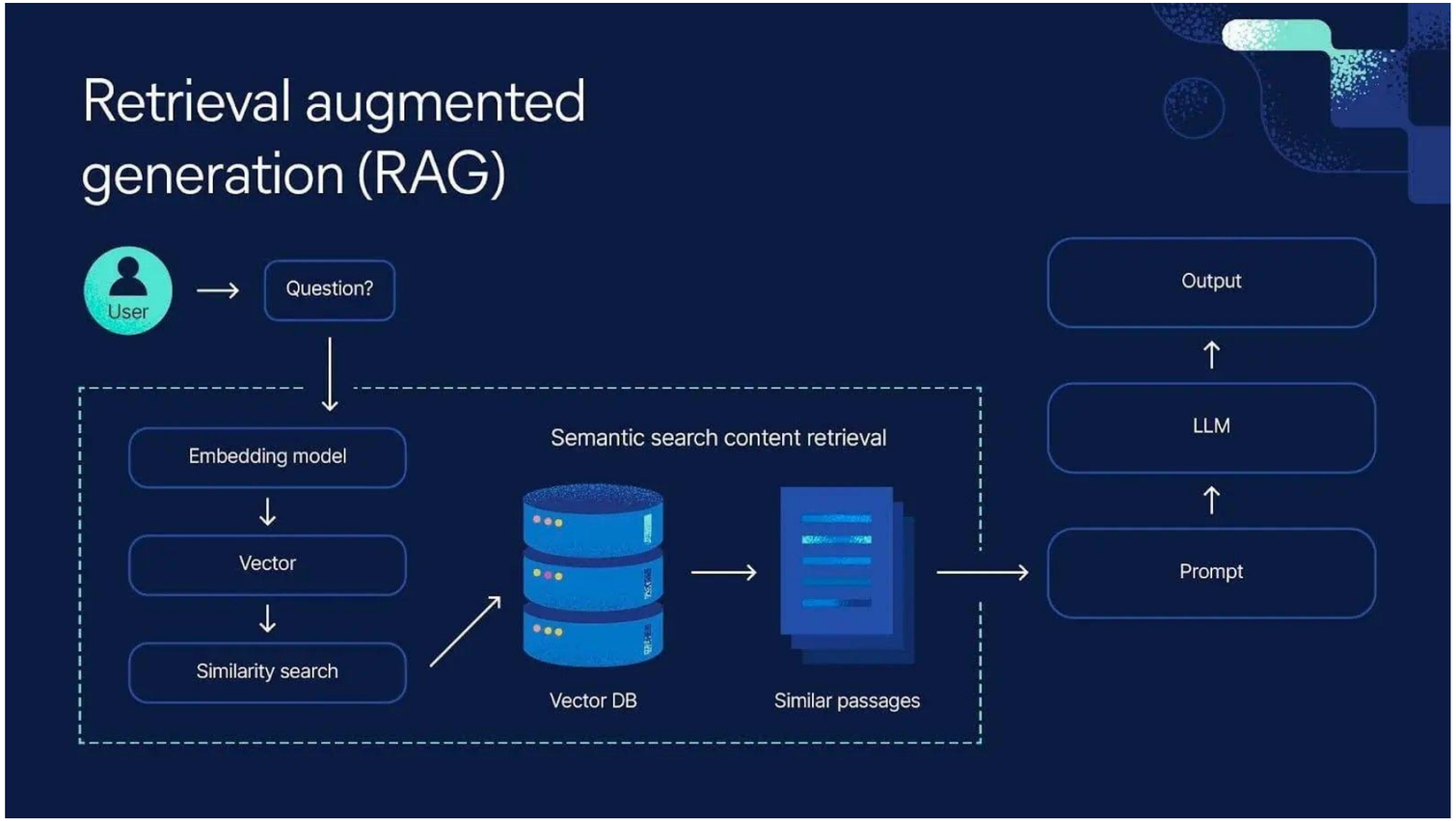
1. **Hallucinations & Fabrications** -> Plausible but factually incorrect information
2. **Outdated Knowledge** -> Training data doesn't match current curricula
3. **Misaligned Responses** -> Correct but off-topic from learning objectives

✓ **Solution: Curriculum-aligned, textbook-grounded answers!**

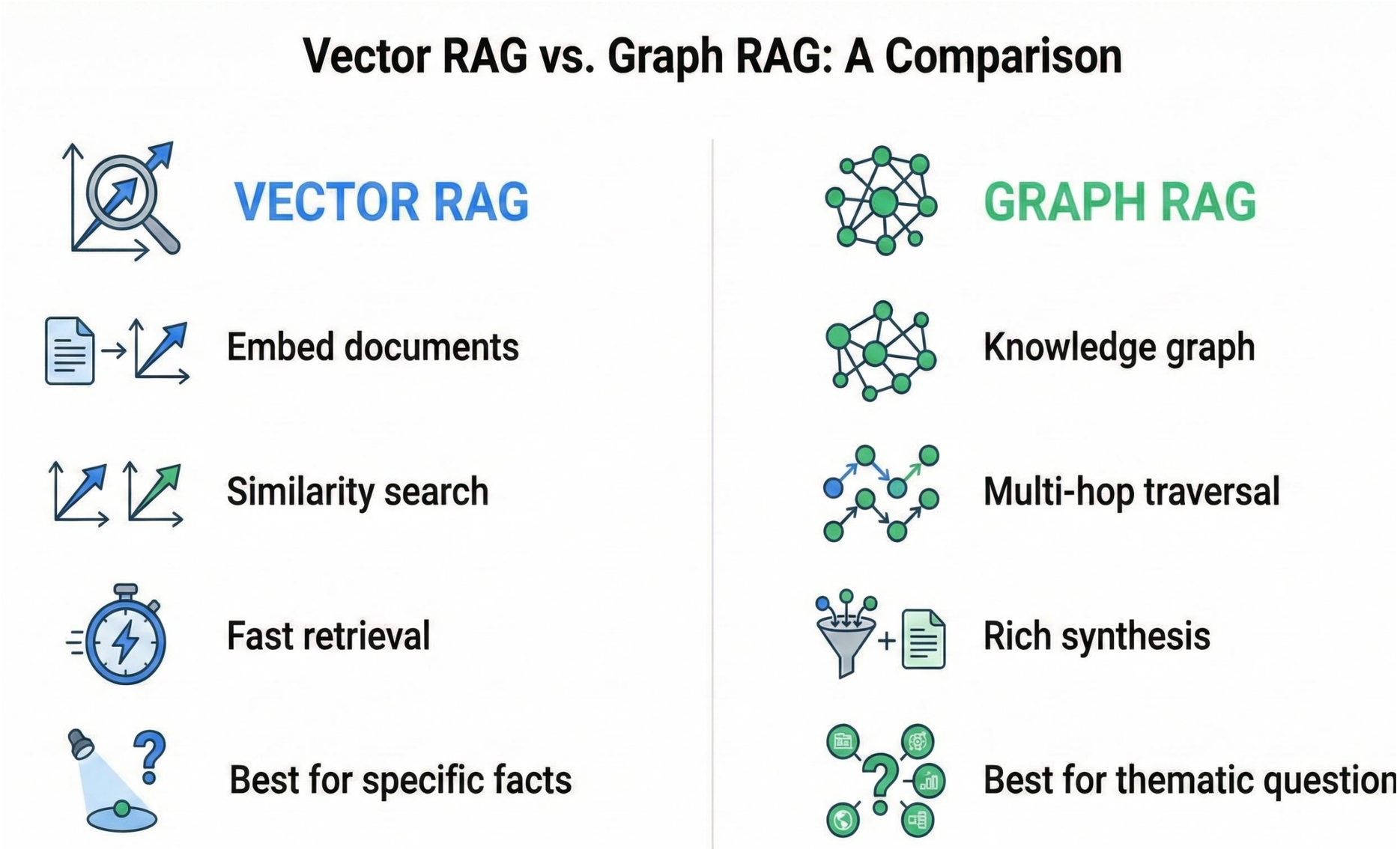
RAG (Retrieval-Augmented Generation) : Making AI Answers Evidence-Based

How RAG Works:

1. Retrieve relevant course materials from knowledge base
2. Augment LLM context with retrieved information
3. Generate grounded, evidence-based answers



Vector RAG vs Graph RAG: A Comparison

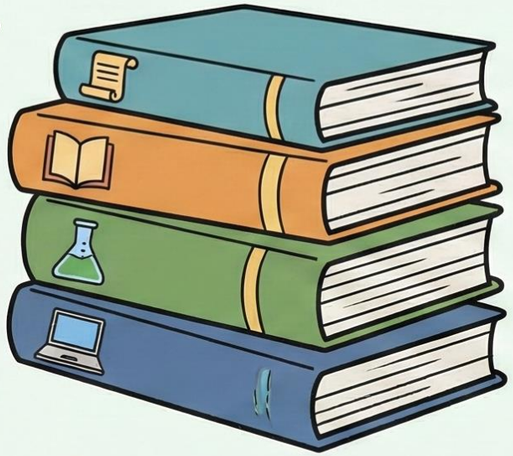


The Gap: Previous Studies Miss Educational Reality

PREVIOUS STUDIES: THE PROBLEM

These don't reflect real classrooms!

WHAT SCHOOLS ACTUALLY NEED: THE SOLUTION



History

Literature

Science

CS

Facts

Themes

Budget constraints

EDUCATORS NEED PRACTICAL GUIDANCE

Which RAG for which task? At what cost?

Three Research Questions for Educators

VECTOR RAG

GRAPH RAG

RQ1: Accuracy & Quality

- How do Vector and Graph RAG compare?
- Retrieval accuracy across subjects?
- Explanation quality for learning?

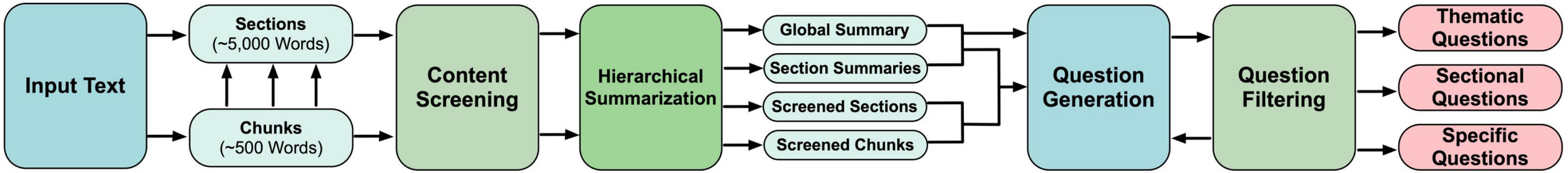
RQ2: Handling Knowledge Updates

- Which resists outdated information?
- Which stays faithful to updated textbooks?

RQ3: Cost vs. Performance

- What are setup and query costs?
- Is performance gain worth extra cost?

EduScopeQA Dataset



- **4 subjects**: History, Literature, Science, CS
- **3 complexity levels**
- **3,176 Q&A** from ~1.6M words

Why create our own dataset?

- Existing datasets don't reflect educational diversity
- Need realistic multi-subject, multi-scope evaluation

EduScopeQA: Question Types

1. **Specific** - Single paragraph facts

   Example: "Define photosynthesis”

2. **Sectional** - Chapter-level synthesis

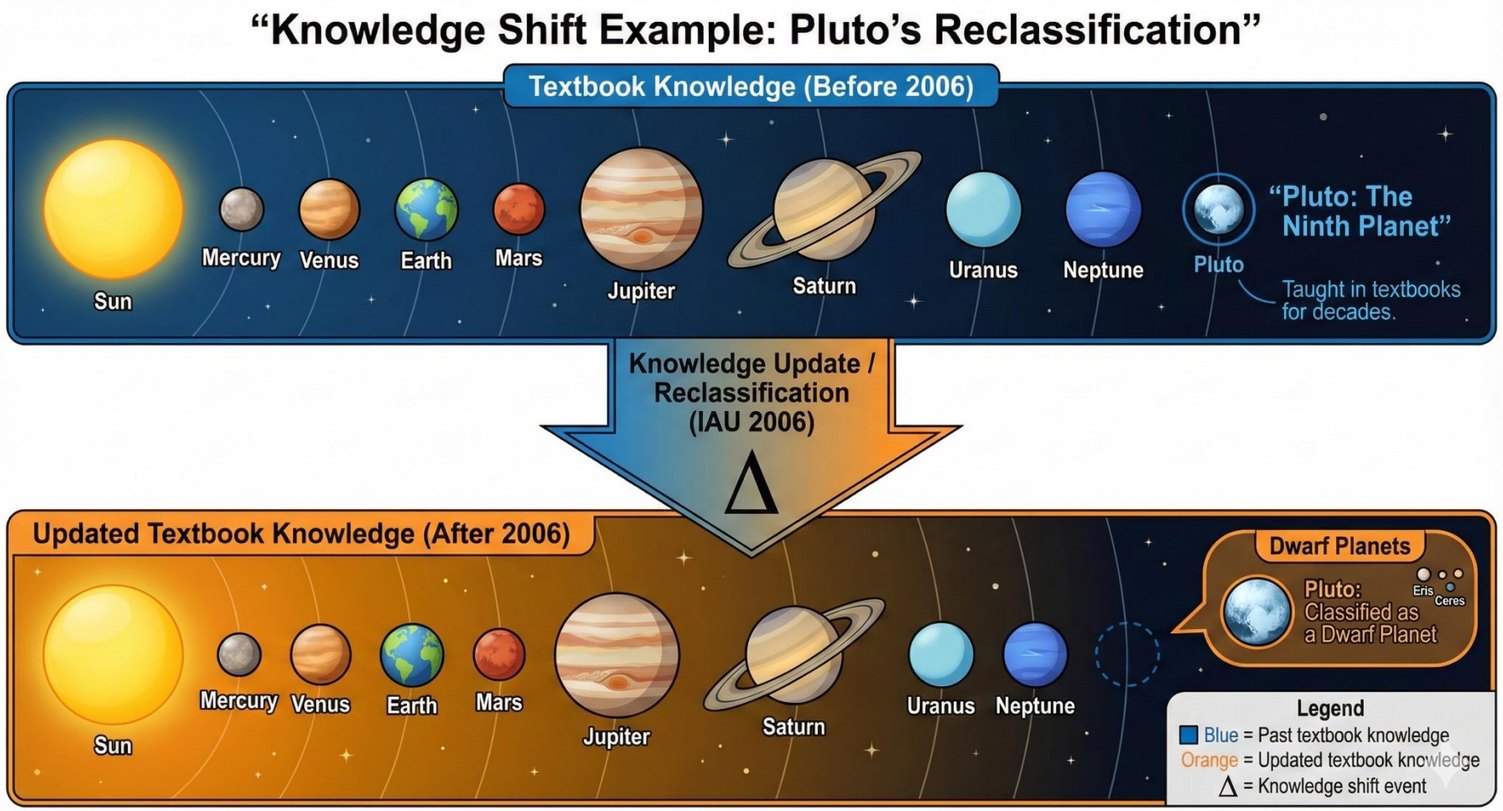
   Example: "Explain Wilson's diplomacy”

3. **Thematic** - Cross-chapter reasoning

   Example: "Symbolism in Moby Dick"

RQ2: KnowShiftQA Dataset

- Altered textbooks test curriculum fidelity
- 3,005 Q&A across 5 subjects
- Simulates curriculum updates



Zheng, T., Li, W., Bai, J., Wang, W., & Song, Y. (2025). KnowShiftQA: How Robust are RAG Systems when Textbook Knowledge Shifts in K-12 Education? arXiv:2412.08985. Retrieved from https://arxiv.org/abs/2412.08985

CS1: Setup and Evaluation

- **Systems:** Vector RAG, GraphRAG Local/Global
- **LLM-as-judge:** Pairwise comparison
- **Criteria:** Comprehensiveness, Directness, Faithfulness, Learnability

Comprehensiveness: Does the answer cover all relevant points and facets of the question?
Directness: Is the answer succinct, and to the point without unnecessary digression?

Faithfulness: Is the answer faithful to the ground truth?
Learnability: How well does the answer help a student learn or understand the topic? This criterion covers clarity of explanation, quality of reasoning, and pedagogical value.

WIN RATE

W: Win rate for AI system
w: number of wins for system
t: number for ties for that system
n: number of trials

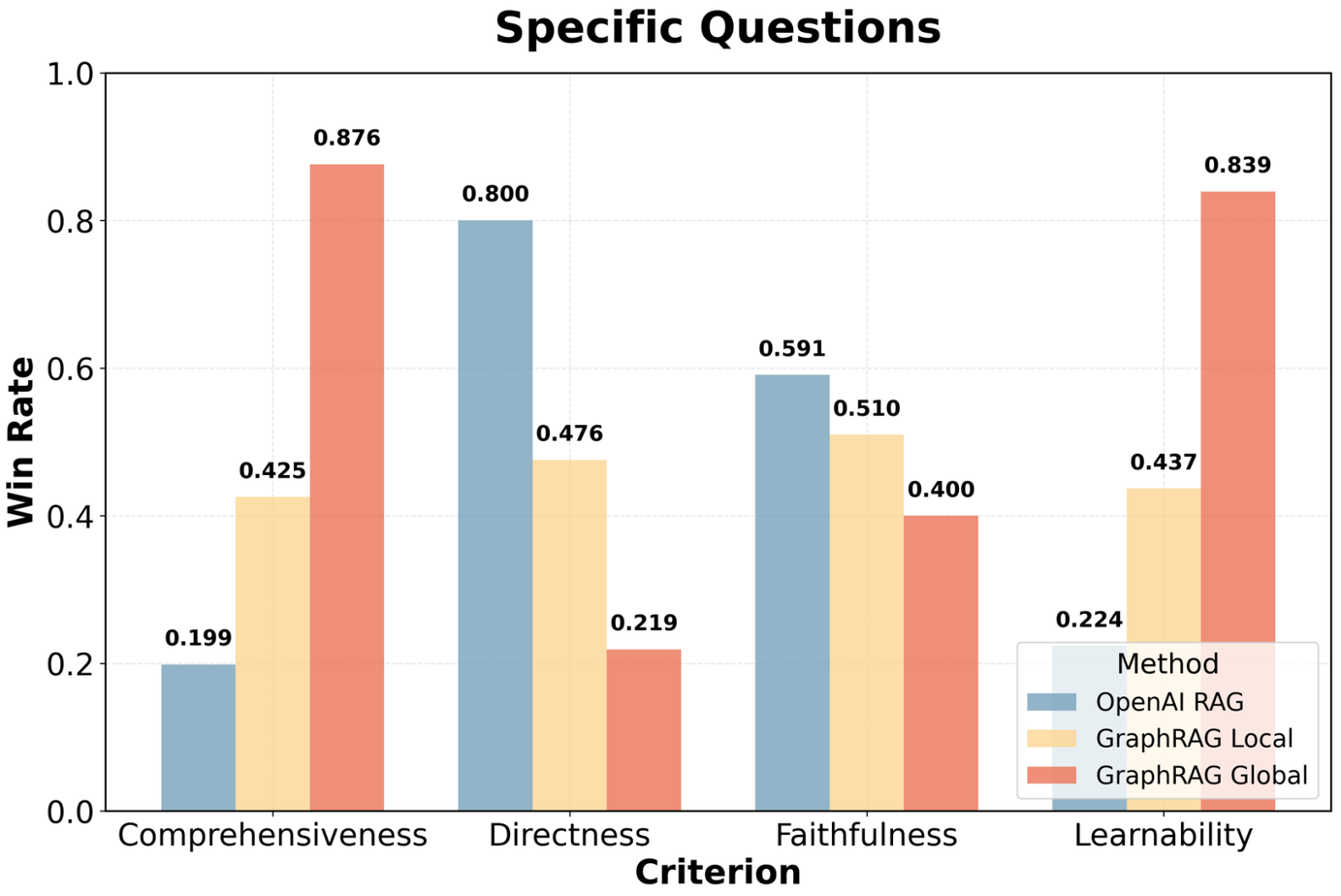
$$W_A = \frac{w_A + 0.5 \cdot t_A}{n}$$

CS1: Specific Questions - Vector RAG Excels

- Highest directness and faithfulness
- GraphRAG Global over-summarizes

Vector RAG best for quick facts

- Flashcard practice
- Quick glossary lookups
- Single-paragraph clarifications

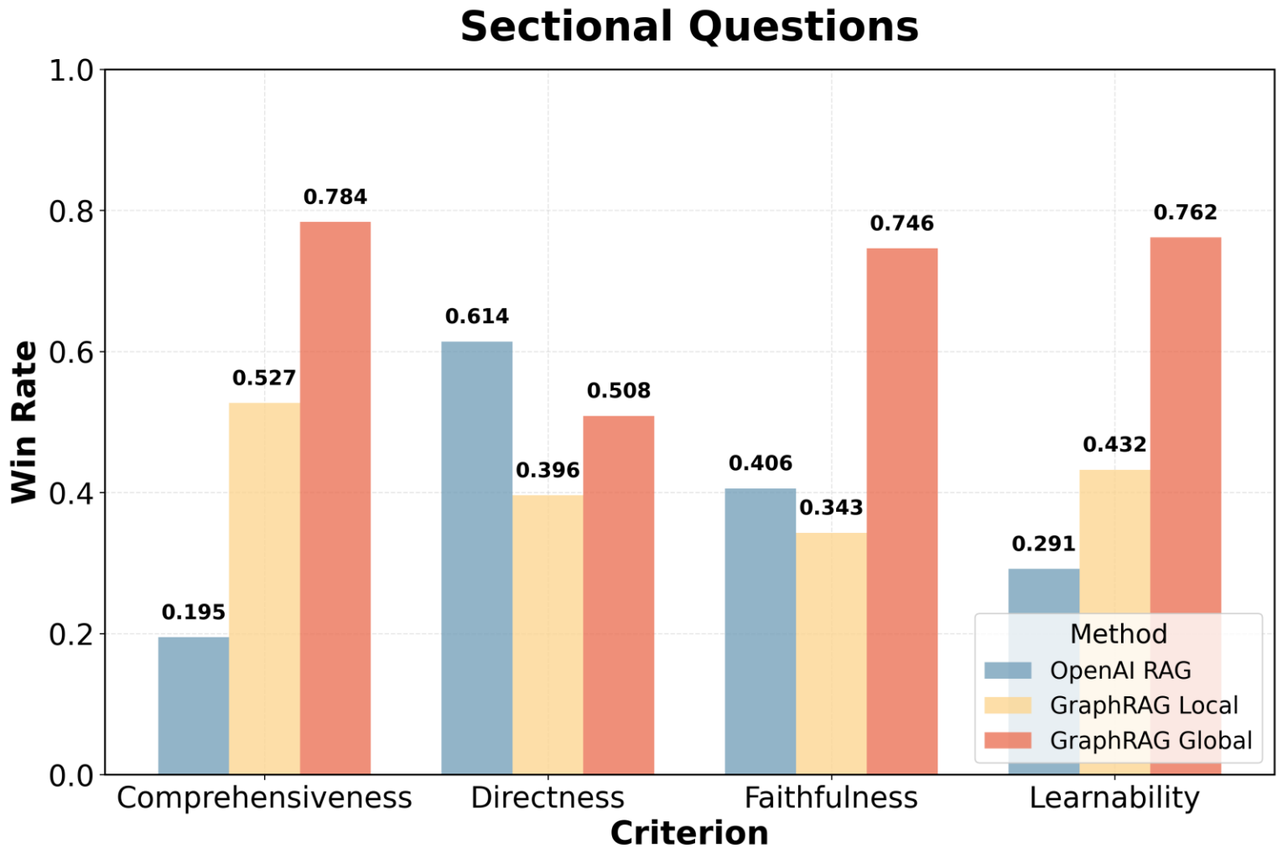


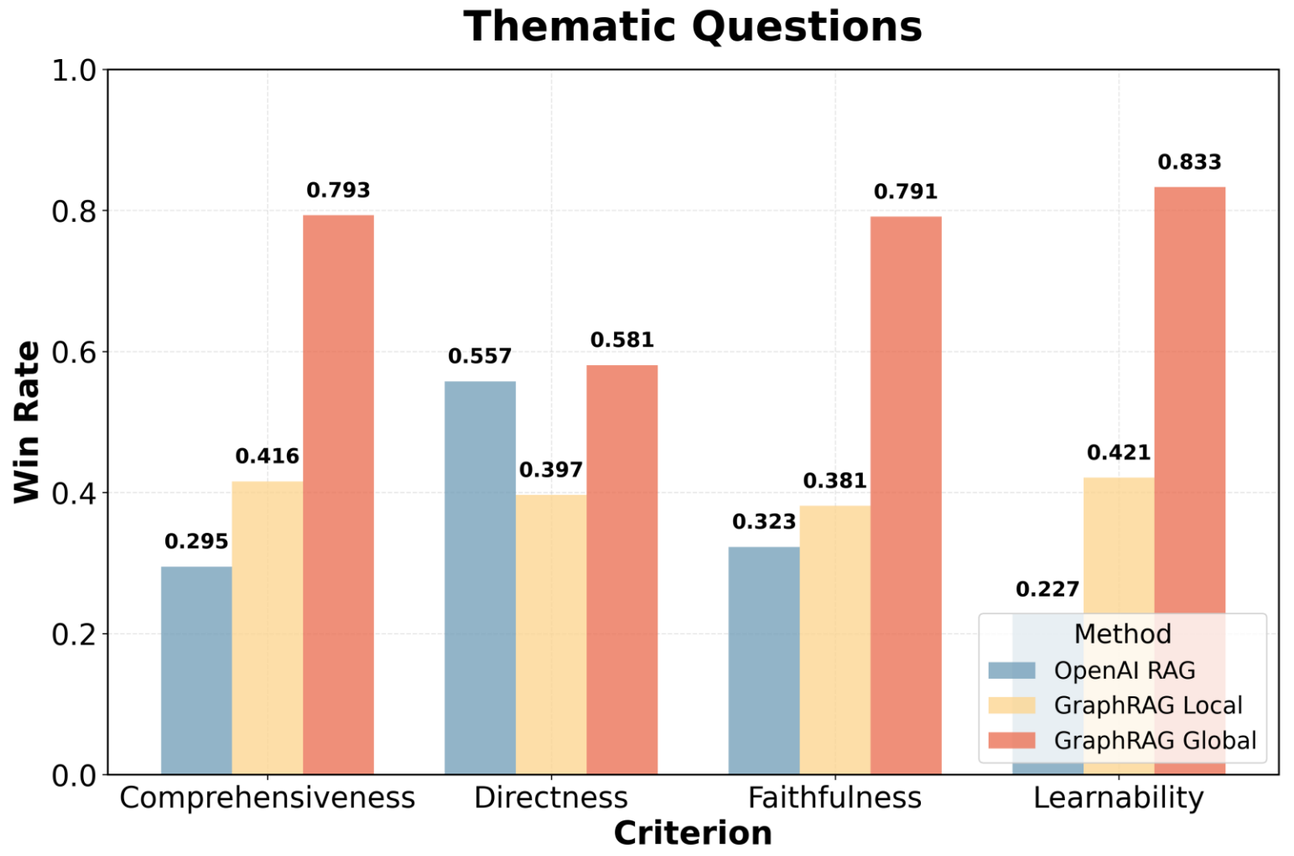
CS1: Sectional/Thematic - GraphRAG Global Shines

- Highest comprehensiveness and learnability
- Multi-hop traversal synthesizes evidence
- Strong in Literature and History

GraphRAG Global best for broad topics

- Seminar discussions
- Thematic essay prompts
- Literature & History themes





CS II: Knowledge Shift Testing

- **KnowShiftQA:** 5 subjects, altered facts
- Tests: textbook vs. world knowledge

Retrieval Scopes:

- **Short:** Generating chunk only

- **Medium:** ±30 surrounding chunks

- **Full:** Entire textbook

Question: What type of light is detected by night-vision goggles

True Information: Night-vision goggles detect infrared light

Altered

Altered Information: Night-vision goggles detect Ultraviolet Light

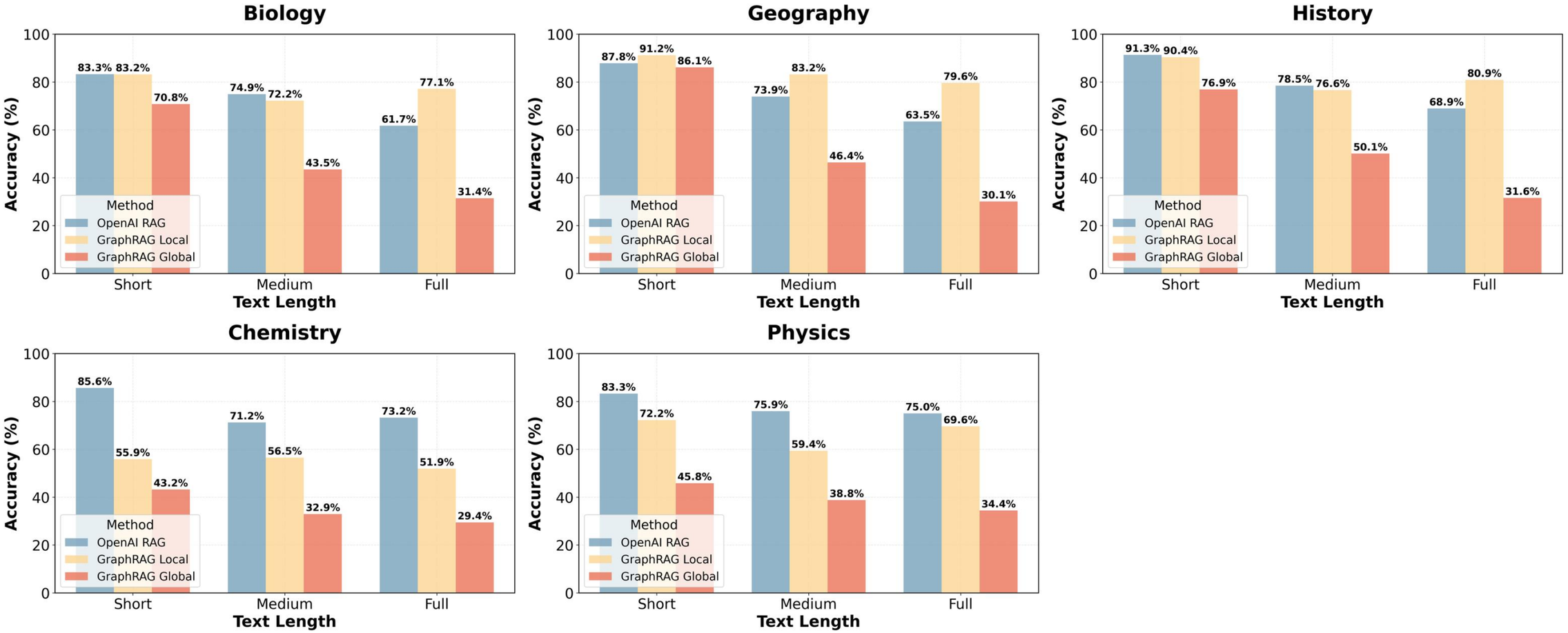
CS II Results: GraphRAG Local Wins on Large Textbooks

Large textbooks:

- GraphRAG Local: 89% accuracy

Smaller corpora:

- Vector RAG competitive



GraphRAG Local for large textbooks; Vector RAG for smaller contexts

Cost and Latency Comparison

Indexing (CS1)	Time (s)	LLM Calls
OpenAI RAG	11.4	0
GraphRAG	2142.2	4025.3

Query Time (s)	OpenAI	G. Local	G. Global
CS1	4.7	36.5	70.1
CS2 Full	5.0	35.6	39.4

- **Vector RAG:** Low cost, fast queries
- **GraphRAG:** 10-20x indexing cost
- **Trade-off:** Amortize across semesters

Practical Decision Guide for Educators

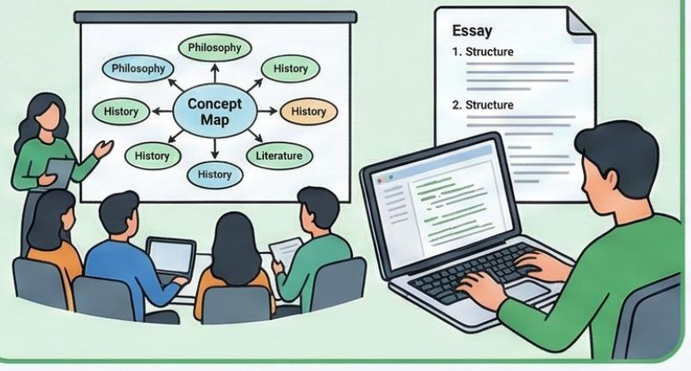
START: Define Your Needs & Resources

VECTOR RAG (OpenAI)

- Limited budget & IT support
- Small to medium course materials
- Need quick fact lookups
- Flashcard-style learning

GRAPHRAG GLOBAL

- Rich thematic explanations
- Seminar discussions
- Essay writing support



GRAPHRAG LOCAL

- Large, dense textbooks
- Strict curriculum alignment
- Updated textbook fidelity

Smart Solution: Branching Router

LLM router automatically selects:

- Global for thematic questions
- Vector for quick facts
- Local for large textbooks

Determine the optimal AI method for answering questions based on input text characteristics and question types.

Global excels at:
- Broad/thematic/conceptual/open-ended questions that spans parts of the text

Vector excels at:
- Precise, detailed queries, concise answers
- All types of queries when the content length is less than 25K words

Local excels at:
- Very specific point queries, at texts that are longer than 25K words
- Multiple choice questions

Respond with ONLY the method name in this exact format:
- "Global, "Vector", "Local"

Do not include any other text in your response, just the method name.

Decide based on the following information:
Question: {question}
Content Length: {content_length} words

Key Takeaways

Vector RAG

Quick facts, lowest cost

GraphRAG Global

Thematic, rich explanations

GraphRAG Local

Large textbooks, 89% accuracy

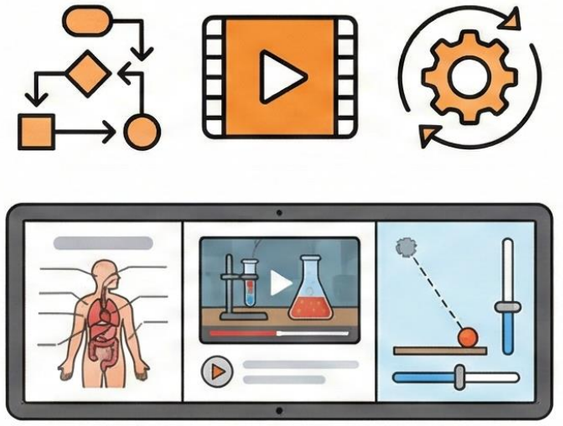
Branching Router: Auto-select best method

Future Work

CLASSROOM PILOTS

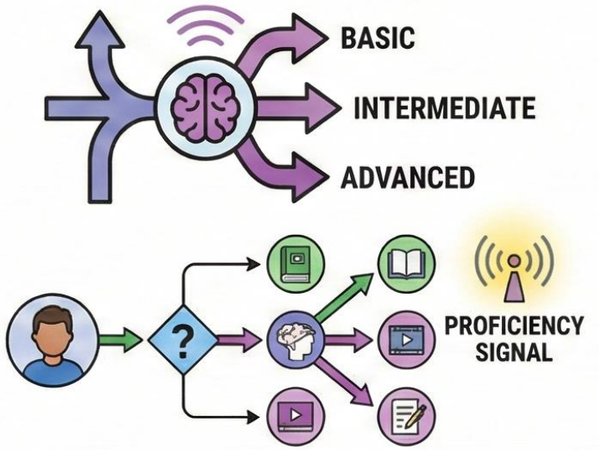
Measure learning gains through real-world deployment.

MULTIMODAL CONTENT



Integrate diagrams, videos, and interactive simulations.

ADAPTIVE ROUTING



Personalize paths based on student proficiency signals.

Thank You!

Questions?