






# AI Tutor: Complete Learning System for Data Science and GPU Acceleration

## Overview

The **AI Tutor** is an intelligent, interactive learning system designed to teach students core and advanced topics in **Data Science** and **GPU Acceleration**. Built using **Gradio**, **Transformers**, **FAISS**, and **Sentence Transformers**, it leverages state-of-the-art **RAG (Retrieval-Augmented Generation)** pipelines, **Socratic tutoring**, and **gamified practice modes** to engage users in meaningful, structured learning.

---

## ✨ Features

Mode	Description
 Learning Mode	Structured learning paths for Python, Pandas, cuDF, Machine Learning, etc.
 Tutor Mode	Socratic-style interactive Q&A with RAG-enhanced LLMs
 Game Mode	XP-based gamified practice including flashcards, quizzes, and coding puzzles

## 📖 Learning Mode

Learners can select a topic and follow a structured, curated learning path. Each path includes detailed content breakdowns, and direct links to trusted resources (e.g., RAPIDS, scikit-learn, CUDA docs).

Additionally, a built-in search bar lets users discover relevant external resources dynamically, enhancing exploratory learning.

Structured learning paths across 6 major domains:

- Python for Data Science
- Pandas Deep Dive
- GPU Acceleration with cuDF
- Machine Learning Fundamentals
- RAPIDS Ecosystem
- Data Visualization

Each topic includes:

- Learning objectives
- Topic prerequisites
- Duration estimates
- Trusted resource links
- Suggested next steps

## Tutor Mode

This mode follows a semi-Socratic approach—the tutor poses guiding questions before offering insights.

It streams clear, concise, and contextual responses in real time, encouraging deeper understanding. The system intelligently weaves in concepts related to GPU acceleration, parallelism, and data science tooling, adapting to the learner's queries.

Powered by **Socratic questioning** and **RAG-based guidance**, it responds with:

- Polished, structured explanations
- Clarifying questions
- Analogies and examples
- Summary takeaways
- Contextual AI-generated quotes
- Documentation assistant for 12+ data science/GPU libraries

## Game Mode

Designed to reinforce learning through XP-based gamification, Game Mode features three components:

### Flashcards

Learners generate flashcards on specific topics (e.g., cuDF, pandas, CUDA). Cards include randomized flip logic with two cards out of which one has correct answer. This enhances retention through interactive recall. Features include,

1. AI-generated via RAG.
2. Flip cards to see the correct answer.
3. Tracks correct/wrong choice of card with some fun quote on picking wrong card.

### Quiz Challenges

Users are tested with multiple-choice questions. Each correct answer earns XP, and learners track their progress visually as they level up. Features include,

1. Topic-based MCQs (2 per round)
2. Auto-evaluated, XP points awarded
3. Feedback with correct answers and explanation

### Coding Puzzles

Learners are given CPU-based code snippets and are tasked with rewriting them using GPU-accelerated libraries like cuDF, cuML, or CuPy.

Each puzzle includes:

1. Convert pandas code to cuDF for GPU optimization
2. Difficulty levels: Beginner, Intermediate, Advanced
3. Benchmarks execution time
4. Visual and numeric speedup metrics

---

## System Architecture

- **LLM:** NousResearch/Hermes-2-Pro-Mistral-7B (supports Socratic and tutoring dialogue)
  - **Embedder:** all-MiniLM-L6-v2 via SentenceTransformer
  - **RAG Engine:** FAISS + custom document chunking
  - **Interface:** Gradio Blocks
  - **Caching:** Chunks and index stored locally as `chunks.pkl` and `faiss.index`
  - **Sources:** Configurable via `custom_urls.txt`
- 

## How It Works

### Initialization

- Detects GPU and sets optimal model configurations
- Loads or builds FAISS index from scraped academic and technical documents
- Downloads HTML, parses clean text, chunks into semantic segments
- Uses MiniLM to embed, FAISS to index, and Hermes-2 LLM to generate responses

### Tutor Response Flow

1. User enters question
2. System retrieves top-k context from RAG
3. Injects into prompt for LLM
4. Streams Socratic-style output
5. Enhances with contextual quote

### Flashcards

- Generated on-the-fly from retrieved content
- Strict JSON validation to avoid malformed outputs
- Two cards shown, one correct, one distractor
- Feedback for both right and wrong flips

### Coding Benchmark

- Compares user-submitted cuDF code with ground truth
- Uses `perf_counter` for timing

- Uses cProfile for profiling
  - Generates a performance summary with visual bars
  - Gives feedback, tips, and real-world insights
- 

## Requirements

- Python 3.10+
  - Transformers, Sentence Transformers
  - FAISS (GPU/CPU)
  - cudf (optional for real GPU mode)
  - Gradio
- 

## How to Use

1. Launch via: `python ai_tutor.py` or notebook cell
  2. Navigate to local server or use public Gradio link
  3. Explore tabs:
  4. **Learning Mode:** Click on a topic to view learning path
  5. **Tutor Mode:** Ask a question for Socratic response
  6. **Game Mode:** Practice with flashcards, quizzes, and coding
- 

## Future Enhancements

- Multi-GPU inference
- LLM fine-tuning on DataScience education
- Student progress tracking
- Real-time visualization panel
- Team leaderboard integration