<Codelab/>

The AI helps you learn. It won't do your exam.

An online judge where the teacher sets how much AI a session allows — and where allowed AI use is graded, not ignored.

TRY IT LIVE
codelab.ai.vn

WHILE LEARNING
The assistant explains. It refuses to write the solution.

IN THE EXAM
The teacher picks the level. Where AI is allowed, prompts are graded too.

FOR THE TEACHER
The same models write exercises, verify them, build papers, mark them.

INTRODUCTION

Who actually wrote this code?

An AI finishes most coursework in seconds. Ban it or ignore it — either way, the grade stops showing who learned.

RESEARCH QUESTIONS

RQ1 Can "how much AI help" be a configured, auditable setting?

RQ2 Can an LLM pipeline fill an exercise bank without losing quality?

RQ3 What must a judge do beyond stdin and stdout?

METHOD · PLATFORM

One judge for the whole curriculum

Student browser
editor · timer · assistant chat

Platform API — gate
sessions · grading · audit log

MongoDB · Redis
submissions · logs

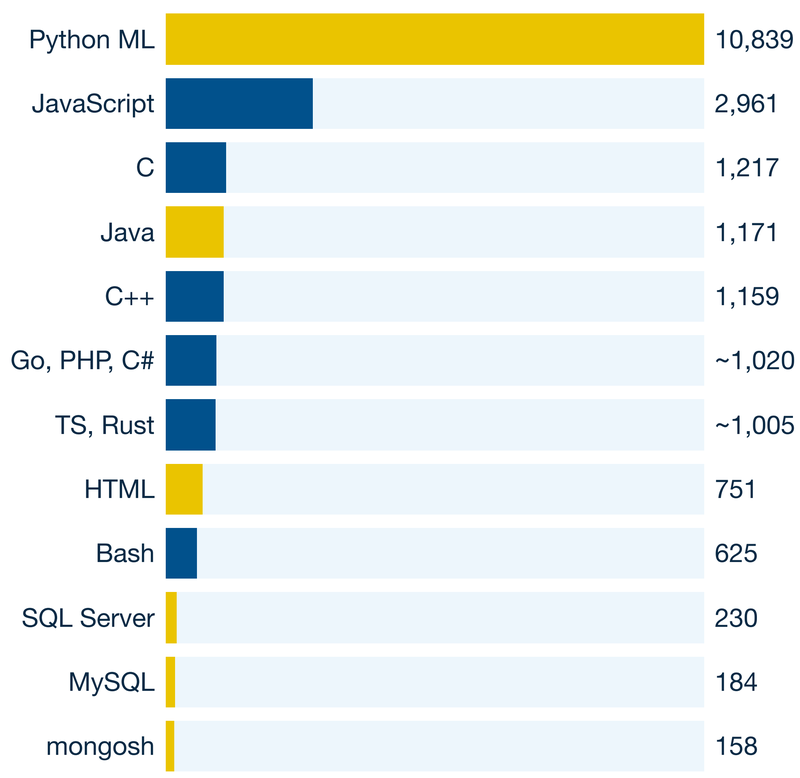
LLM providers
self-hosted · API

Judge0, customized
19 targets · CPU, memory and time caps per run
+6 engines: Java 21 · Python+NumPy · mongosh
SQL Server · MySQL · HTML — sandbox per submission

A database exercise gets a real server, seeded and destroyed per submission.

RESULTS · COVERAGE

Where the bank sits



Yellow = the six engines we built on top of Judge0 · blue = languages Judge0 ships with.

METHOD · CONTRIBUTION 1

Help while learning, never the answer

```
1 def total(xs):
2   s = 0
3   for x in xs: s = x
4   return s
```

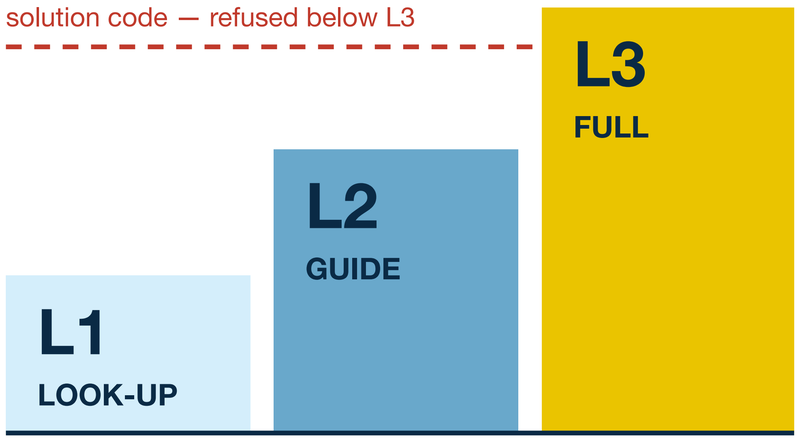
case 1 ✓ | ✓ | ✗ | ✗ | score 2 / 4

STUDENT
Just write the correct function for me.

ASSISTANT · LEVEL 2
I won't write the solution. After each step, what should **s** hold — this item, or everything so far?

METHOD · THE LEVELS

The teacher sets the level



L1 concepts and syntax · L2 strategy in words · L3 code allowed, because using AI well is the skill. The session configuration decides — never the chat window.

METHOD · EXAMS WITH AI

At L3, the prompt is part of the answer

Every turn is stored with the submission. The teacher marks the code the judge scored **and** how the student steered the assistant.

METHOD · CONTRIBUTION 2

The same AI works for the teacher

specify → generate → critique → tests → verify

from a topic slot — **on the students' own judge**

Two models draft and critique; nothing ships until the reference solution passes.

Builds one paper per seat

Marks & re-grades in bulk

RESULTS · IN PRODUCTION

164,370
submissions judged
128,660 in exams

291,132
sandboxed runs
one per test case

24,096
exercises
116,325 test cases

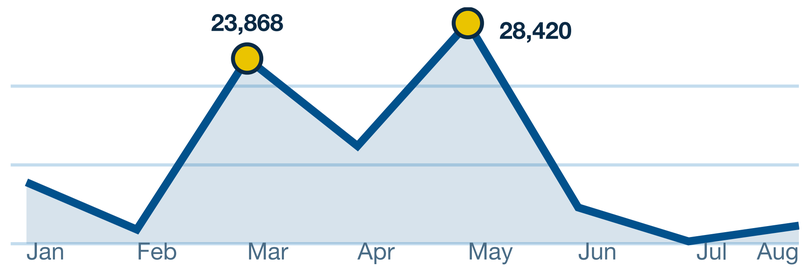
2,192
accounts
1,998 students

181
exam sessions
largest one 310 students

14,185/day
busiest exam day
4,372 in an hour

RESULTS · RHYTHM

Peaks are exam weeks



Exam submissions per month, 2026 — 81,989 so far, against 46,639 in all of 2025.

RESULTS · OUTCOMES

Exams are measurably harder

Yet **80.2 %** of exam student–exercise pairs end accepted, after nine attempts.

RESULTS · ASSISTANT PILOT

163 conversations, 31 students

- 6 turns asked it to solve the task — all refused, all logged.
- 4.5 s median reply on a self-hosted 14B model.
- 144 of the 163 turns fell in the last 30 days.
- Practice mode only — no learning-gain claim yet.

CONCLUSIONS

- **RQ1** — levels + logs make AI help auditable.
- **RQ2** — verification, not generation, is the bottleneck.
- **RQ3** — a curriculum judge needs stateful sandboxes.

Faculty of Information Technology · NEU EdTech Lab
College of Technology, National Economics University, Hanoi
Open to learners outside NEU: **codelab.ai.vn**

Phạm Xuân Lâm · lampx@neu.edu.vn
Campus platform **fit.neu.edu.vn/codelab**
All figures queried from the production database, 24 Aug 2026.