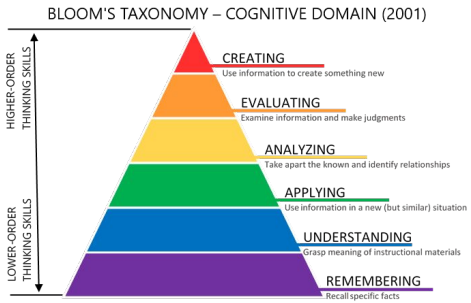
Qanqa: Personalized Tutor System

Multi-agent educational platform aimed to deliver personalized learning for K12 students by tailoring assessments and activities to unique needs and progress

Problem

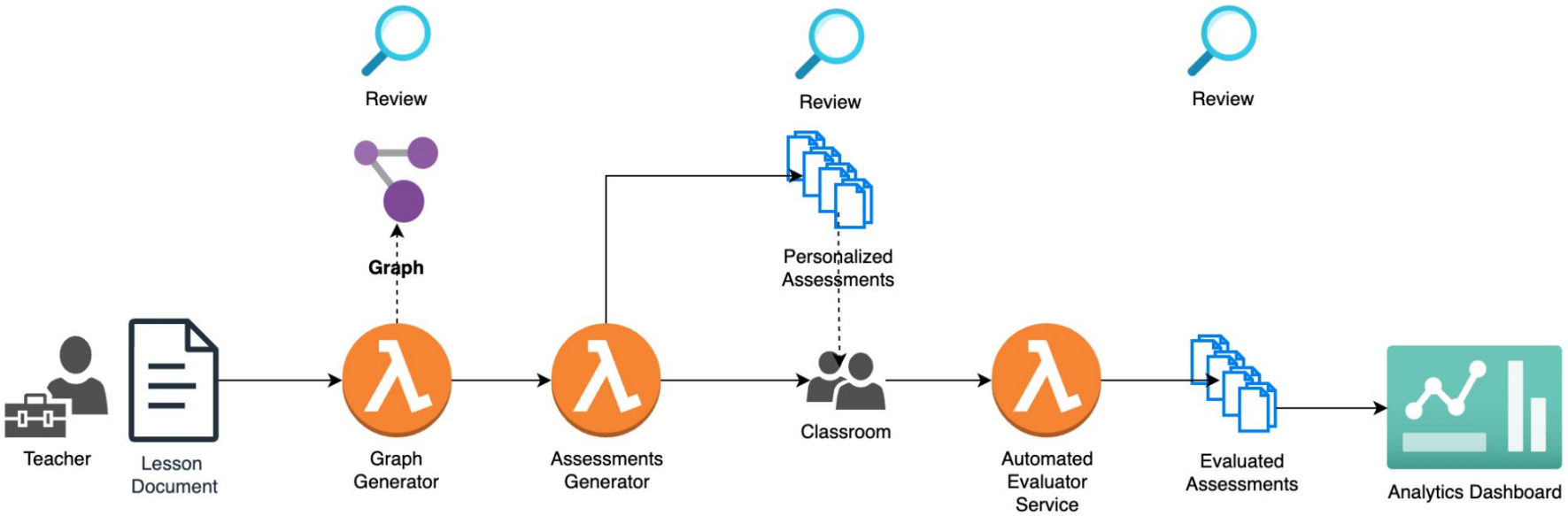


1. Traditional learning environment struggle to provide personalized support due to time/resource constraints.
2. Teachers can be overwhelmed by the effort required to tailor exercises and assessment to individual students
3. Current education system, especially in Latin America, lack detailed measurement of student understanding.
4. Multiple efforts are being made to reduce the gap in education quality existing in Latin America schools

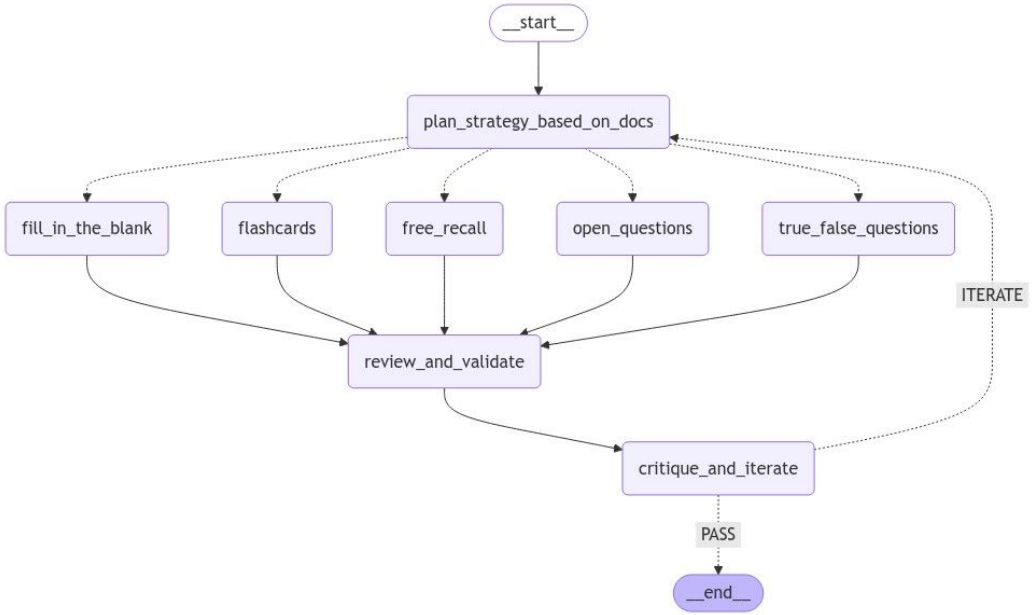
An educational platform designed to empower teachers

- ☑ Qanqa **automates repetitive tasks**, such as assessments, enabling teachers to focus on creating engaging and effective classroom experiences.
- ☑ Teachers can make **data-driven decisions** using dashboards that provide high-quality insights into their student's learning progress.
- ☑ With diagnostics and interactions of autonomous AI agents, teachers can design personalized strategies and **address students' learning gaps**.

— DEMO


Review
Review
Review
Graph
Personalized Assessments
Teacher
Lesson Document
Graph Generator
Assessments Generator
Classroom
Automated Evaluator Service
Evaluated Assessments
Analytics Dashboard

LangGraph

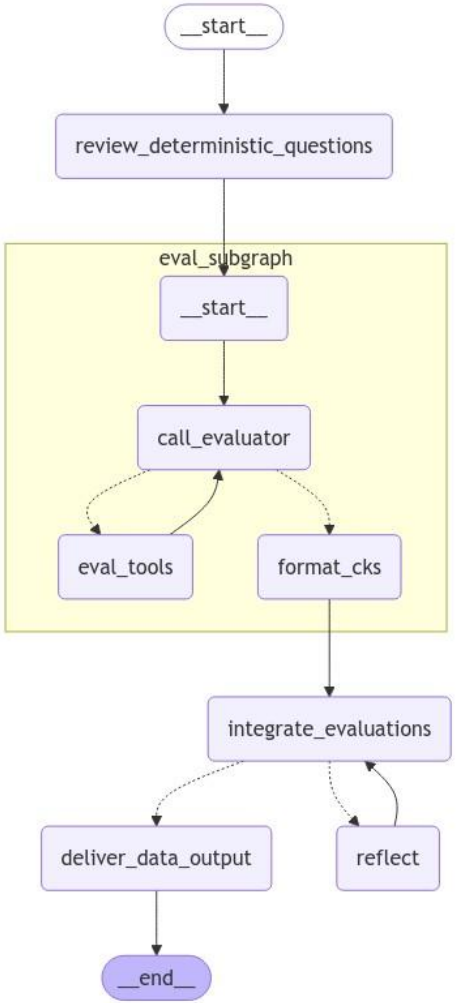


Metrics: What matters most?

- Does the system incorporates users past responses effectively?
- Are all exercises non-redundant and designed strategically?

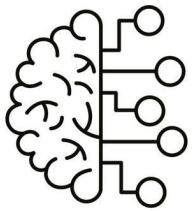
Data Slices: What is our system performance regarding:

- Topics
- Levels of difficulty
- Different Students Knowledge State
- Course Goals Types



LangGraph

Automatic Answer Grading alleviate one of the main key challenges faced by teachers.



LLM’s already demonstrate significant world knowledge and it will keep growthing

The high scores in MMLU, HumanEval, MGSM, and DROP indicate that the model has a comprehensive and deep understanding across various domains.

However, these benchmarks primarily evaluate structured or multiple-choice questions rather than open-ended ones. And this raises the question:

Can LLMs serve as effective automated graders for open-ended student assessments? And if not, what role RAG would play improving their grading capabilities?

Results

Mohler dataset: Open-source dataset that includes questions and graded assignments. The scope is only for Data Structure.

Methods	MAE	RMSE
Gpt-4o-mini		
One-Shot	0.81	1.25
Reflect	0.82	1.33
Basic Rag	0.82	1.37
Gpt-4o		
One-Shot	0.87	1.35
Reflect	0.82	1.34

Next Steps:

Tune Prompts with DsPy ;)

Collect Expert feedback or Synthetic data

Create broader benchmarks with other topics

Measure other metrics (Fairness, Feedback consistency, etc)

Create new features [Audio interface, Dashboards]

We are going to run a pilot test in March, at the beginning of the school year.

Agentic Workflow

Key Generated Items:

Lesson taxonomy [Key concepts]

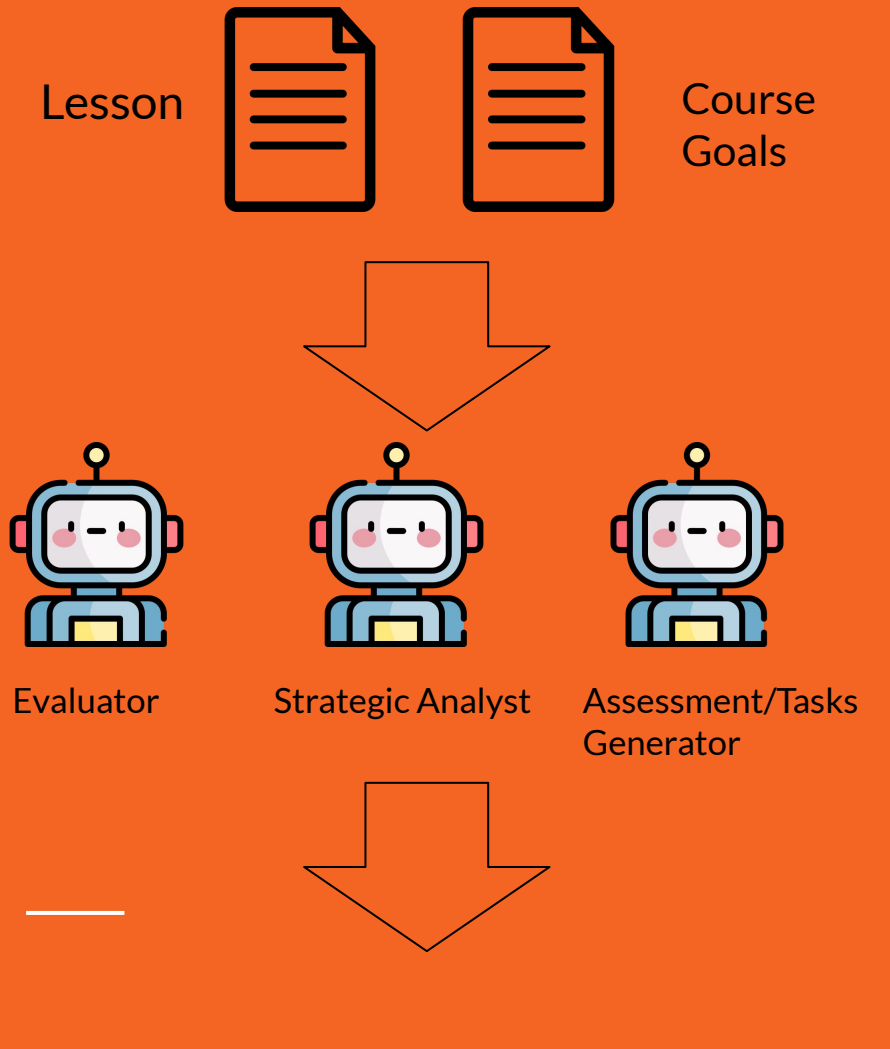
Class ----------

Student

Current Knowledge State: Blooms Schema score per concept

Strategic Learning Document [Strengths - Areas of improvements]

Tasks and Assessments [Reinforce Concepts - Measure new acquired knowledge]

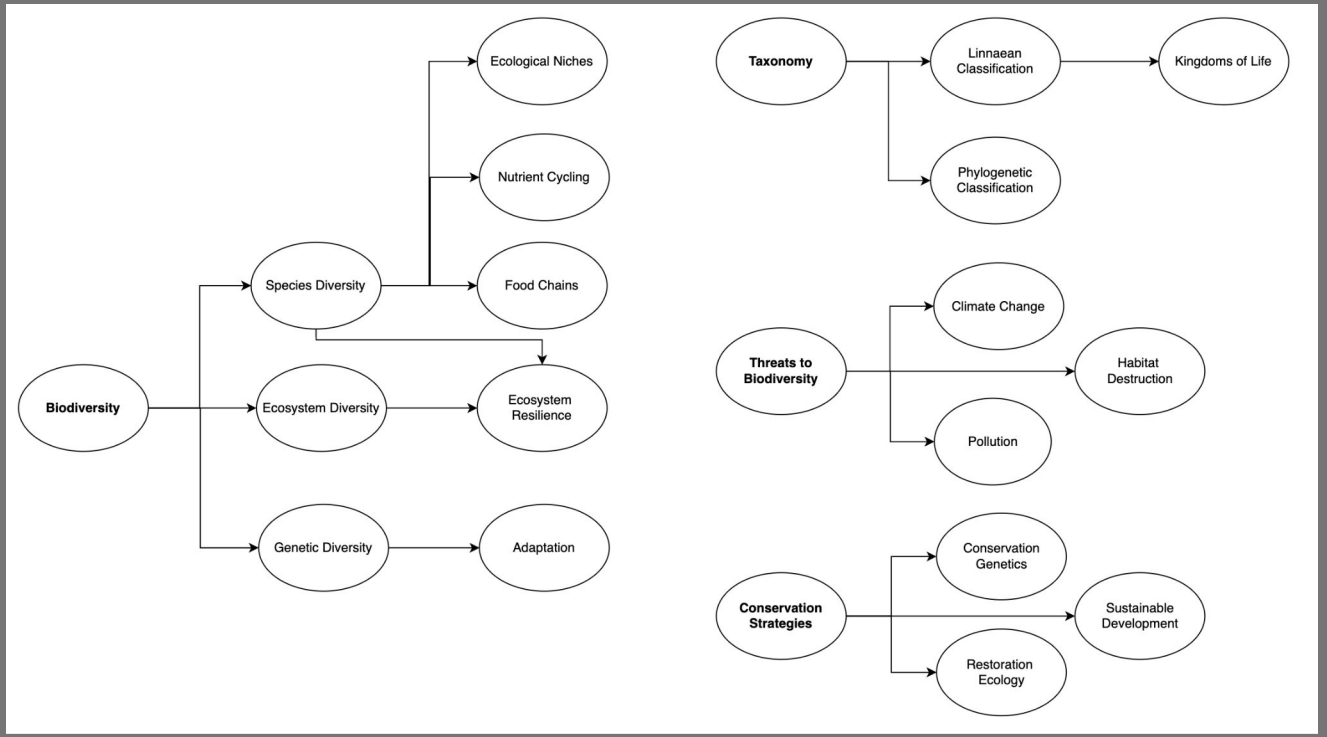


Demo

Fixed Taxonomy

Generalized Strategic Document

Generalized Assessments



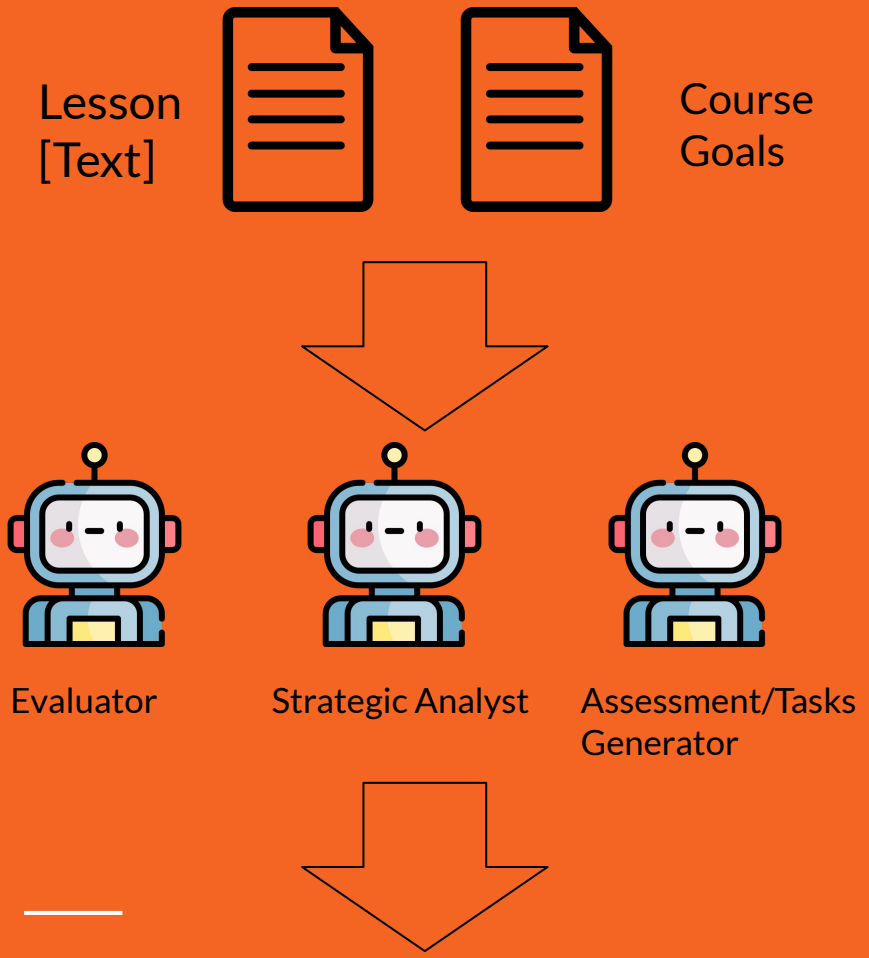
Agentic Workflow

Key Tools:

rag_on_lesson: Validates answers and review lesson doc

align_with_course_goals
update_blooms_score

create_flashcards
create_fill_in_the_blank
create_true_false_questions
create_quizz
create_free_recall



Demo

Simulated Student with specific knowledge gaps

- CKS
- Strategic Document
- Assessments