

Glite

AI-Powered Education

London
October 6, 2026



Meetup vibe

- It is **not** about teaching AI
- Practical applications of AI in education
- Place where EdTech business and academia meet
- Discussions about the opportunities and challenges in creating AI-powered learning solutions
- Networking for people interested in this area
- Technical angle



Meetup series topics

- **RL** Reinforcement Learning in Education
- **KT** Knowledge Tracing, Student Modeling, Cognitive Modeling
- **ITS** Intelligent Tutoring Systems, Adaptive Learning
- **LA** Learning Analytics, Educational Data Mining, Predictive Analytics
- **CD** Curriculum Design
- **AA** Automated Assessment & Feedback, Grading



Agenda

18:45 | Welcome & Opening Remarks

19:00 | Presentation 1: Pavel Katunin

“N+1 Approach to Education”

19:30 | Pizza + Networking

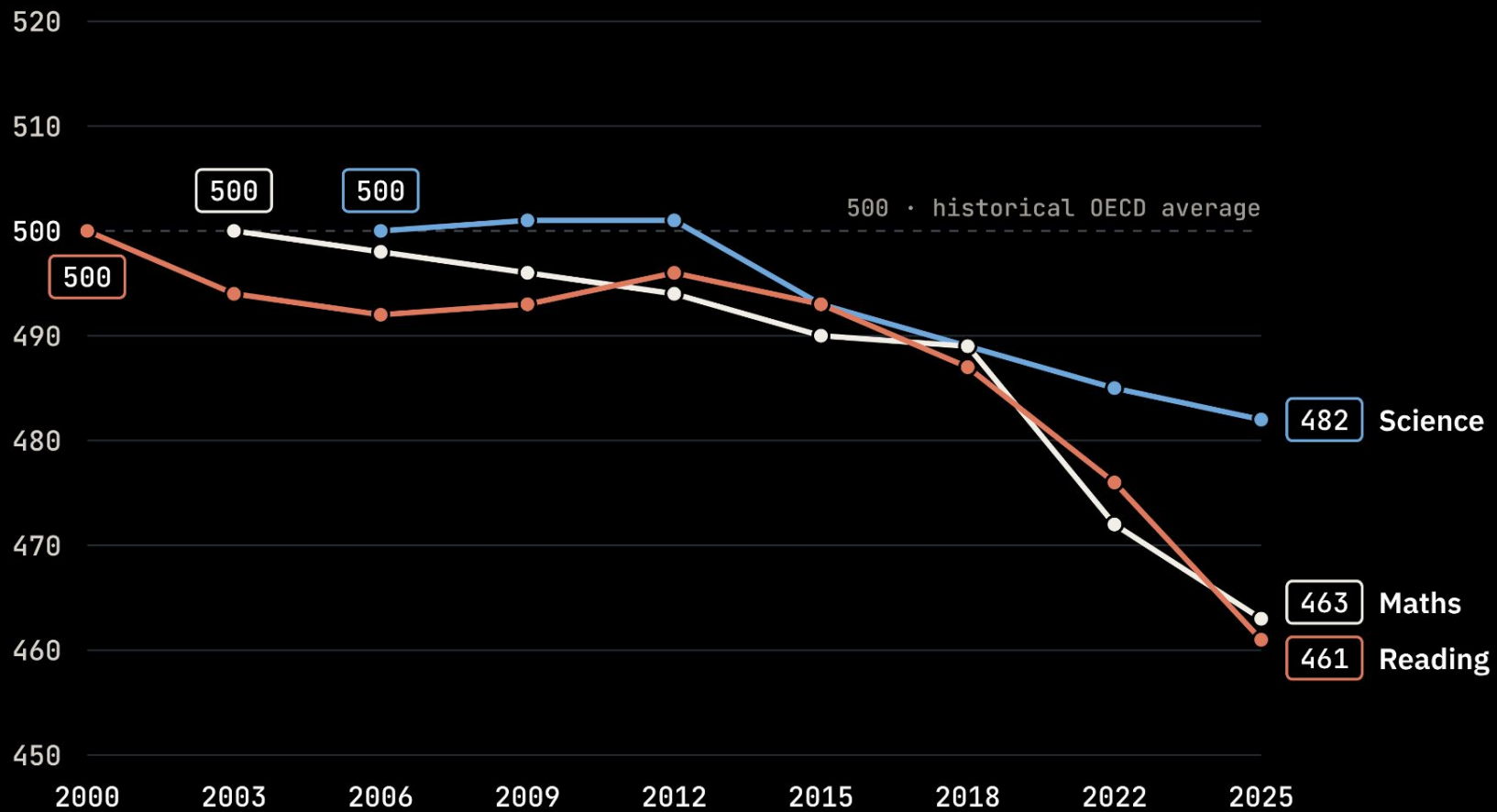
20:15 | Presentation 2: Vassili Philippov

“Finding concepts in learning materials and student answers”

20:45 | Closing Remarks & Next Steps

PISA OECD Average

● Science ● Maths ● Reading



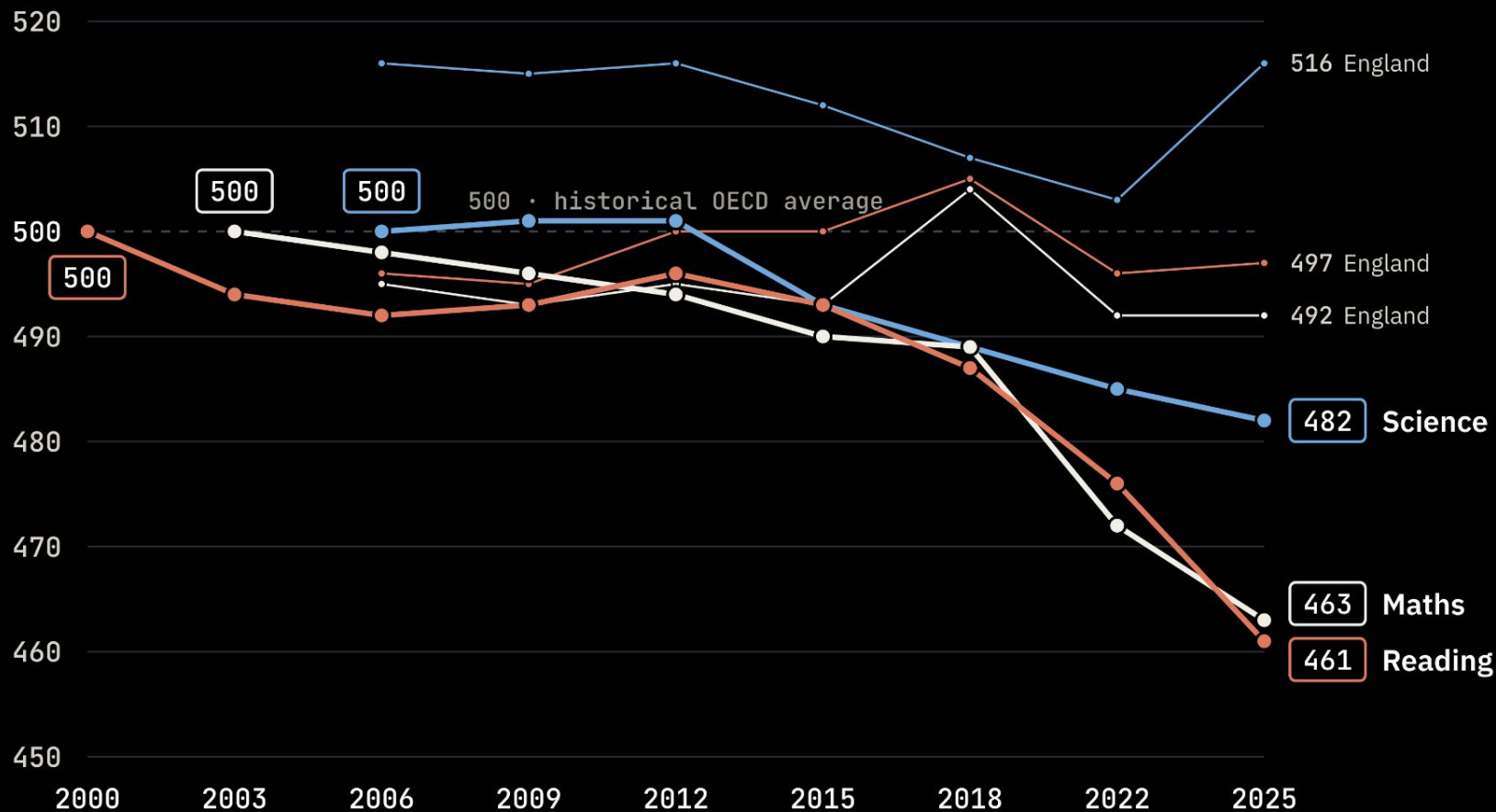
PISA OECD Average

— England

● Science

● Maths

● Reading



1. AI Tutoring with Khanmigo: Two-Year School Experiment

ITS

By Philip Oreopoulos 🇨🇦, Nina Low 🇺🇸

18 schools, two-year randomized trial. Students got Khan Academy with the Khanmigo AI tutor in their daily 25–40 min remedial maths sessions.

Maths +0.06–0.08 SD per year.

AI dose: about 1 meaningful tutoring exchange per week.

The gain looks like practice, not AI.

2. Better Problem Order: Same Chatbot, +0.15 SD

ITS

RL

By Angel Tsai-Hsuan Chung 🇺🇸, Hamsa Bastani 🇺🇸, et al.

770 Taipei high-school students, 5-month Python course, randomized.

Same AI chatbot for everyone; only the order of practice problems differed:
adaptive vs fixed easy-to-hard.

+0.15 SD on an exam without AI. Beginners +0.22 SD, experienced +0.01.

"RL" $\approx$ greedy difficulty matching; no simple adaptive baseline.

3. Cognitive Surrender: Study Time Fell After ChatGPT

LA

By Sina Rismanchian 🇺🇸, Eyad Kurd-Misto 🇺🇸, et al.

10 years of ALEKS logs (3.2M practice sessions). After ChatGPT, time on word problems (pasteable into a chatbot) fell 27–31% in college and high school; graph problems flat.

The effect vanishes under proctoring.

Proctored retention on word problems: odds –25% relative to graph problems ($p=0.04$).

4. "Act Like a 5th Grader" Is Not Enough

KT

By Krisztian Balog , Arild Michel Bakken 

Real reading-test answers from 2,359 Norwegian pupils (grades 4–6).

LLMs told to act like a 5th grader score 96–99%; real pupils score 69%.

Random guessing is closer to real children than any LLM.

Limiting the model's memory fixes the average score, not which questions are hard.

5. StudentSim: Simulators Trained on Real Learners

KT

RL

By Ke Yang 🇺🇸, Jianfeng Gao 🇺🇸, et al.

One fine-tuned simulator per real learner: chess, English writing, maths (60 learners).

Predicts a learner's real next move better than Maia2 and GPT-5.4 (chess: 0.52 vs 0.45 vs 0.23).

Used as an RL reward, it trained a chess tutor rated more accurate (90.5% vs 75.7% without RL).

No real learners were tutored.

6. Who Am I?

Simulated Students With Real History

KT

RL

By Zhangqi Duan 🇺🇸, Andrew Lan 🇺🇸, et al.

Eedi maths: 670 students, 1,775 real tutoring chats.

An RL-trained profile summarises each student's past answers and chats; a simulator predicts their next message.

Beats GPT-5.4 given the raw history.

Honest, modest results: it matches the student's actual misconception only 15% of the time.

7. Knowledge Tracing Inside the Chat

ITS

KT

By Shuyan Huang 🇺🇸, Andrew Lan 🇺🇸, et al.

Predicts whether a student will answer the tutor's next question correctly, turn by turn.

An LLM reads the chat and outputs two readable numbers: student ability and question difficulty (IRT).

LLM methods beat classic KT on real Eedi chats (AUC 65 vs 56).

Only +0.4 AUC over the previous LLM method; the value is interpretability.

8. Stable and Faithful: Simple Model vs Deep KT

KT

By Praveena Padi 🇺🇸, Pradeep Kumar Dolabehera Kakitapelli 🇺🇸, et al.

XGBoost with 13 teaching-meaningful features (spacing, mastery, response time) vs DKT, SAKT, AKT, SimpleKT.

Same information: the simple model ties deep KT (AUC 0.717 vs 0.720).

Better features: it beats them (0.786).

Bonus: duplicated rows in the standard ASSISTments 2009 dataset leak the answers.

9. Which Hint to Give?

Offline RL From Real Logs

RL

AA

By Badmavasan Kirouchenassamy 🇫🇷, Vanda Luengo 🇫🇷, et al.

AlgoPython logs: 2,672 learners, 25K submissions. The old system picked feedback at random, which is ideal for offline RL.

Learned which of 123 expert-written messages to show (Conservative Q-learning).

Large estimated gains, but offline only: no student has used the learned policy yet.

Glite



**How will the next breakthrough
in education happen?**

What qualities will the next best educational product have?

What qualities the new best education product will have?

1.

**1 to 1 education
(0.7 sigma problem)**

2.

**Has objective measurable
metrics of success**

3.

Super human level of teaching

4.

Full Multimodality

Today you will learn:

Today you will learn:

1.

**The most efficient approach
to education**

Today you will learn:

1.

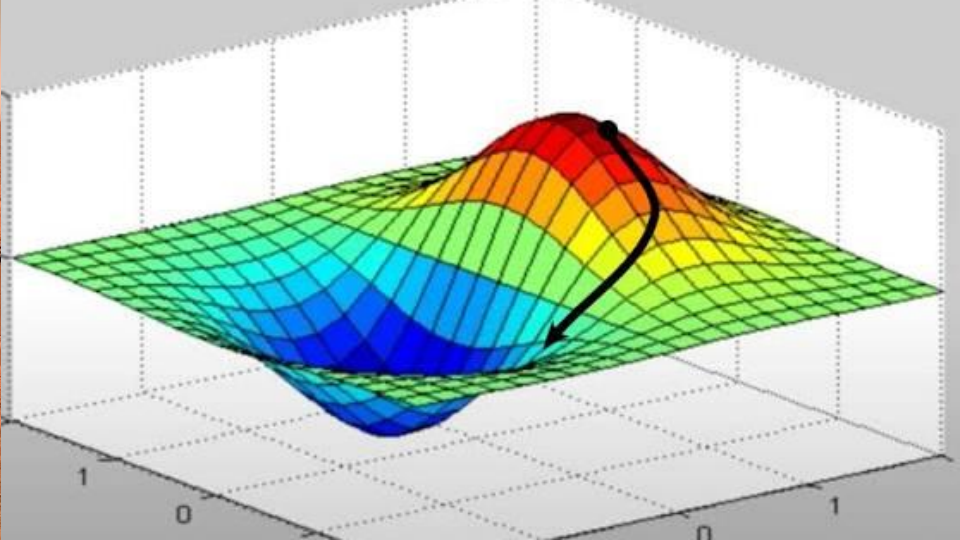
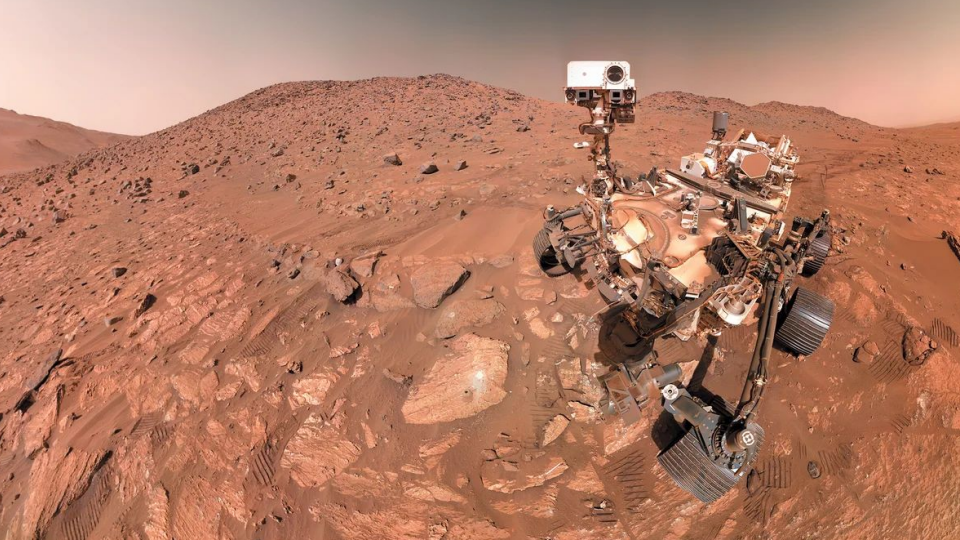
**The most efficient approach
to education**

2.

**Where LLMs perform poorly
in education**

N + 1

Approach to Education

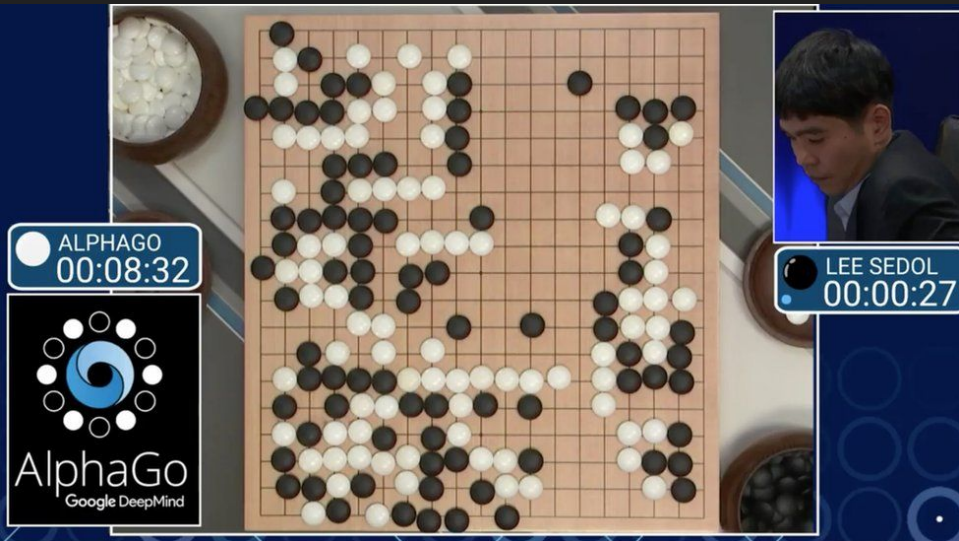


What is Reinforcement Learning?

- | | | |
|----------------------------------|-----------------------------------|------------------------------|
| 1. | 2. | 3. |
| Learning by trial and error | Agent + environment | Rewards provide feedback |
| 4. | 5. | 6. |
| Goal: maximize cumulative reward | Policy = decision-making strategy | Exploration vs. exploitation |

**Will RL get us to
a superhuman tutor?**

RL is superhuman in



2016
AlphaGo



2019
OpenAI Five





Tutor



Expertise



Tutor



Expertise

Questions, Tests



Student



Goal – Uni Maths Exam



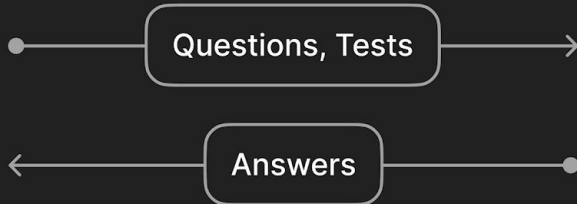
Tutor



Expertise



Model



Student



Goal – Uni Maths Exam



Tutor



Expertise



Model



Log

Questions, Tests

Answers



Student



Goal – Uni Maths Exam



Tutor



Expertise



Model



Log

Questions, Tests

Answers

Tutor understands what the student
knows and doesn't know



Student



Goal – Uni Maths Exam



Tutor



Expertise



Model



Log

Questions, Tests

Answers

Tutor understands what the student
knows and doesn't know

Tutor selects the next concept to teach



Student



Goal – Uni Maths Exam



Tutor



Expertise



Model



Log

Questions, Tests

Answers

Tutor understands what the student
knows and doesn't know

Task for a concept

Tutor selects the next concept to teach



Student



Goal – Uni Maths Exam



Tutor



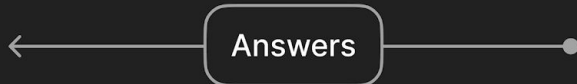
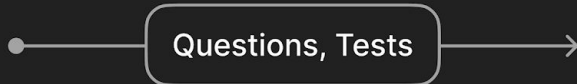
Expertise



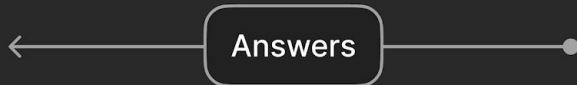
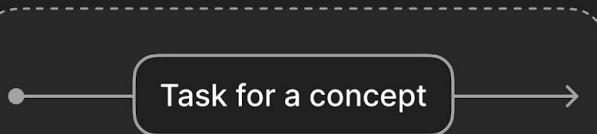
Model



Log



Tutor understands what the student
knows and doesn't know



Tutor selects the next concept to teach



Student



Goal – Uni Maths Exam



Tutor



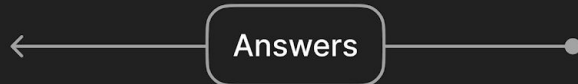
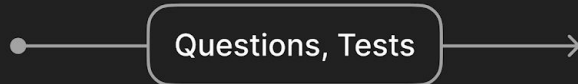
Expertise



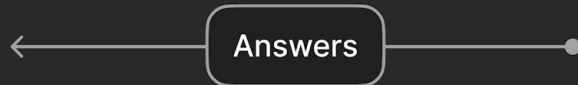
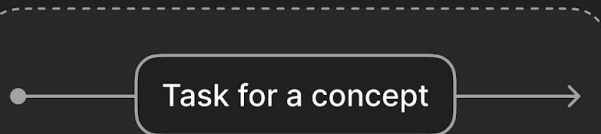
Model



Log



Tutor understands what the student
knows and doesn't know



Tutor selects the next concept to teach

× N

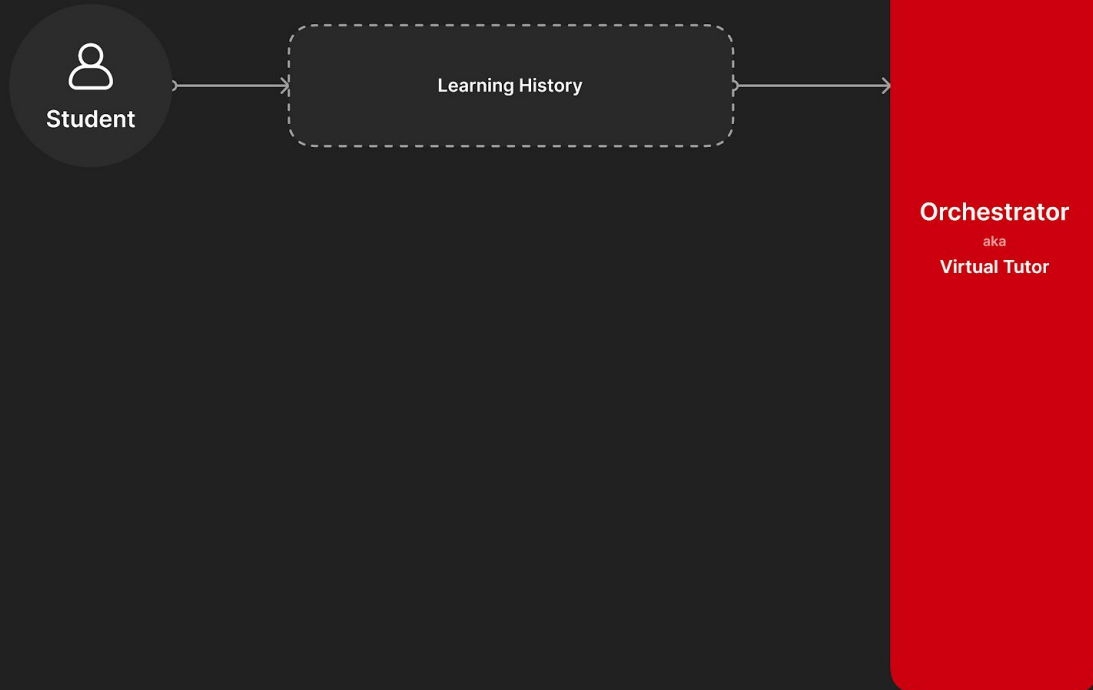


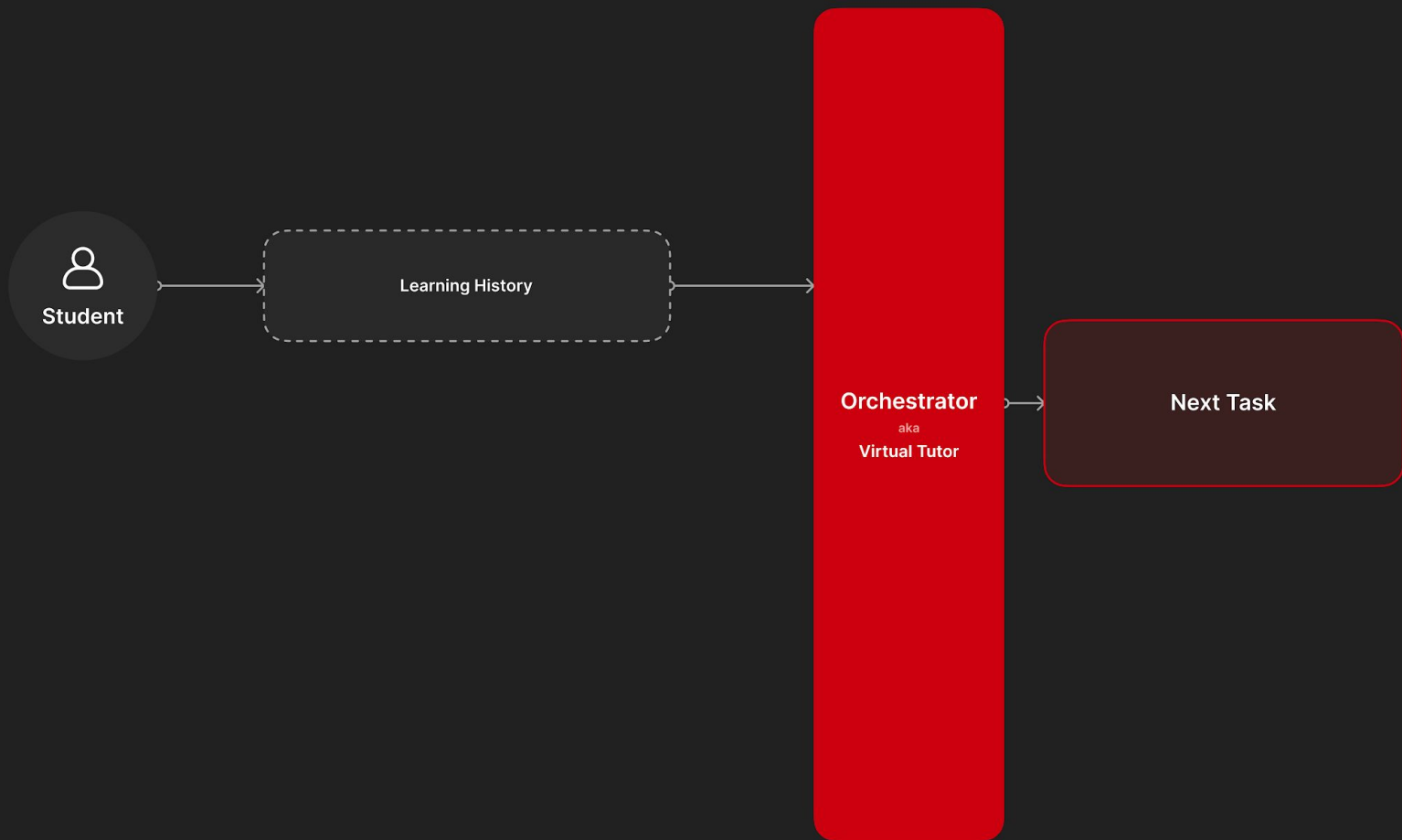
Student

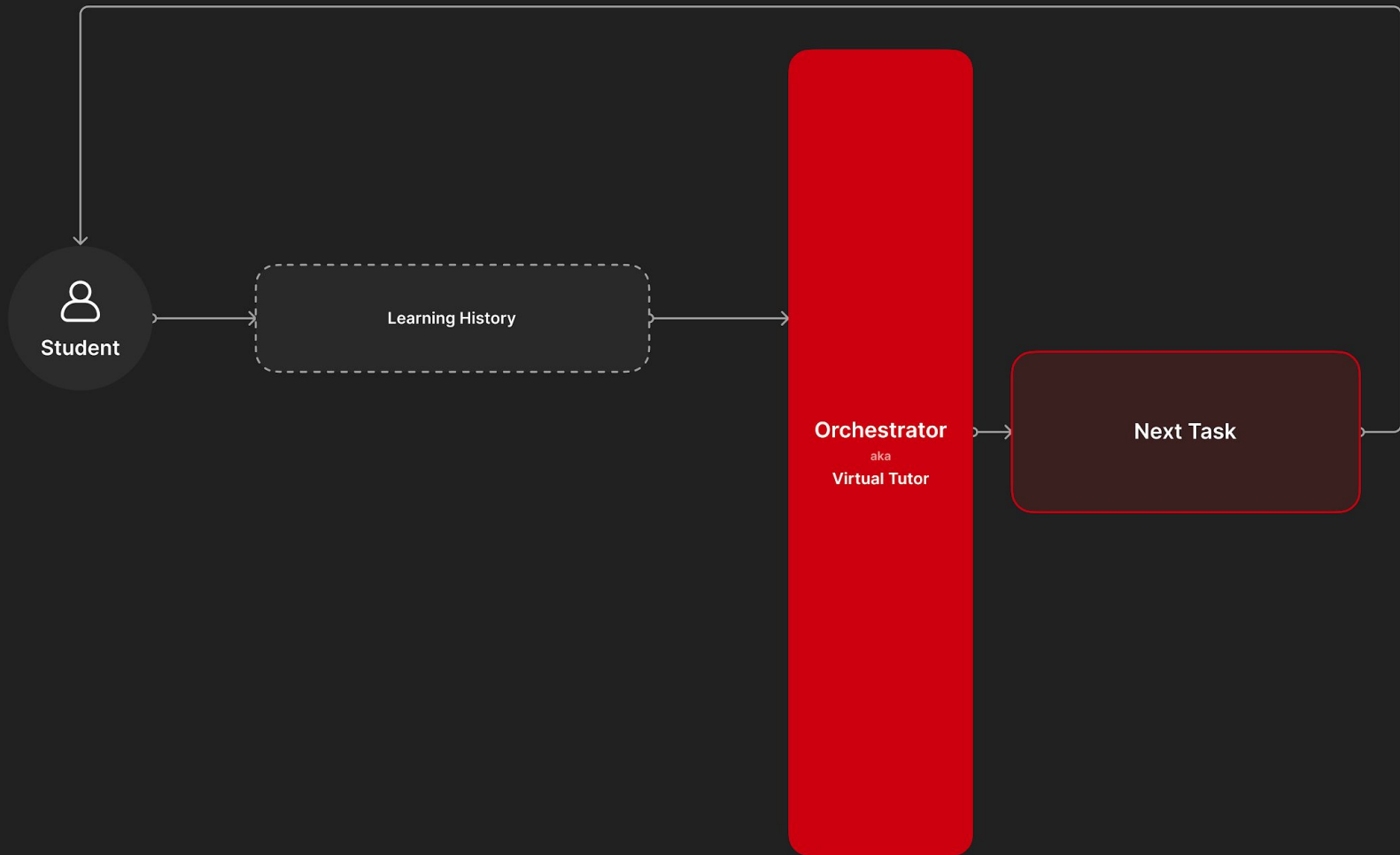


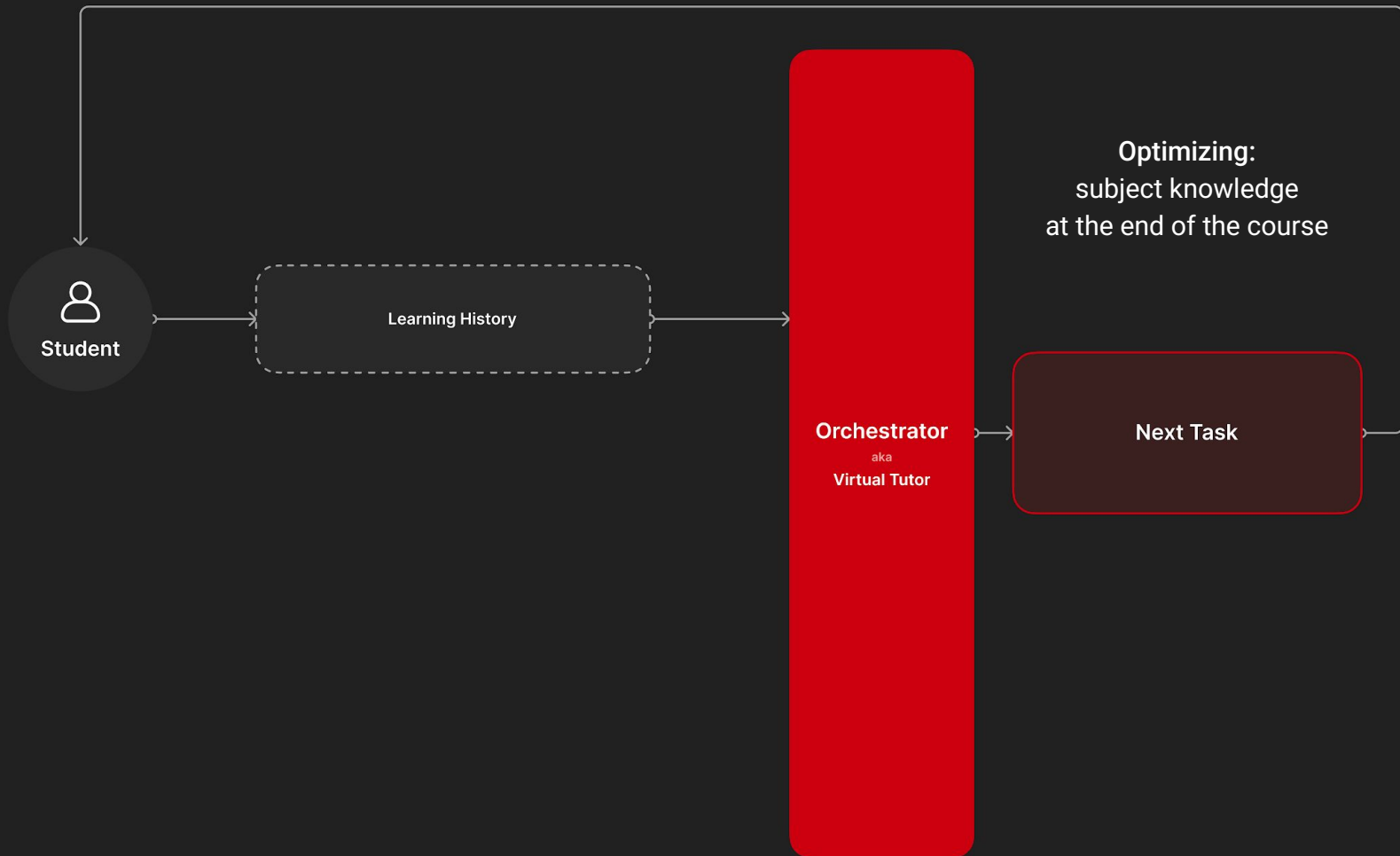
Goal – Uni Maths Exam











IELTS

~3 hours

TOEFL

~2 hours

TOEIC

~3-4 hours

Cambridge

~4 hours

~15 min Glite Quick General English Test

Many people treat being constantly online as a harmless habit, but the effects are more insidious than they first appear. Notifications fragment our attention into ever-smaller pieces, making it difficult to sustain deep concentration. Over time, this can erode our capacity for patience and reflection, leaving us restless when nothing demands our immediate response. Ironically, the same devices that promise connection can also intensify feelings of isolation. We compare our unfiltered lives to the carefully curated images of others and conclude, wrongly, that we fall short. None of this means technology is inherently damaging. Rather, its influence is subtle and cumulative: each small interruption, each glance at a glowing screen, nudges our minds toward distraction. To use technology well, we must set deliberate boundaries—moments in the day when we are fully offline—so that our attention, and perhaps even our sense of

What is the main point of the passage?

Constant online activity can quietly weaken our attention and well-being unless we control it.

Social media images are the main cause of modern loneliness.

Technology is dangerous and should be avoided whenever possible.

Continue



Listening...

What is the main focus of the conversation?

Why office noise makes work impossible

How to schedule more video meetings

How remote work affects creativity and connection

Continue

Many people say that technology has made our daily lives easier, but others think it has created new problems. Write an essay of 80–150 words in which you explain your own opinion about this. Describe how you use technology in your everyday life, and say whether you feel it helps you more than it causes difficulties. Give reasons and examples from your own experience to support your ideas.

Write your answer here...

0 words - aim for 80–150

80 more words

Many people feel that modern technology has changed our daily lives a lot. Please talk about how technology affects your own life. You could mention: (1) which devices, apps, or websites you use most often and what you use them for; (2) one or two positive effects technology has on your daily life; (3) any problems or disadvantages you have noticed, and how you try to deal with them. Give examples from your own experience.

0:00



Skip this one

YOUR ESTIMATED LEVEL

B2

You can interact fluently enough for regular conversation, and argue a point of view.

GLITE BAND · IELTS-SCALE ESTIMATE

5.0 – 7.0

a range, not a single score (± 0.71 standard error)

By section

Indicative only: each section is a few items, so the ranges are wide.

Overall language
Stronger

C2
7.5 – 9.0

Reading
Weaker - low confidence

A2
1.5 – 4.5

Listening
Weaker - low confidence

A1
1.0 – 3.0

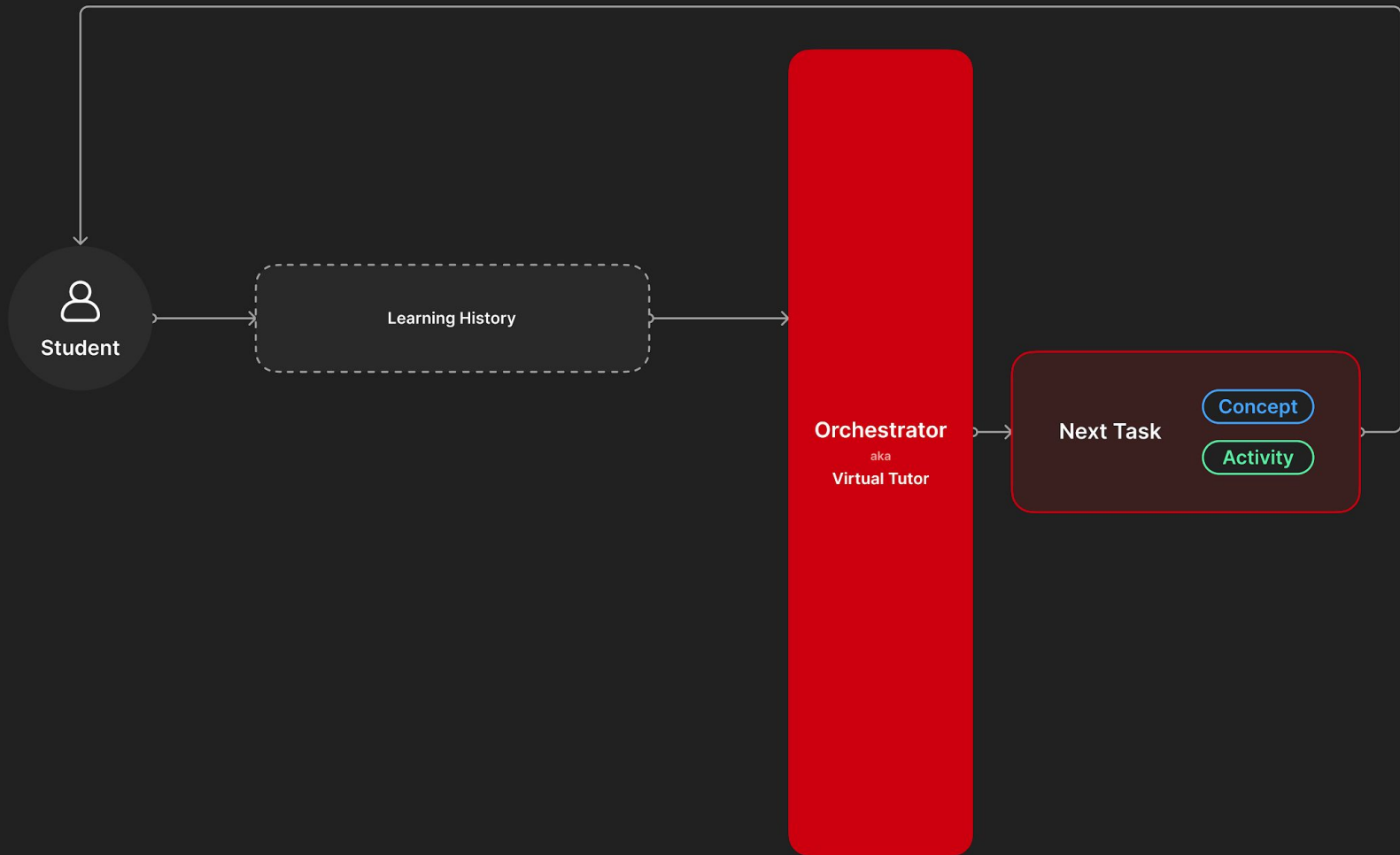
Writing

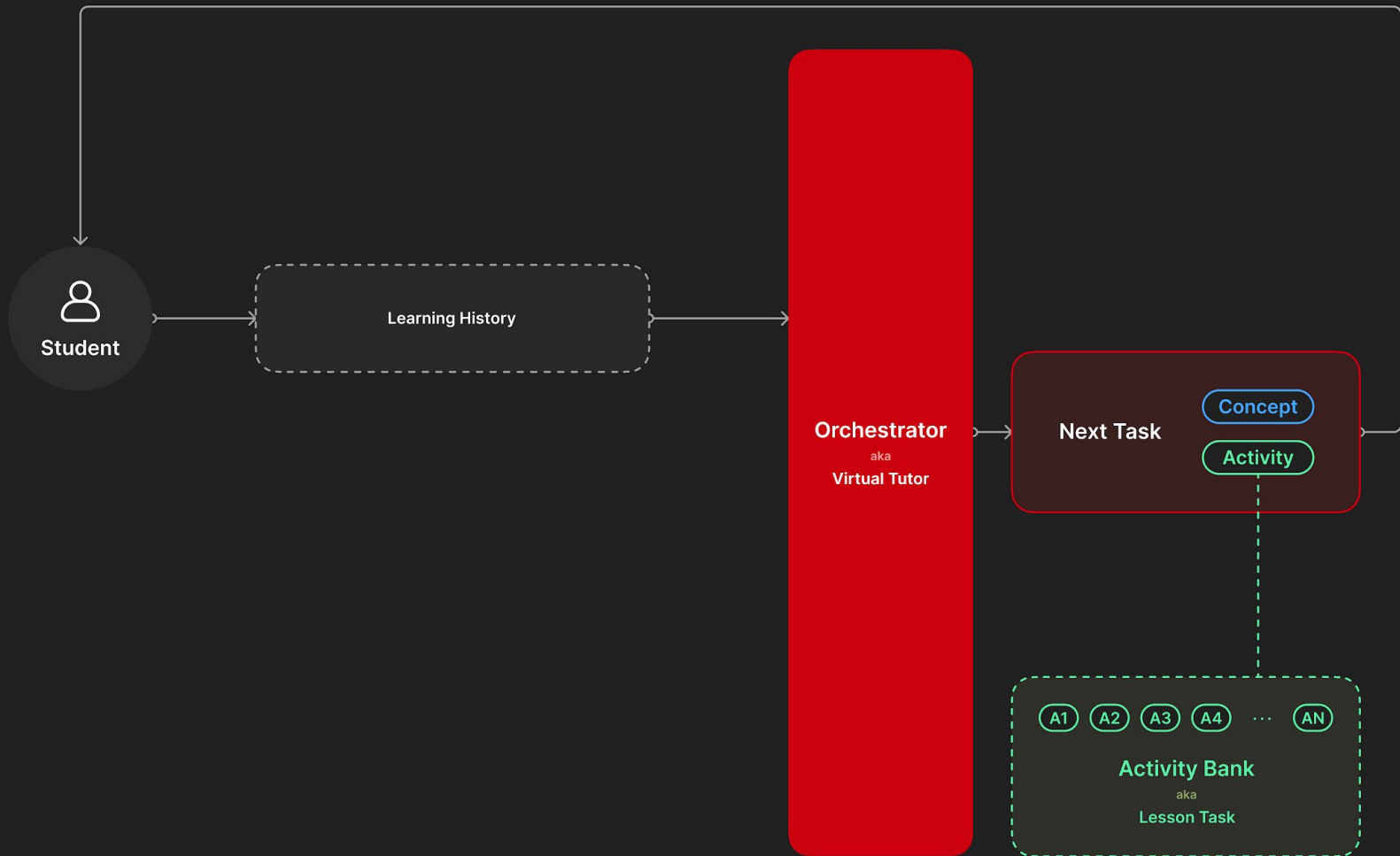
B1
4.5 – 6.0

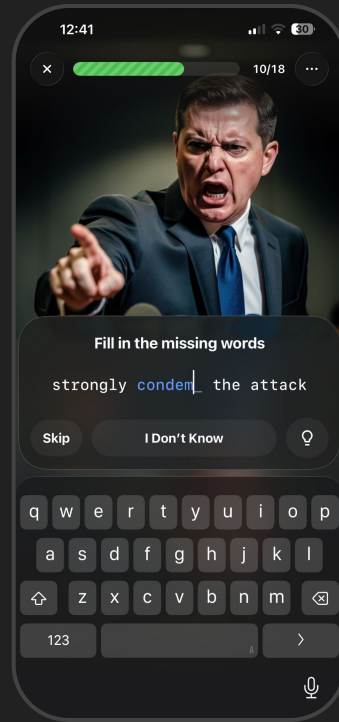
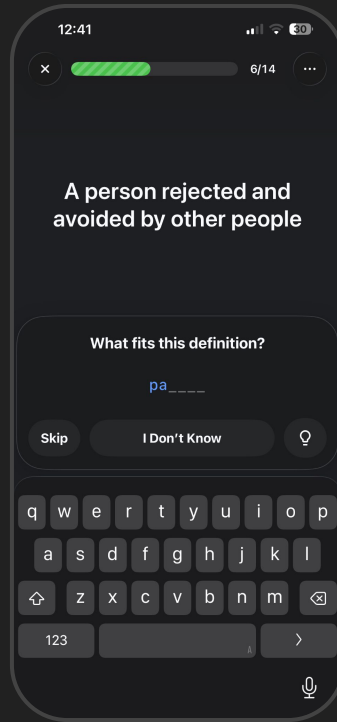
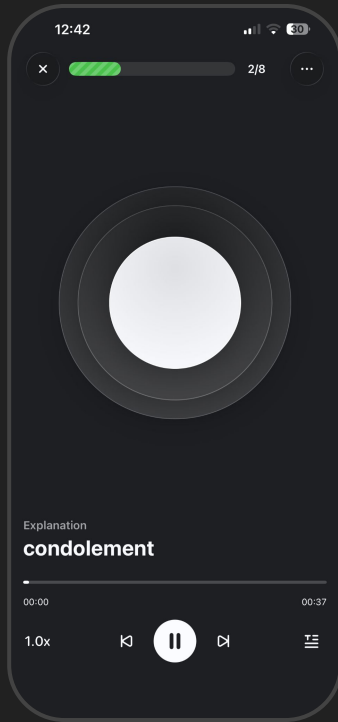
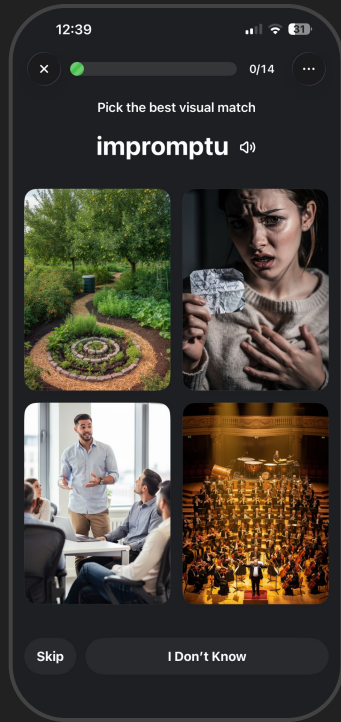
Speaking

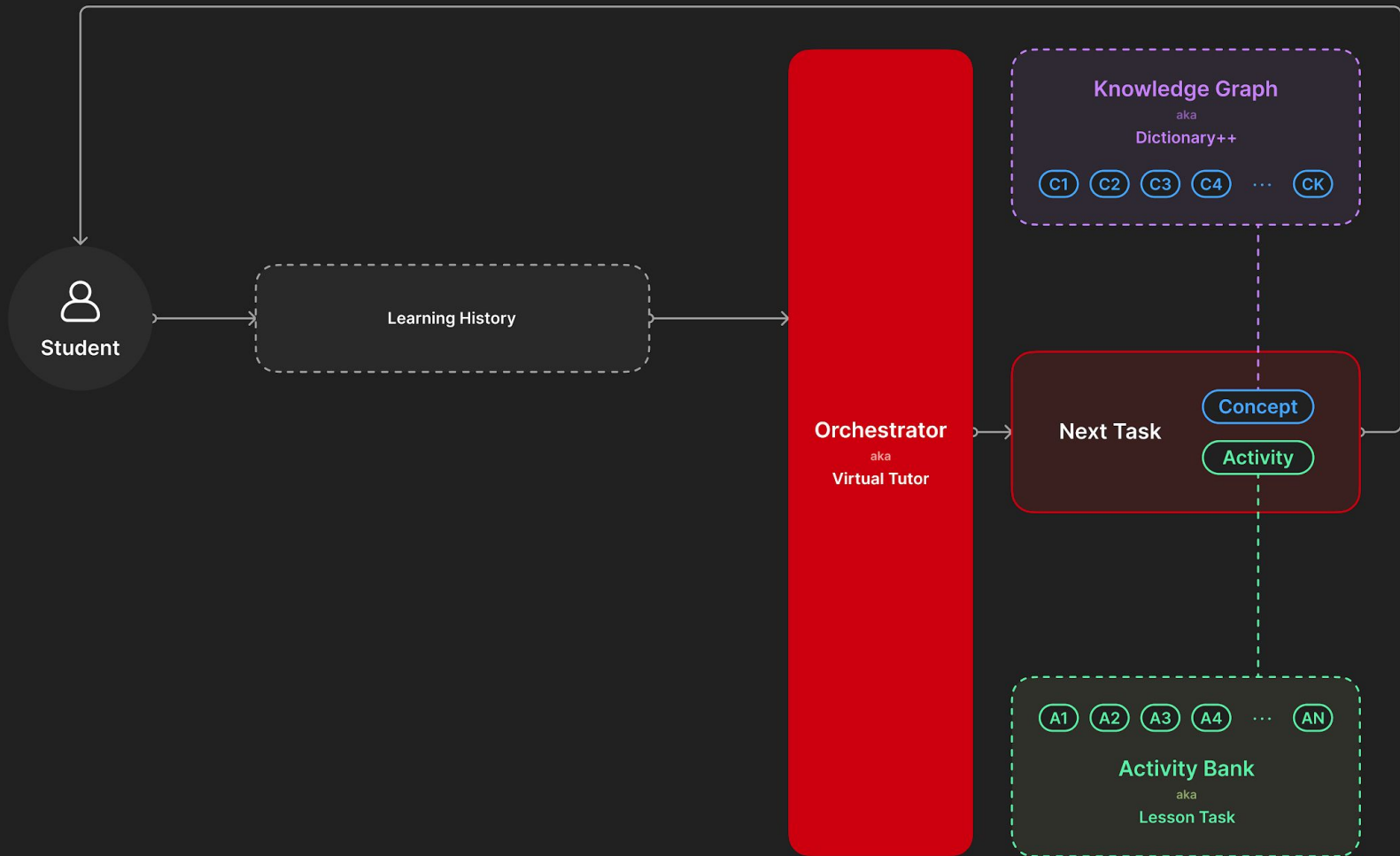
C1
7.5 – 9.0

This is a ~11-minute estimate, not an official IELTS score. IELTS is a three-hour exam and this test cannot match its precision. Use this as a rough guide to where to start, not as evidence of your level.

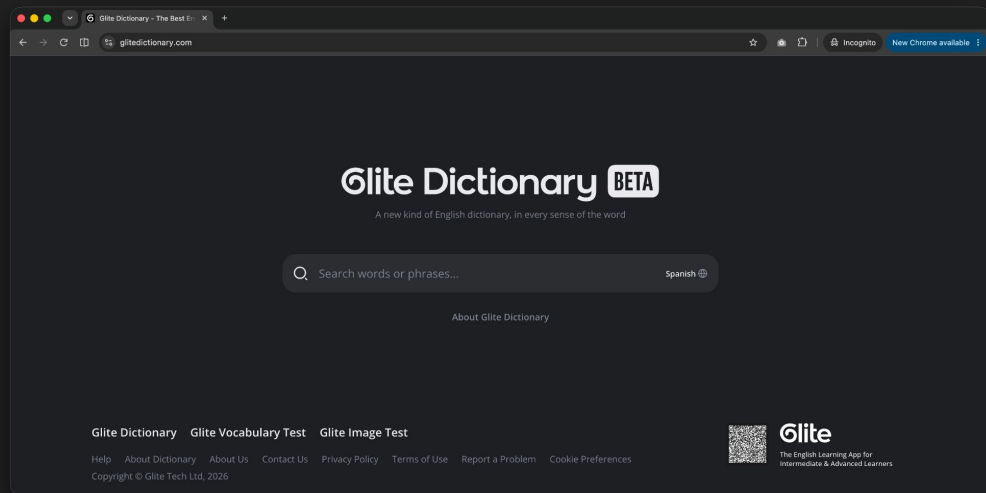








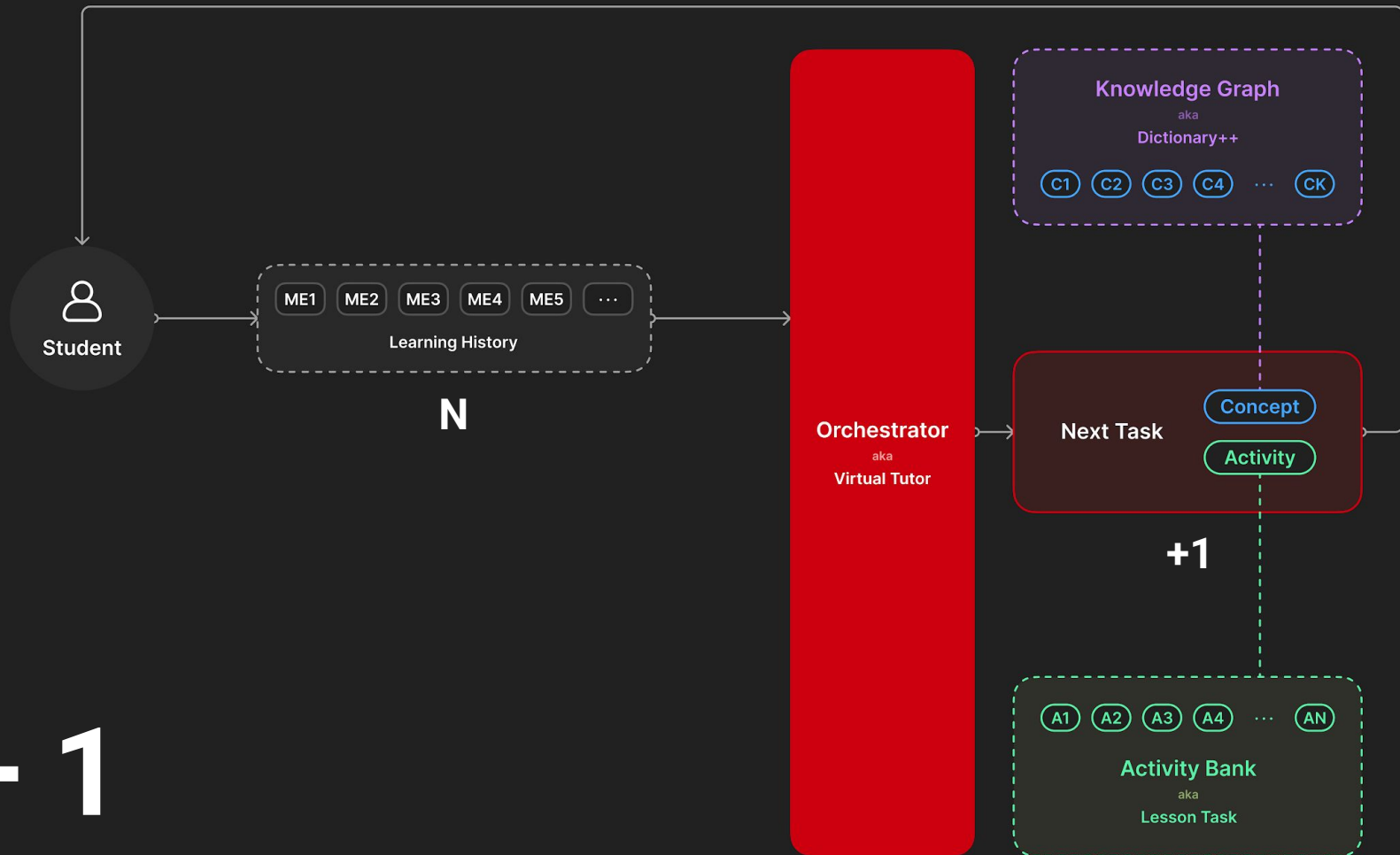
glitedictionary.com



- + Grammar Rules
- + Mistakes
- + Pronunciation concepts



N + 1



Simulations in RL

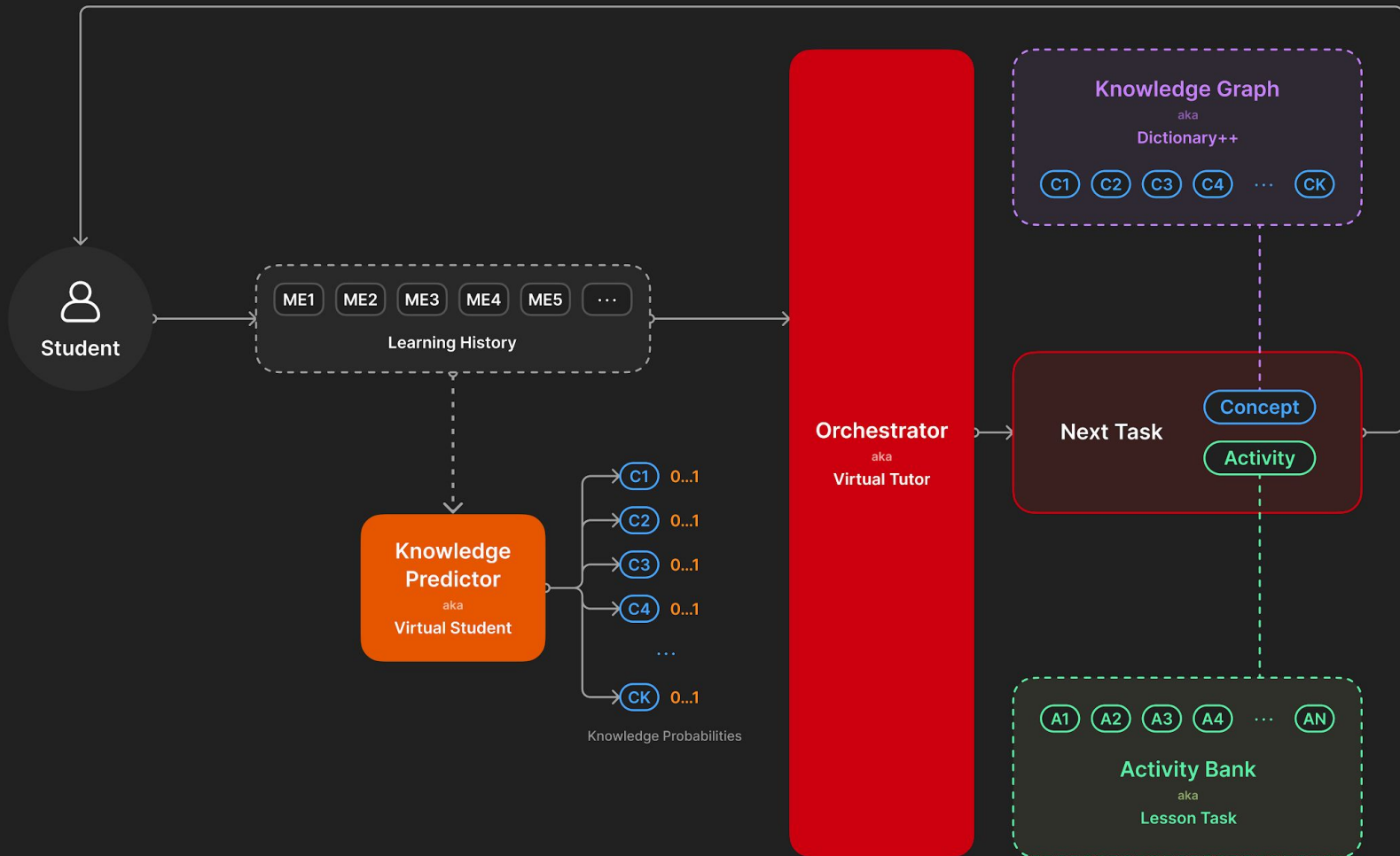
SIMULATING ROBOTS

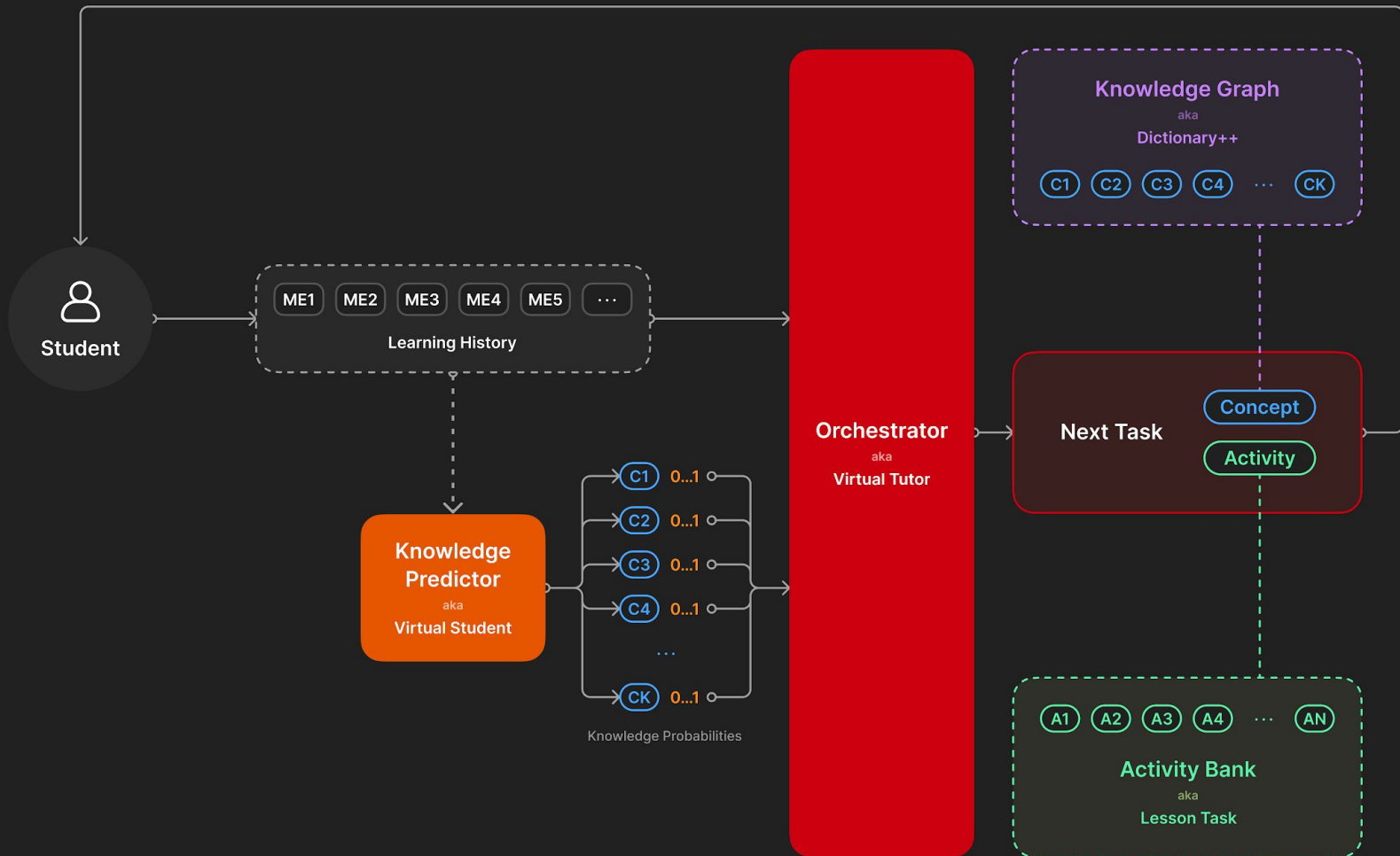


Part 2

Dagger Policy







**How well do you think LLMs
succeed in this task?**

Prompted LLMs lose to small, specialised KT models

Study	Metric	Classic KT	LLM, prompted only	LLM, fine-tuned
Neshaei 2024, 3 datasets	AUC	DKT 0.75–0.83, BKT 0.63–0.71	GPT-3.5: 0.49–0.50	GPT-3.5: 0.68–0.71
Hooshyar 2025, ASSISTments	AUC	DKT 0.83	Llama 3 8B: 0.69	Llama 3: 0.77
Bhattacharyya 2026, Eedi	Accuracy	SAKT 72.7%, DKT 71.8%	GPT-4o-mini: 58.6%	Llama-1B: 71.0%
Huang 2026, tutoring chats	AUC	Best deep KT: 0.56 / 0.63	Not tested	0.65 / 0.77

0.49–0.50

Zero-shot GPT-3.5 AUC: chance level. On Eedi, GPT-4o-mini (58.6%) is below a per-question average baseline (66.5%).

–6 pts

AUC gap between fine-tuned Llama 3 (0.77) and DKT (0.83), after about 198 hours of training.

615–12,400×

Lower inference cost for KT models than for LLMs on Eedi. Latency under 0.25 s, against 3 s or more.

+9–13 pts

Where LLMs win: fine-tuned on tutoring-chat text, over deep KT. Absolute AUC is still 0.65–0.77.

A persona prompt does not make the model know less

Study	What is measured	Reference point	Prompted LLM “student”
Balog & Bakken 2026	Reading test score, 2,359 pupils	Real pupils: 68.7%	95.6–98.5%
Balog & Bakken 2026	Score-distribution gap, JSD	Real 0.006, random 0.52	0.63–0.77
Srivatsa et al. 2025	NAEP percentile when told the grade	Average student: 50th	o3-mini ~99th, Qwen-7B 18th
Yang et al. 2026	One learner: chess / L2 / maths	Naive: 0.45 / 0.51 / 0.49	GPT-5.4: 0.23 / 0.51 / 0.61
Duan et al. 2026	Same misconception as the student	History + RL: 15%	GPT-5.4: 8–11%

96–99%

Score of six models told to act as 5th graders; real pupils got 69%. Random guessing matches the score spread better.

<1%

Change in accuracy when 20+ models take on low, medium or high proficiency personas (Li et al., ACL 2026).

$\rho = 0.34$

GPT-5's correlation with real item difficulty, below GPT-4.1 (0.44). Stronger models simulate worse.

0.7 → 0.09

What works: JSD after a working-memory limit is built in. Constraint and per-student training help; prompts do not.

How can learners be exposed to the concepts?

How can learners be exposed to the concepts?

NETFLIX



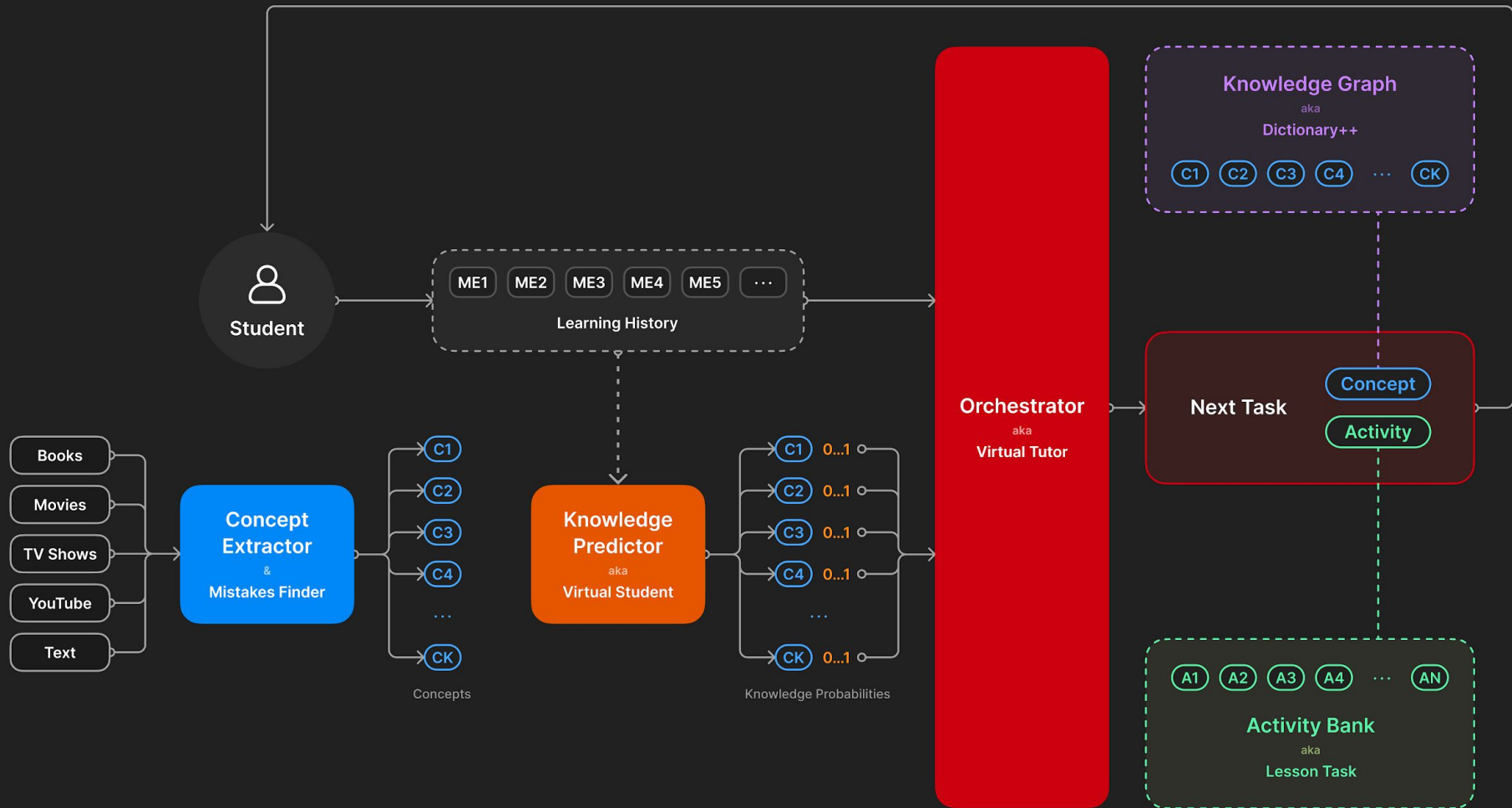
Content

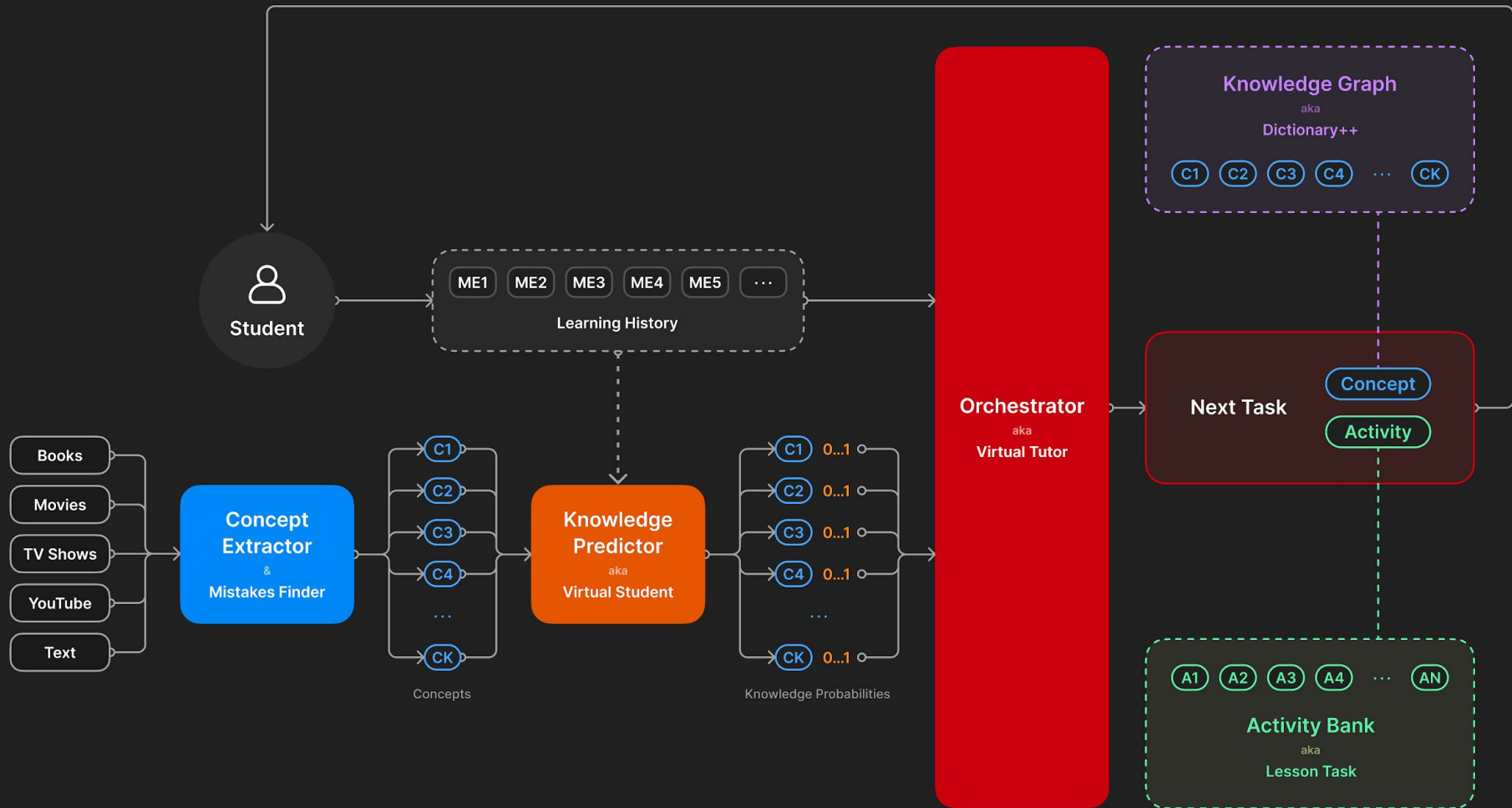


amazonkindle



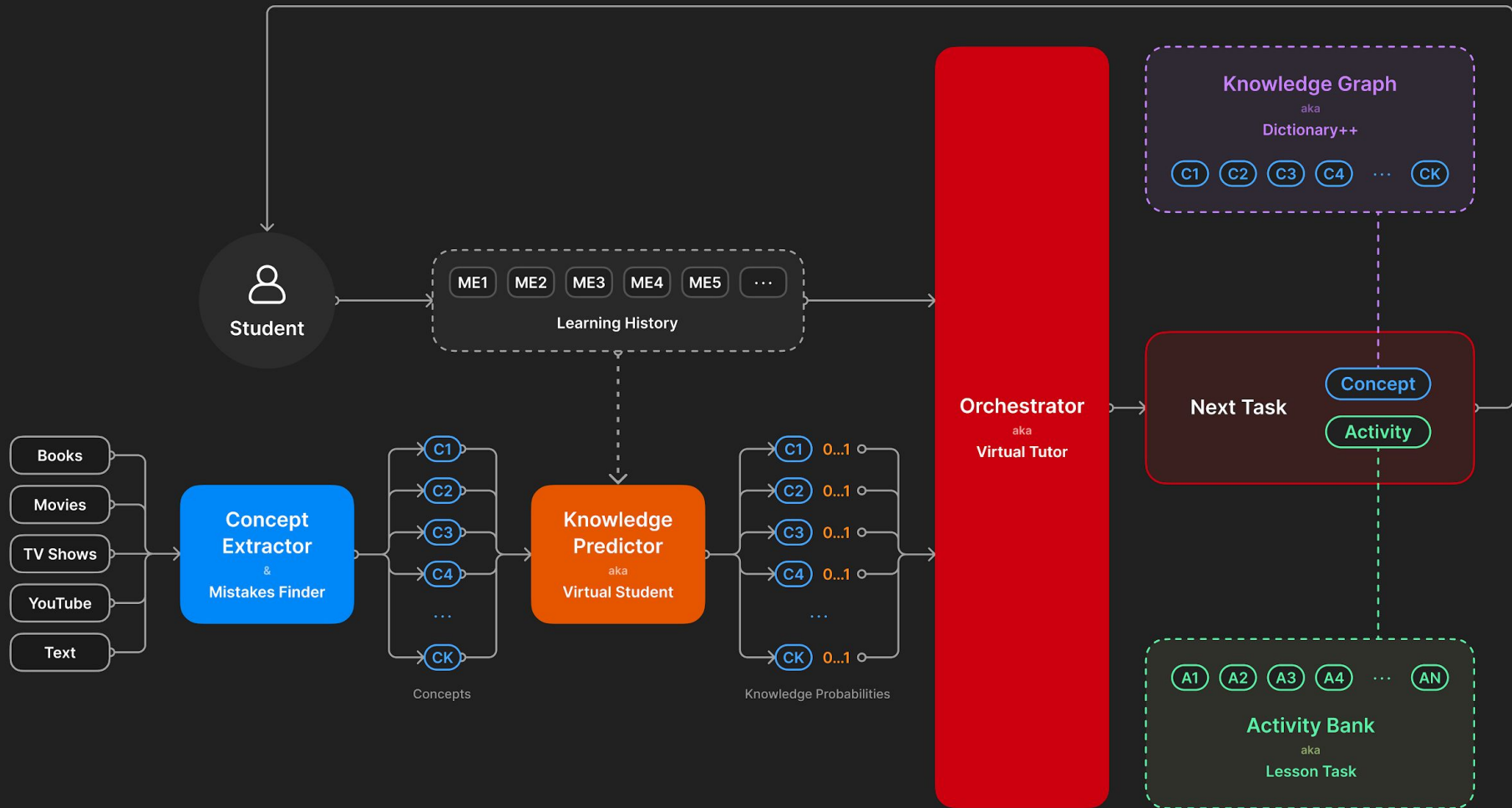
YouTube





N + 1

Approach to Education



Quiz



Pavel Katunin

Co-founder, CPTO

linkedin.com/in/pavelkatunin

pavel@glite.ai



Slides Deck



How do we automatically find concepts in learning materials and student answers?

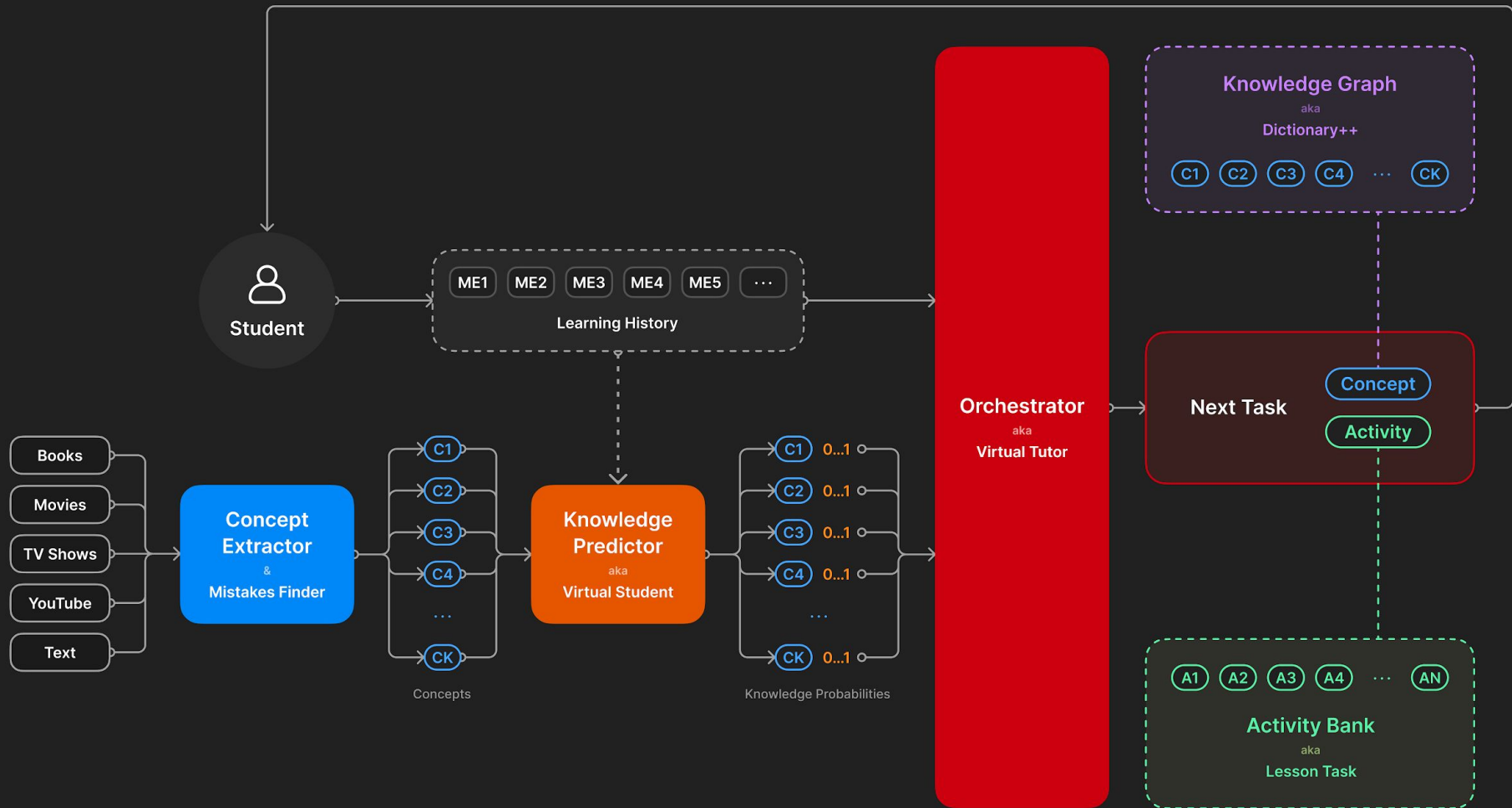


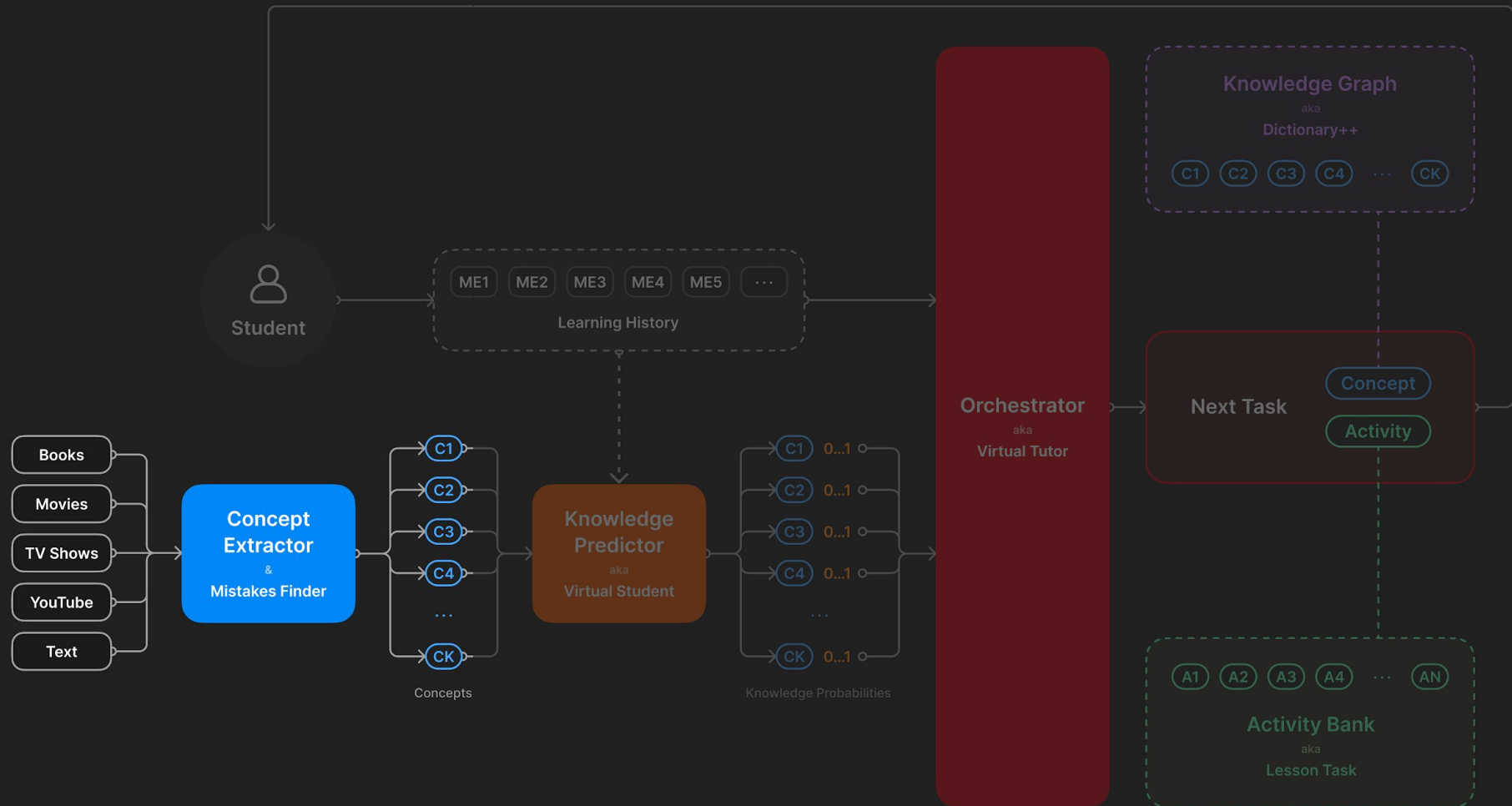
Vassili Philippov

Co-founder, CEO Glite Tech

[linkedin.com/in/vassiliphillippov/](https://www.linkedin.com/in/vassiliphillippov/)

vassili@glite.ai

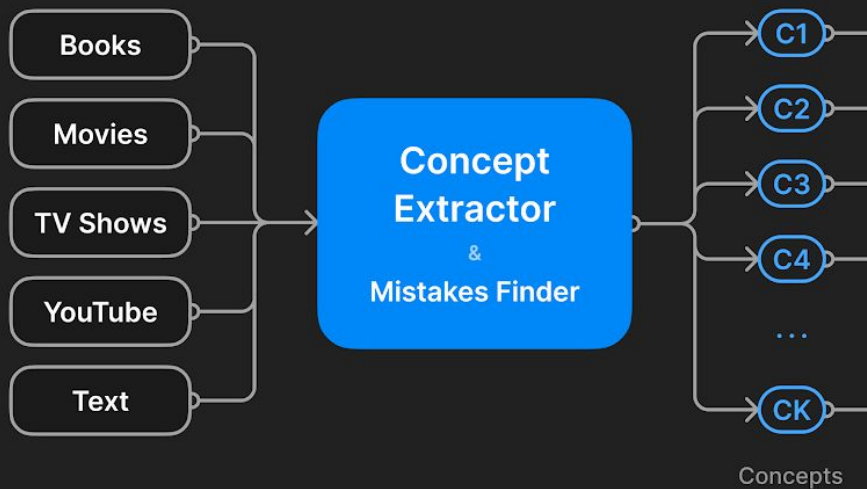




Find concepts here?

Kahoot time

Why do you need concept extractor?



Imagine, you already have this. **How would you use it?**

Why do you need concept extractor?

1. Find mistakes

2. Indexing
materials

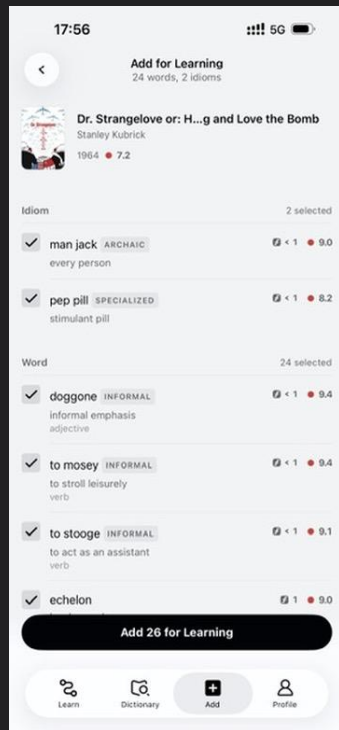
3. Importance/
frequency

4. Spontaneous
practice

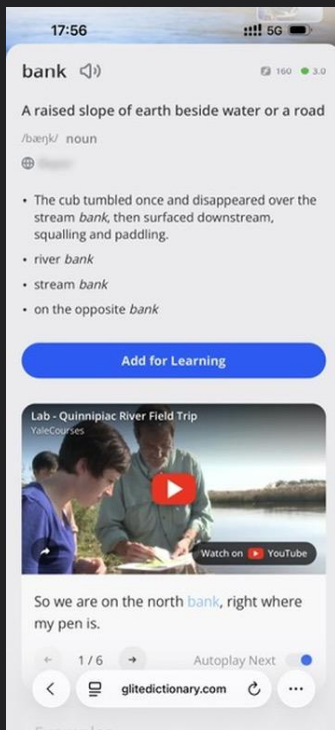
5. Cross subject

6. Check LLM tasks

Concept Extractor @ Glite



Words from movies



Video examples

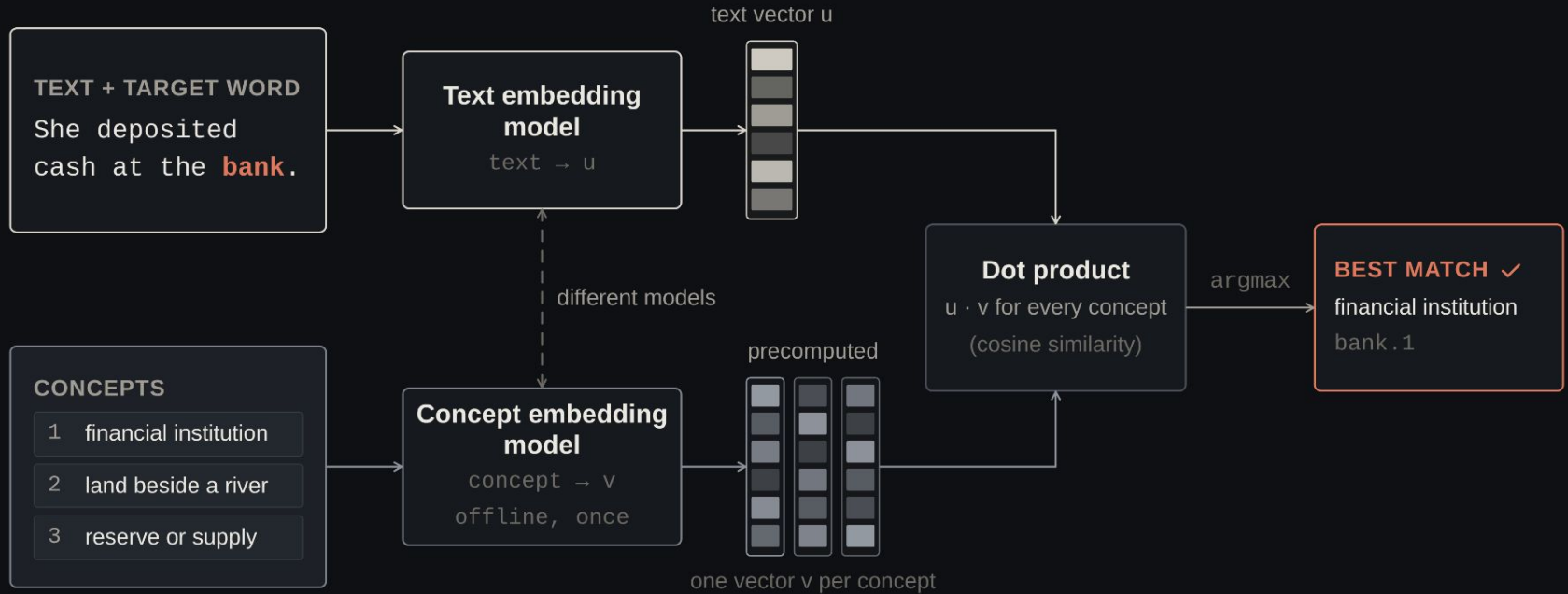


Importance

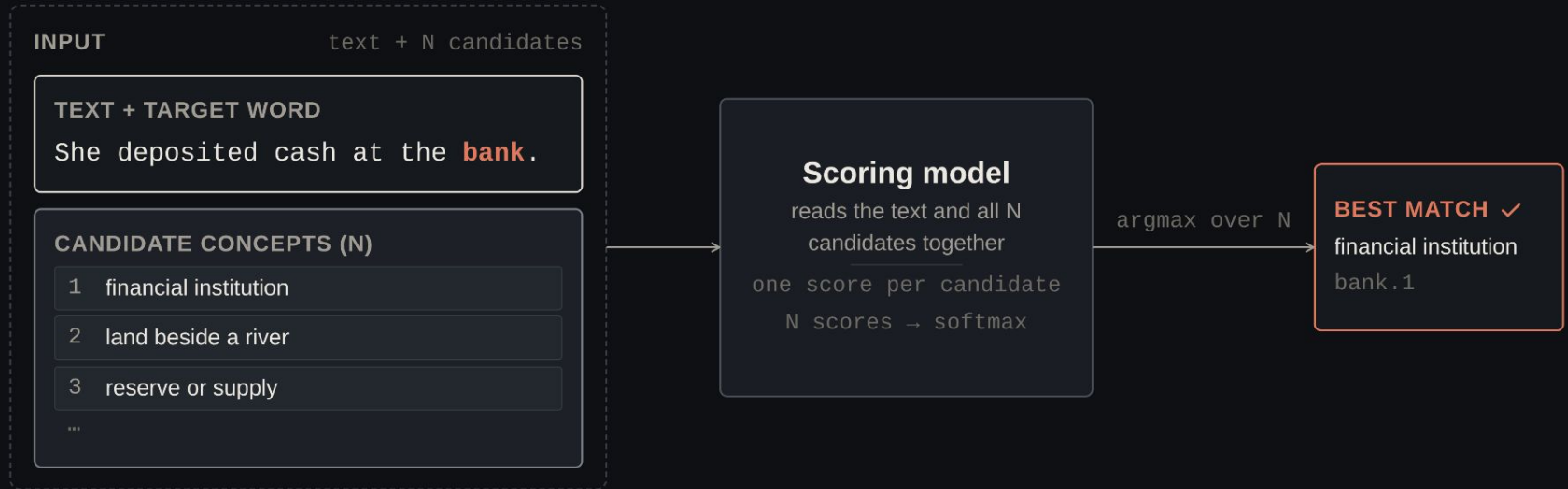
What you will learn?

How to create State Of The Art concept extractor for English, Maths and other subjects

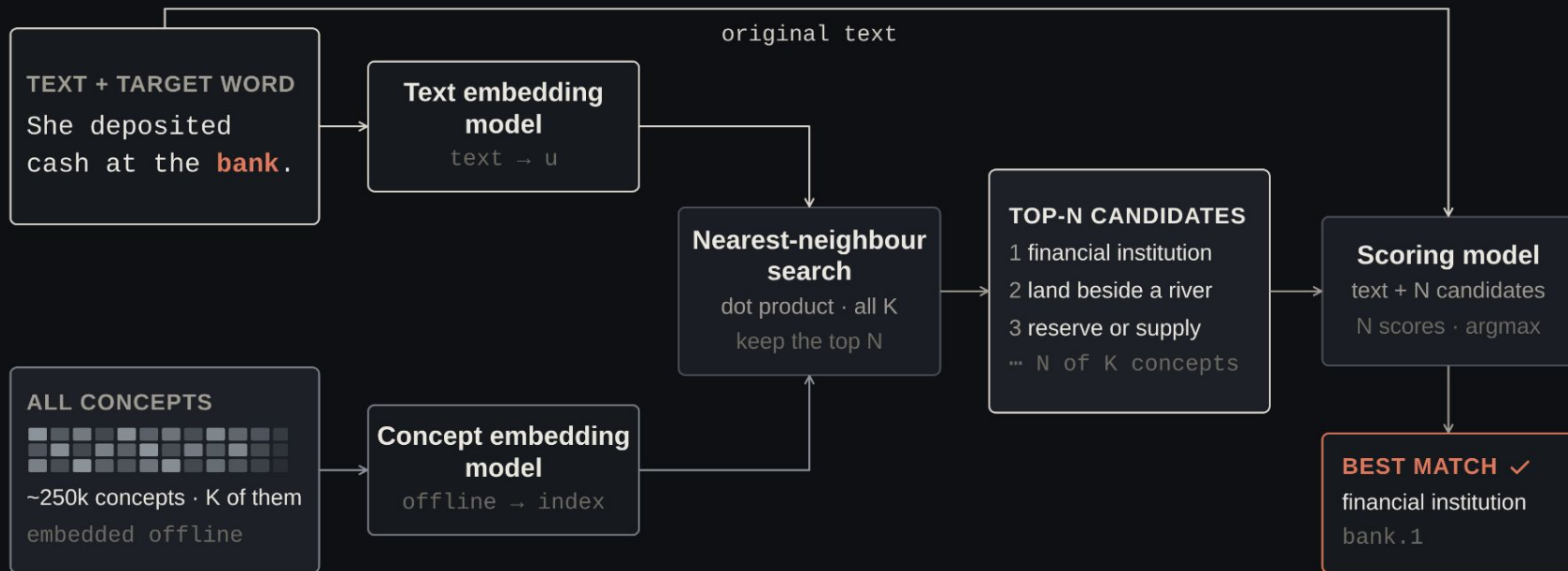
Bi-Encoder



Cross-Encoder



Hybrid



What about LLMs?

arXiv

Search Submit Donate Log in

Computer Science > Computation and Language

[Submitted on 19 Jul 2026]

English Word Sense Disambiguation in 2026: When the Labels Become the Bottleneck

Vassili Philippov, Amro Salman, Dmitrii Andreev, Penny Hands, Emil Kaiumov, Pavel Katunin, Anton Nikolaev

In English all-words word sense disambiguation (WSD), the labels, not the models, have become the bottleneck: frontier LLMs are accurate enough that the errors surviving in the gold standard decide benchmark rankings -- in the test sets we score on and, as we show causally, in the corpus we train on. We release lexEN, a WSD evaluation benchmark built as a conservative, human-adjudicated correction layer over Maru2022's ALL_NEW benchmark (211 labels changed, 56 removed), and SenseBench, an auditable LLM WSD evaluation harness and living leaderboard (57 models, 192 runs). The task is inventory-constrained multiple choice (the model picks from the supplied WordNet senses), so the reported accuracies are a ceiling on what models achieve without that help. On lexEN-v1 the frontier LLMs converge near 95% (best, 95.6%), the top three families are statistically indistinguishable, and accuracy trades off against reasoning effort and cost across a ~2,500x price span. Relabeling SemCor with frontier models and retraining BEM, ESCHER, and ConSeC unchanged lifts them by several F1 points on test sets the relabeling never touched; we release the relabeled corpora and Glite LENS, a 298M bi-encoder trained on the repaired labels -- to our knowledge the strongest reported (83.6 Raganato ALL, 87.4 Maru ALL_NEW) -- serving at ~\$0.13 per million items. On hard items, fine-grained WordNet senses are partly ill-posed even for experts (three-reviewer Fleiss kappa=0.537); coarsening raises annotator agreement and model accuracy together across four inventories, placing a top model inside the expert agreement band at coarse granularity (statistically equivalent under three of four) but significantly below it at fine. The binding constraint is now cost.

Comments: 48 pages, 10 figures. Code, data, and live leaderboard: [this https URL](#)

Subjects: **Computation and Language (cs.CL)**

Cite as: [arXiv:2609.17554](#) [cs.CL]
(or [arXiv:2609.17554v1](#) [cs.CL] for this version)
<https://doi.org/10.48550/arXiv.2609.17554>

Submission history

From: Vassili Philippov [\[view email\]](#)

[v1] Sun, 19 Jul 2026 23:24:30 UTC (463 KB)

Access Paper:

[View PDF](#)
[HTML \(experimental\)](#)
[TeX Source](#)
view license

Current browse context:
cs.CL
[< prev](#) | [next >](#)
[new](#) | [recent](#) | [2026-09](#)
Change to browse by:
[cs](#)

References & Citations
[NASA ADS](#)
[Google Scholar](#)
[Semantic Scholar](#)

[Export BibTeX Citation](#)

Bookmark

GPT-5-mini	82.2
gemma-4-26B	81.8
GPT-4o-mini	76.2

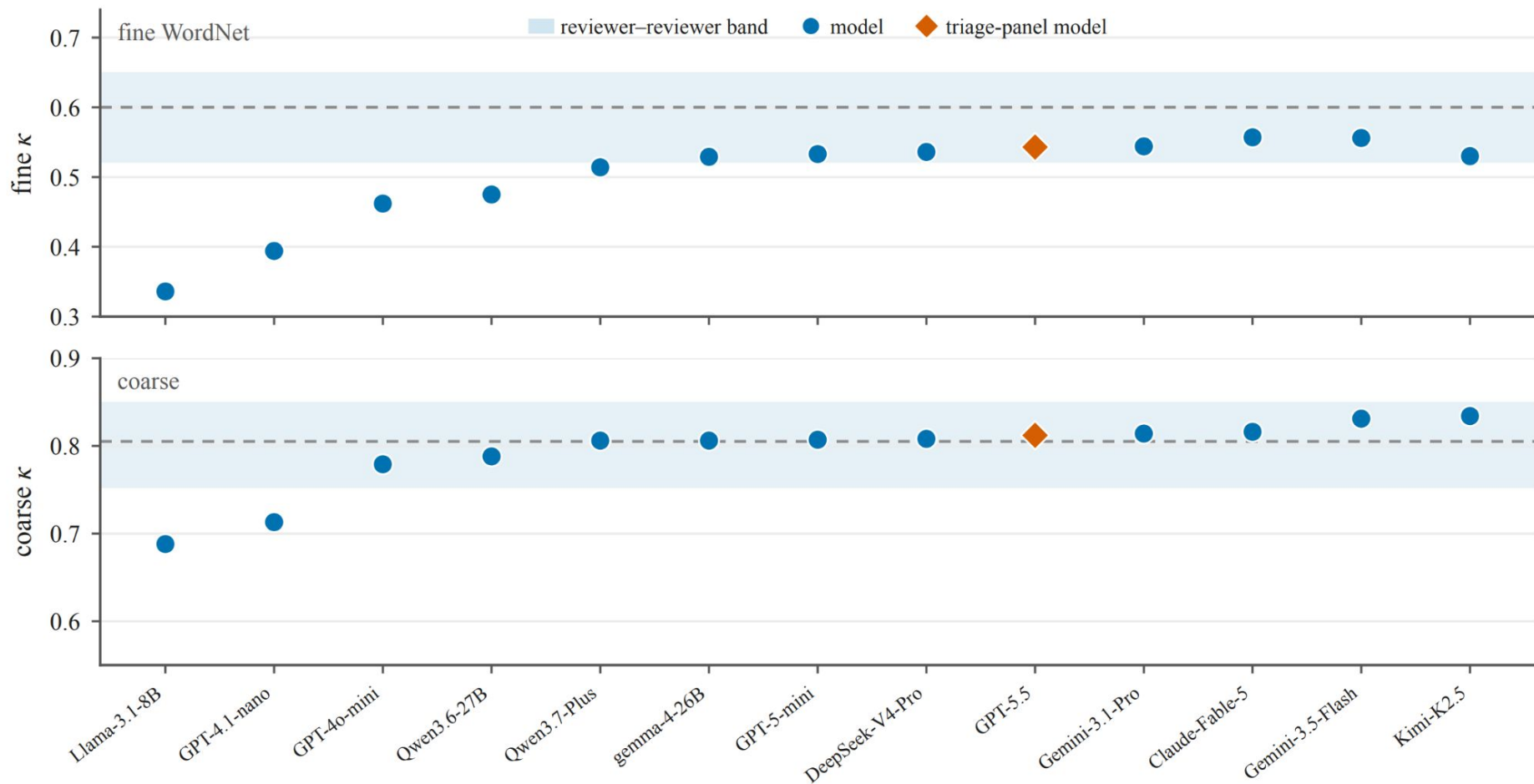
ConSeC	81.8
ESCHER	77.7
BEM	76.6
SANDWiCH	86.6
MFS	58.9

System	Raganato
GPT-5.5	85.4
Gemini-3.1-Pro	85.2
Claude-Fable-5	85.6
GPT-5-mini	82.2
gemma-4-26B	81.8
GPT-4o-mini	76.2
ConSeC	81.8
ESCHER	77.7
BEM	76.6
SANDWiCH	86.6
MFS	58.9

System	Raganato	Maru2022
GPT-5.5	85.4	91.8
Gemini-3.1-Pro	85.2	92.1
Claude-Fable-5	85.6	91.9
GPT-5-mini	82.2	88.1
gemma-4-26B	81.8	87.2
GPT-4o-mini	76.2	81.0
ConSeC	81.8	84.2
ESCHER	77.7	80.8
BEM	76.6	79.3
SANDWiCH	86.6	86.1
MFS	58.9	61.3

System	Raganato	Maru2022	lexEN
GPT-5.5	85.4	91.8	95.6
Gemini-3.1-Pro	85.2	92.1	94.9
Claude-Fable-5	85.6	91.9	95.2
GPT-5-mini	82.2	88.1	90.7
gemma-4-26B	81.8	87.2	89.4
GPT-4o-mini	76.2	81.0	82.7
ConSeC	81.8	84.2	84.9
ESCHER	77.7	80.8	81.4
BEM	76.6	79.3	79.7
SANDWiCH	86.6	86.1	85.2
MFS	58.9	61.3	61.6

Model $\leftrightarrow$ reviewer agreement vs. the human band (307 hard items)



Relabeling of SemCor Using LLM

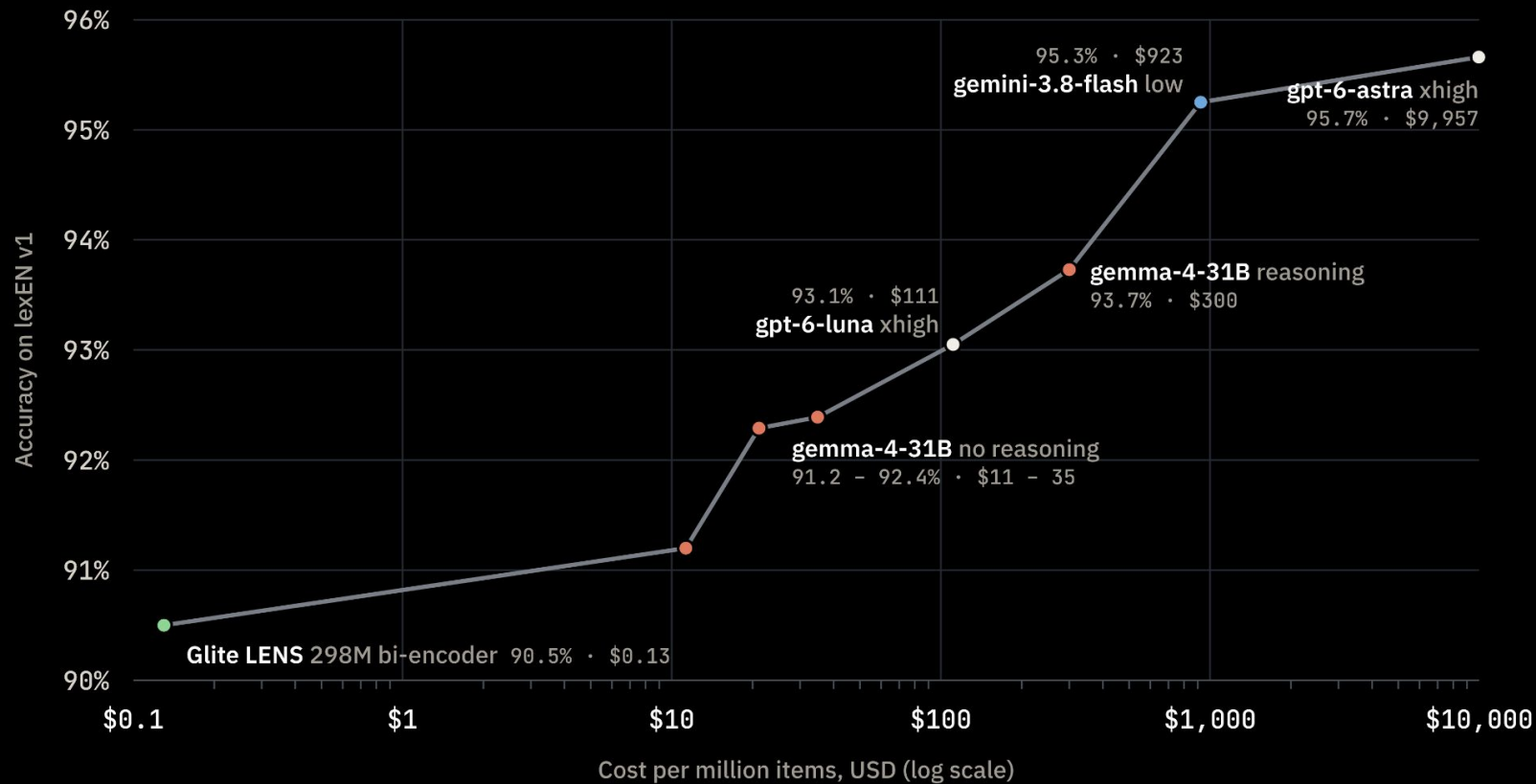
	Traditional SemCor
BEM	78.9%
ESCHER	81.3%
ConSeC	83.3%

Relabeling of SemCor Using LLM

	Traditional SemCor	GPT-5.5-relabeled
BEM	78.9%	86.7% +7.8%
ESCHER	81.3%	88.1% +6.8%
ConSeC	83.3%	90.7% +7.4%

SenseBench Pareto frontier

● Glite ● Gemma ● GPT ● Gemini



~50,000x

cheaper than top LLMs

~100x

cheaper than cheap LLMs

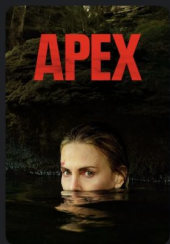
Is Price an issue?

Movies: 37K movies

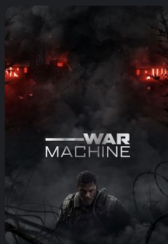
Guess the price

Movies: 37K movies, 200M words

Top LLM: \$1M, Cheap LLM: \$2K, bi-encoder: \$26



Apex
2026
● 7.2



War Machine
2026
● 6.1



Disclosure Day
2026
● 5.1



A Marvel Television...
2026
● 6.2



Tom Clancy's Jack R...
2026
● 5.1



Black Phone 2
2025
● 6.6



Mortal Kombat II
2026
● 5.9



War of the Worlds
2025
● 6.5



Playdate
2025
● 5.5



Predator: Badlands
2025
● 5.6



Project Hail Mary
2026
● 6.1



Shelter
2026
● 6.6



Code 3
2025
● 5.7



The Super Mario Gal...
2026
● 5.9



Office Romance
2026
● 5.9



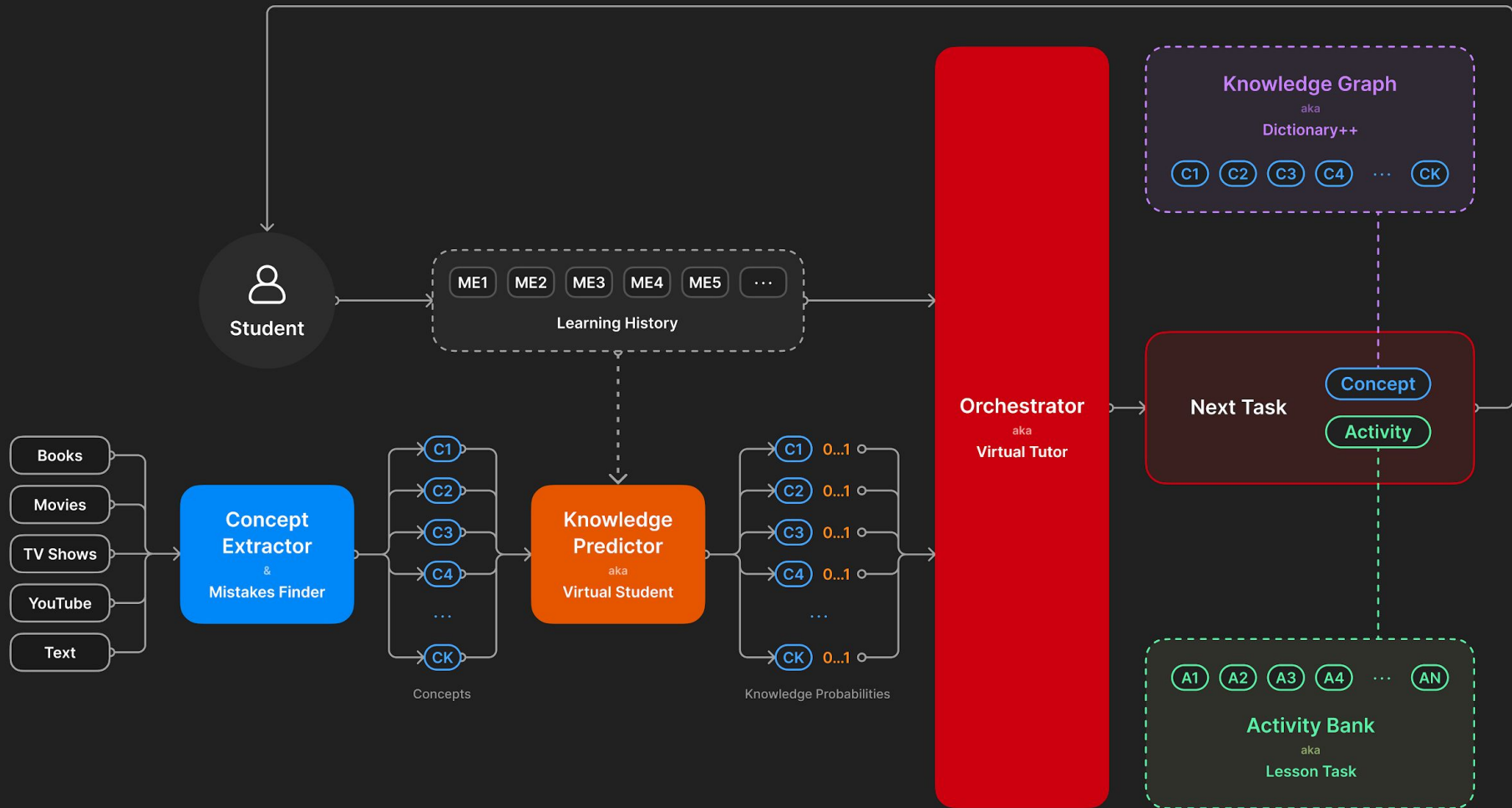
Balls Up
2026
● 6.6



The Bluff
2026
● 7.4



Your Fault: London
2026
● 5.1



TV Series: 78K episodes, 300M words

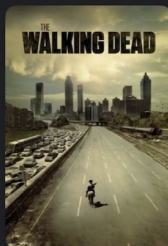
Top LLM: \$1.5M, Cheap LLM: \$3K, bi-encoder: \$40



Game of Thrones
Season 1
● 8.0



Stranger Things
Season 1
● 5.8



The Walking Dead
Season 1
● 6.5



The Good Doctor
Season 1
● 7.2



WandaVision
Season 1
● 6.6



The Big Bang Theory
Season 1
● 6.8



Breaking Bad
Season 1
● 5.9



Lucifer
Season 1
● 5.9



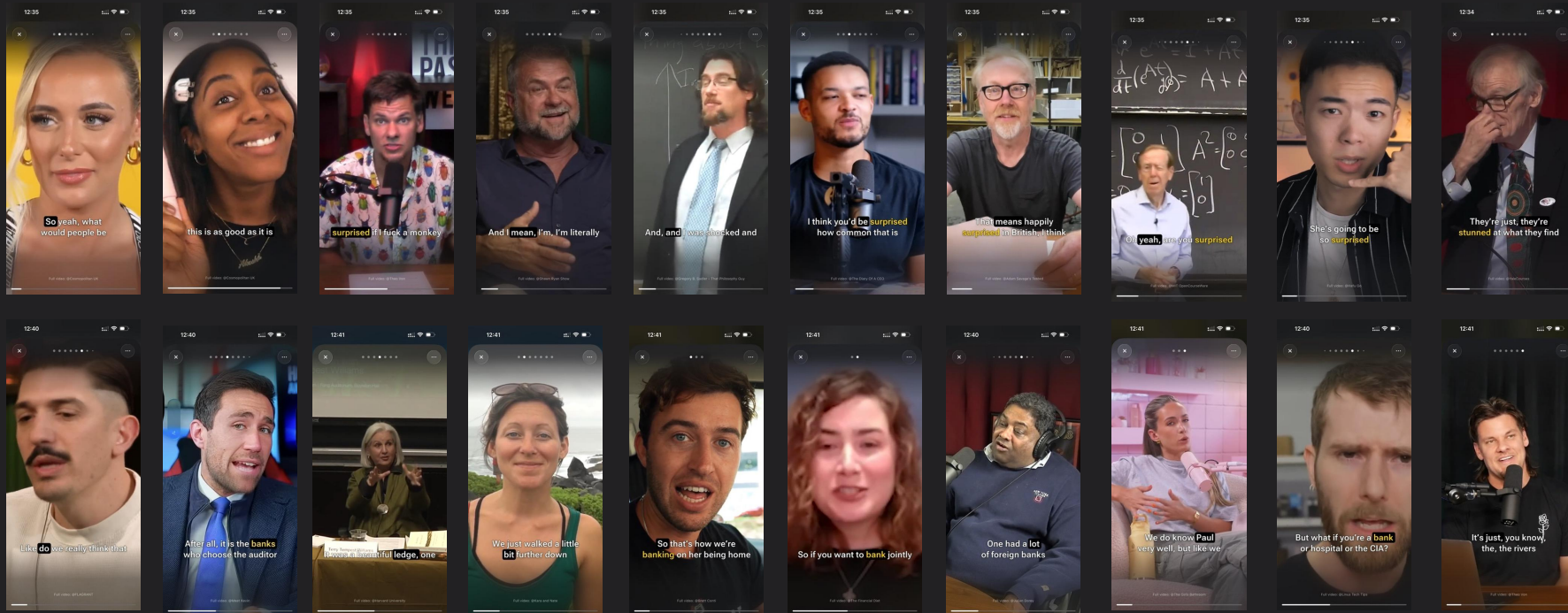
Riverdale
Season 1
● 6.0



Loki
Season 1
● 6.1

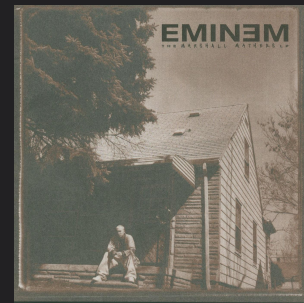
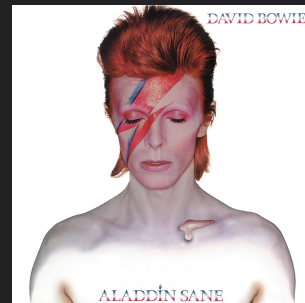
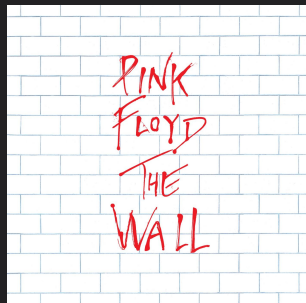
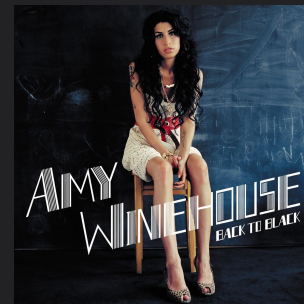
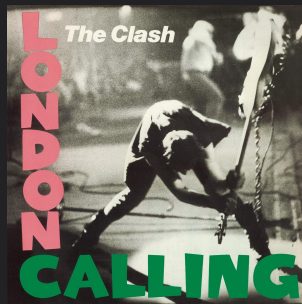
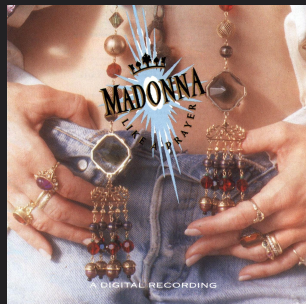
YouTube: 200K videos, 900M words

Top LLM: \$4.5M, Cheap LLM: \$9K, bi-encoder: \$120



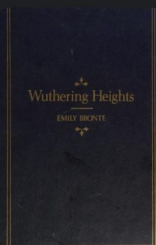
Songs: 7M songs, 2B words

Top LLM: \$10M, Cheap LLM: \$20K, bi-encoder: \$260

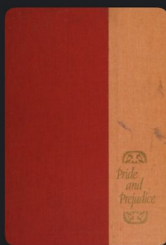


Books: 57K books, 3B words

Top LLM: \$15M, Cheap LLM: \$30K, bi-encoder: \$390



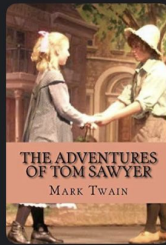
Wuthering Heights
Brontë, Emily
● 7.0



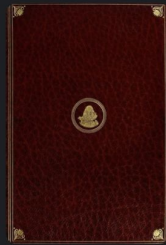
Pride and Prejudice
Austen, Jane
● 6.0



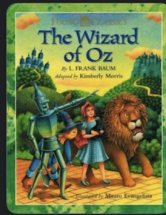
The Iliad
Homer
● 8.5



The Adventures of T...
Twain, Mark
● 5.0



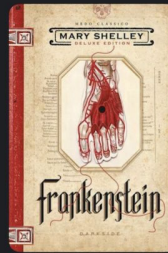
Alice's Adventures i...
Carroll, Lewis
● 5.0



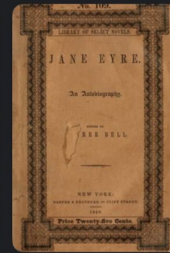
The Wonderful Wizar...
Baum, L. Frank (Lyma...
● 4.5



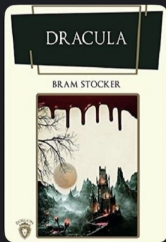
The Time Machine
Wells, H. G. (Herbert...
● 5.5



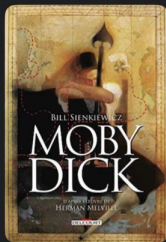
Frankenstein
Shelley, Mary Wollsto...
● 6.0



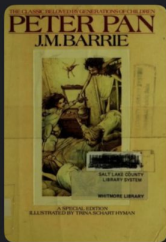
Jane Eyre
Brontë, Charlotte
● 6.5



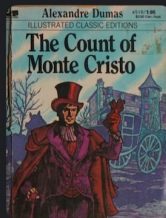
Dracula
Stoker, Bram
● 6.5



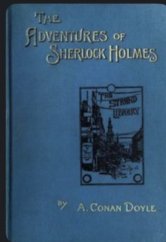
Moby Dick
Melville, Herman
● 7.0



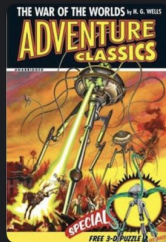
Peter Pan
Barrie, J. M. (James...
● 4.5



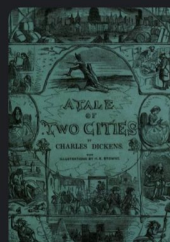
The Count of Monte...
Dumas, Alexandre
● 6.5



The Adventures of S...
Doyle, Arthur Conan
● 6.0



The War of the Worlds
Wells, H. G. (Herbert...
● 6.0



A Tale of Two Cities
Dickens, Charles
● 6.5



Treasure Island
Stevenson, Robert Lo...
● 5.5



Where the Path Breaks
Williamson, C. N. (Ch...
● 6.6

Glite Dictionary v5, 50B words

Top LLM: \$250M, Cheap LLM: \$500K, bi-encoder: \$6500

bank - Glite Dictionary

glitedictionary.com/english/dictionary/bank/ct-ct7ez84F9eJQMZNRe12bank/

Search words or phrases... Spanish

bank

A raised slope of earth beside water or a road

noun /bæŋk/

Spanish:

- The cub tumbled once and disappeared over the stream *bank*, then surfaced downstream, squalling and paddling.
- river bank*
- stream bank*
- on the opposite *bank*

Add for Learning

Key Meanings

- A money place** 1.0
A bank account
noun
- A sloping edge** 3.0
A bank of a river
noun
- Gambling money** JARGON 6.1
The bank at roulette
noun
- A supply kept** 3.4
A bank of blood
noun

Rare & Specialized

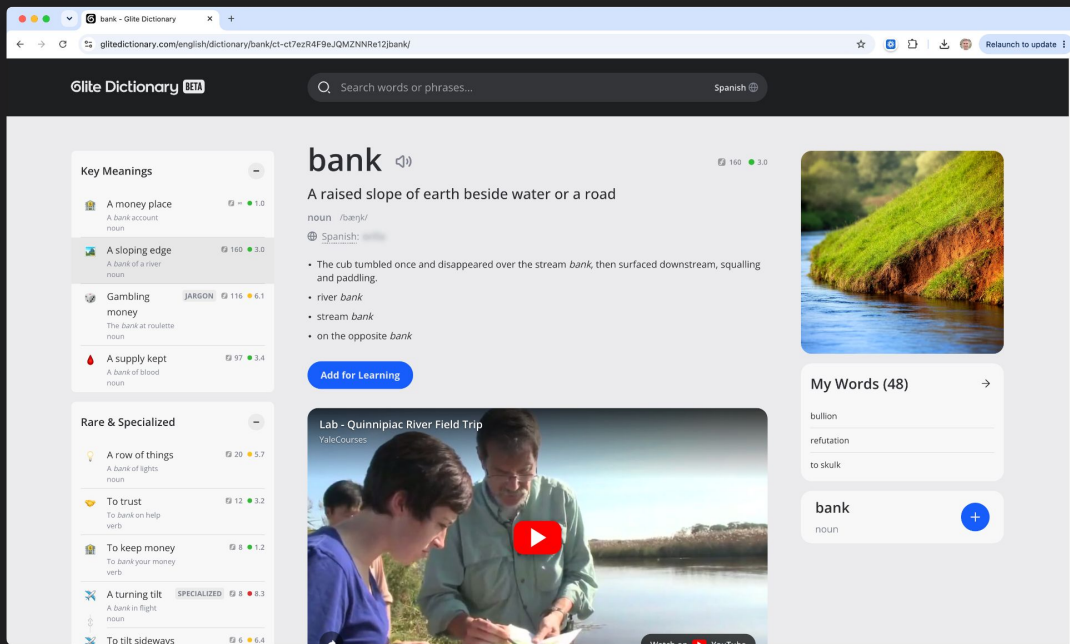
- A row of things** 5.7
A bank of lights
noun
- To trust** 3.2
To bank on help
verb
- To keep money** 1.2
To bank your money
verb
- A turning tilt** SPECIALIZED 8.3
A bank in flight
noun
- To tilt sideways** 6.4

Lab - Quinnipiac River Field Trip
YaleCourses

My Words (48)

bullion
refutation
to skulk

bank
noun



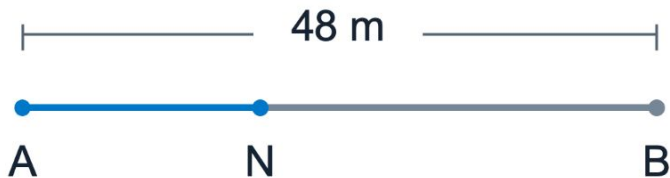
Student Prices

Time of typical task	2-5 mins
Concepts in a typical task	10-50
Concepts per student hour	100-1000
Top LLM concept extraction	\$0.5 - \$5 per student hour

Why did the student choose 16 m?

$$AN : NB = 3 : 5$$

What is the distance AN?



Selected wrong answer

16 m

Which explanation fits best?

A Finds the other part of the ratio

B Divides the whole by the first ratio number

C Stops after finding one ratio unit

1st Place

Misconception Collection (2.5k+)

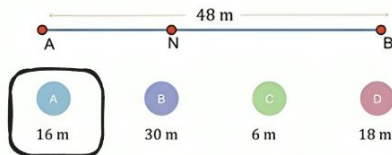
id	MisconceptionName
0	Does not know that angles in a triangle sum to 180 degrees
1	Uses dividing fractions method for multiplying fractions
2	Believes there are 100 degrees in a full turn
3	Thinks a quadratic without a non variable term, can not be factorised
4	Believes addition of terms and powers of terms are equivalent e.g. $a + c = a^c$



Query: Question + Incorrect Answer

I= Eedi

In the diagram, $AN : NB = 3 : 5$
What is the distance AN ?



Incorrect

Retrievers

*To be re-ranked

32 to 64 Misconception Candidates

Stage 1 Ranker (Pointwise)
Qwen/Qwen2.5-14B

Top 8

9 to 25

Stage 2 Ranker (Pointwise)
Qwen/Qwen2.5-32B

Top 5

6 to 8

9 to 25

Stage 3 Ranker (Listwise)
Qwen/Qwen2.5-72B

1 to 5

6 to 8

9 to 25

Public : 0.524
Private : 0.475

Public : 0.611
Private : 0.615

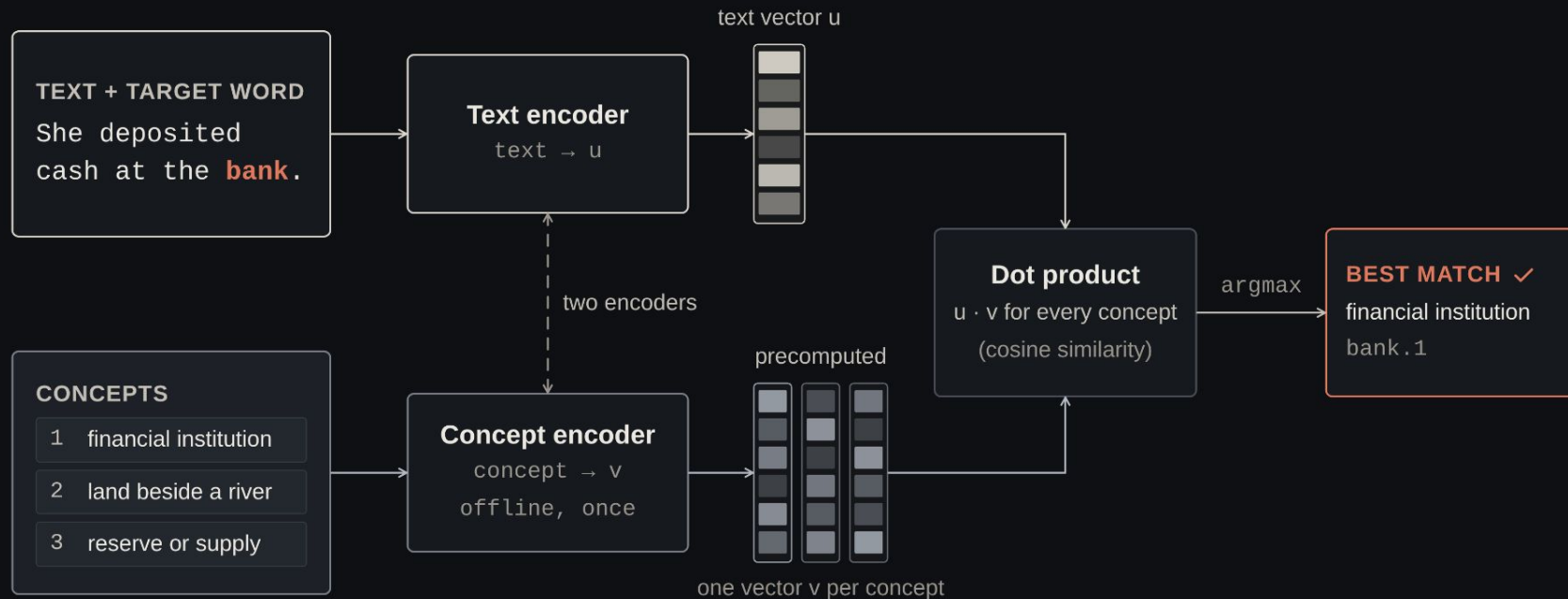
Public : 0.636
Private : 0.625

Public : 0.643
Private : 0.638

Practical Recipe

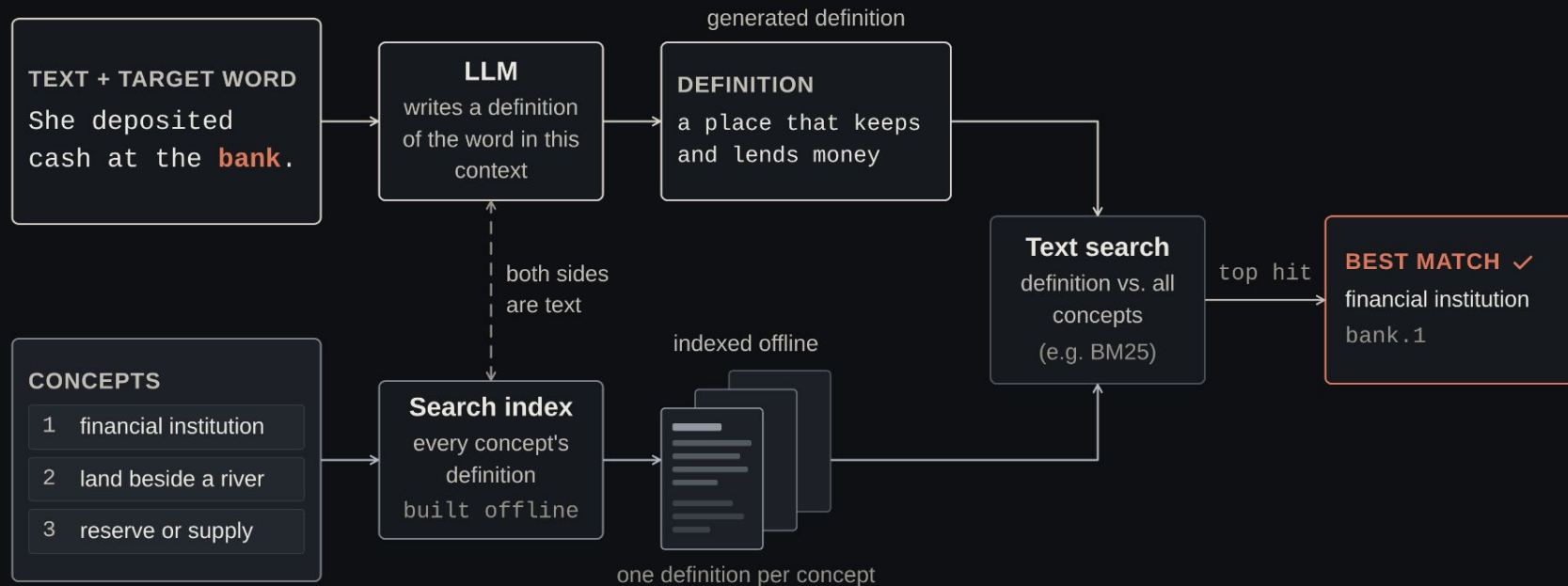
Bi-encoder: embed separately, match by dot product

A text encoder embeds the text as u , a concept encoder embeds each concept as v . The concept with the largest $u \cdot v$ wins.



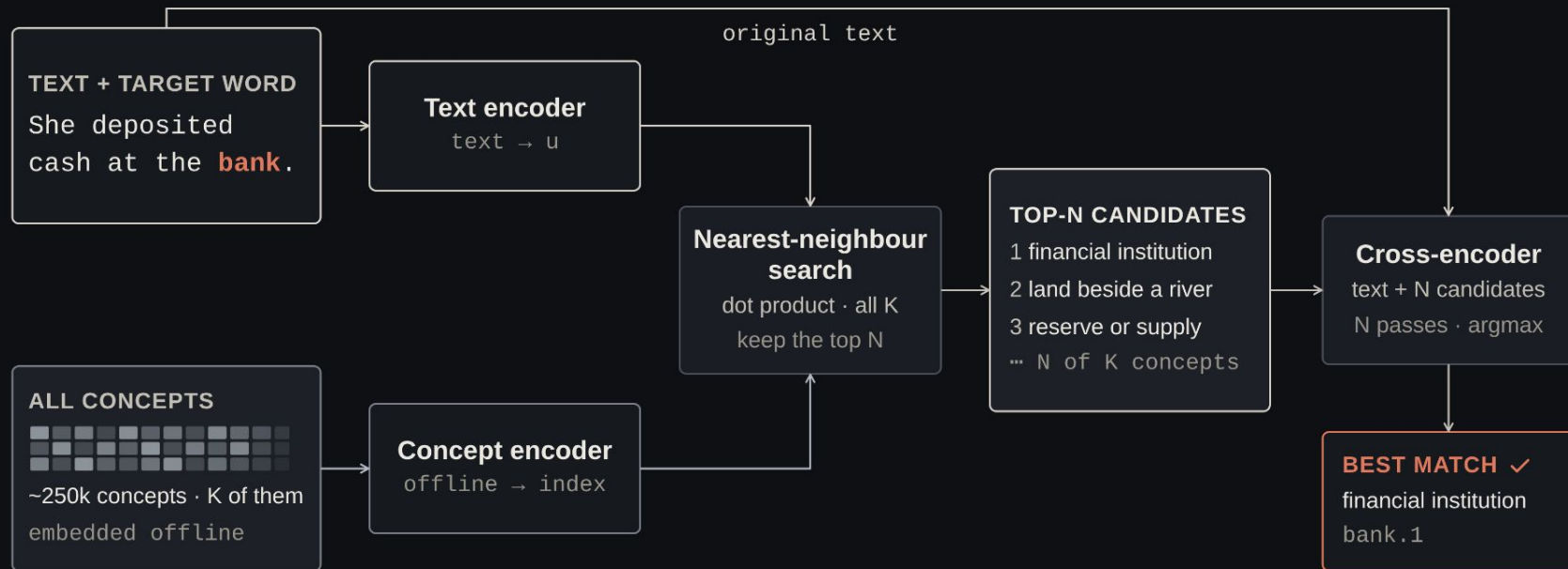
LLM bi-encoder: write a definition, then search for it

The LLM writes a definition of the word in context; text search finds the closest concept definition.



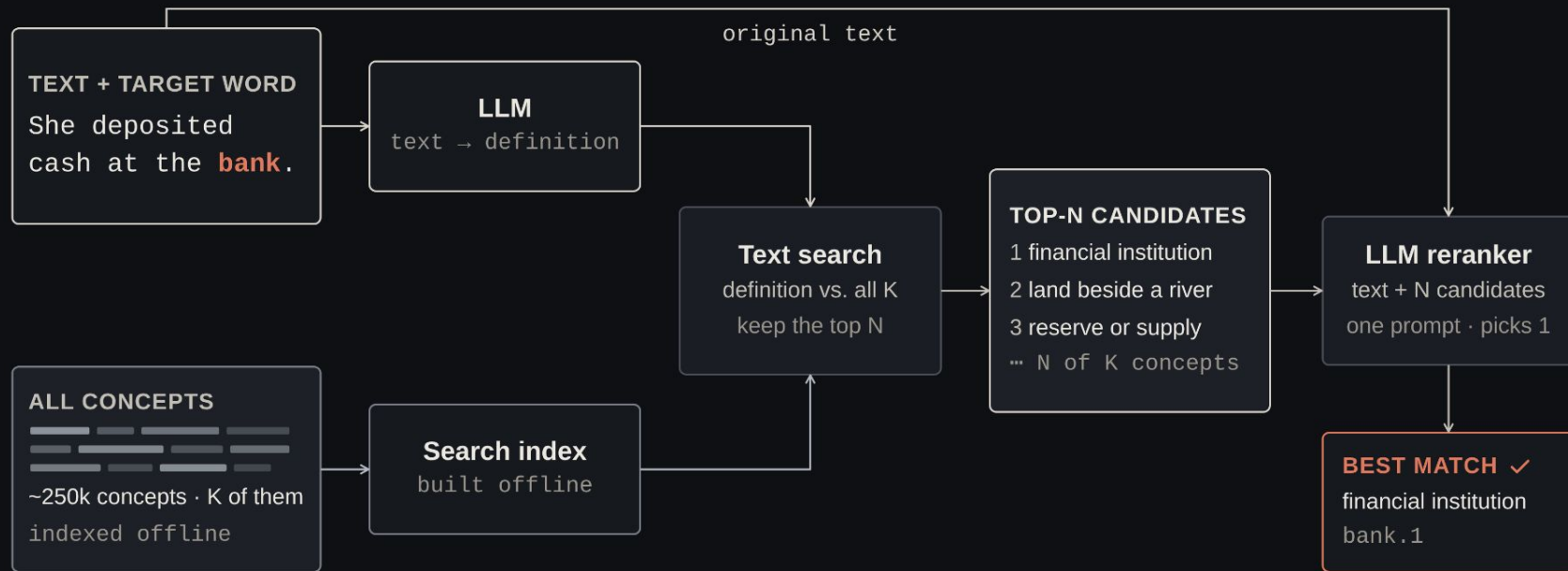
Hybrid: retrieve by dot product, rerank with a cross-encoder

Dot product over the concept index shortlists N candidates; the cross-encoder then scores only those, each together with the text.



LLM hybrid: write a definition, search, then rerank

Text search over the concept definitions shortlists N candidates; the LLM reranker then reads only those, together with the text.



Training

1. Fine-tuning encoders or smaller LLMs
2. Use top LLMs as teachers
3. Active learning: choose which examples are worth paying to label
4. Contrastive training

Discrete concepts: split the space, count the ids

EMBEDDING SPACE · CONTEXTS OF “BANK”



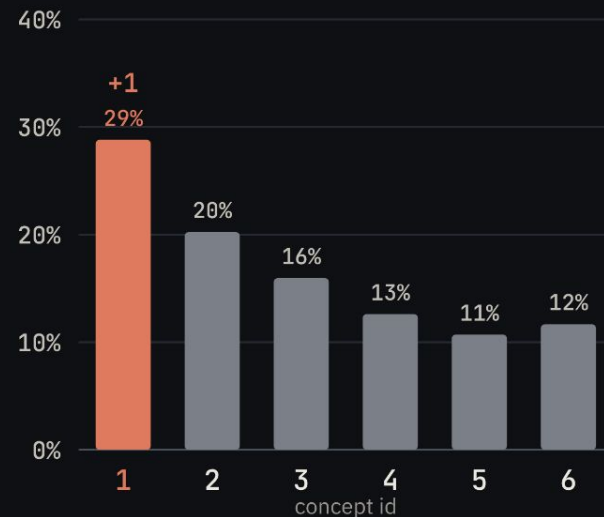
one dot = one context · borders = the knowledge graph

TEXT + TARGET WORD
She deposited
cash at the **bank**.

Concept extractor
context → id

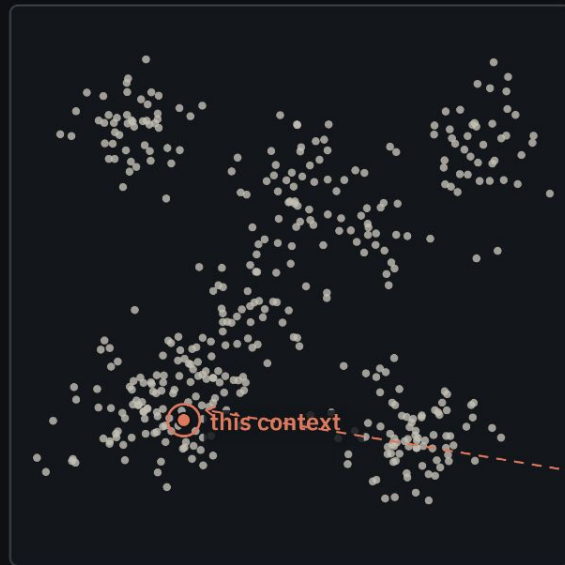
CONCEPT ID
id = 1
one of 6 ids

FREQUENCY · ONE COUNT PER ID



Continuous concepts: keep the space, estimate the density

EMBEDDING SPACE · CONTEXTS OF “BANK”



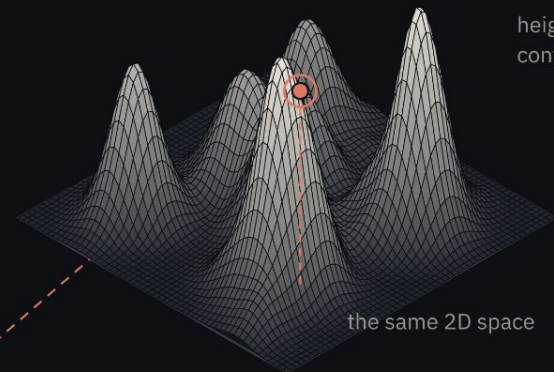
one dot = one context · no borders

TEXT + TARGET WORD
She deposited
cash at the **bank**.

Concept extractor
context → embedding

EMBEDDING
 $e = (0.31, 0.26)$
a point, not an id

DENSITY · A FUNCTION OVER THE SPACE



height = how often
contexts land here

Glite

Thank-you!

We hope to see you at our next meetup next year