tizia.vn

Generating is easy in the EduVerse. Getting it right isn’t.

A self-evolving learning platform for every age — from preschool to pharmacy school — and the pedagogical-accuracy check its AI still can’t do without.

Try it live → tizia.vn · A universe of subjects · The reconcile() experiment

REAL, BUT GRADED
The dream holds up at “AI writes it” — confidence falls fast past that.

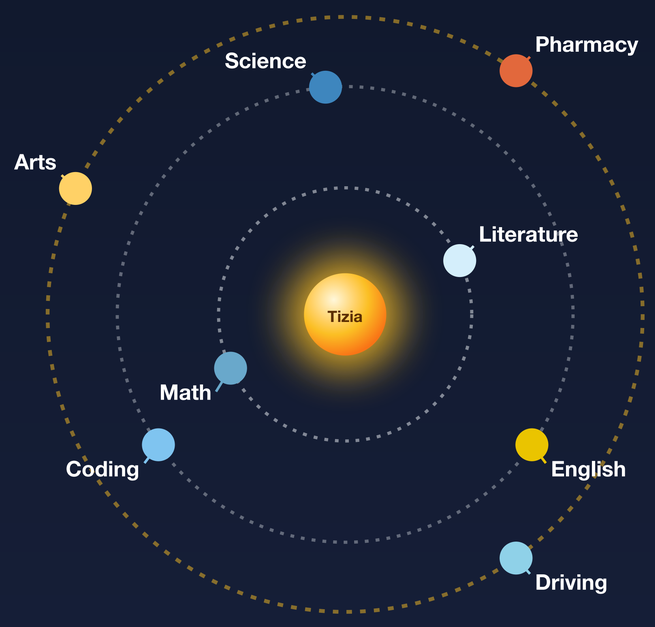
THE BOTTLENECK MOVED
Not “can AI generate it” — “can we verify it’s correct.”

EVERY SUBJECT RAISES THE BAR
From a child’s first fraction to a pharmacy dosage calc — a wrong answer is never just detail.

Try it live
tizia.vn

ONE PLATFORM · A UNIVERSE OF SUBJECTS

A universe where you can learn almost everything



INTRODUCTION

The dream is real. The check isn’t.

Tizia grows the way its own students ask it to: describe a need in plain words, and its AI drafts the code, the lesson, or the practice question. That part already works well. What still can’t be automated is checking whether what it drafted is **correct**.

RESEARCH QUESTIONS

RQ1 Has the bottleneck really shifted from generation to domain verification?

RQ2 How much can one cheap, domain-specific check buy back?

RQ3 Where must a human stay in the loop, and why?

BACKGROUND · WHAT THE FIELD SHOWS

Writing code is solved. Reading intent is not.

- 395 peer-reviewed studies on LLMs for software engineering — the trend is real and growing fast.
- Yet only **~3.9%** of them study requirements, **~0.92%** study design.
- On hard multi-file tasks the strongest agents solve under **25%** — and **35.9%** of failures are wrong reasoning, not bad syntax.

CASE STUDY · MANY WORLDS

5,260 curriculum lessons — and more worlds opening

GENERAL EDUCATION — PRESCHOOL TO GRADE 12

Preschool — Ages 3–5 · play-based	166 lessons	Primary — Grade 1–5 · Math, Vietnamese, Science	1,879 lessons
Lower secondary — Grade 6–9 · full curriculum	1,716 lessons	Upper secondary — Grade 10–12 · exam prep	1,365 lessons

SPECIALIZED SCHOOLS — PROFESSIONAL & LIFE SKILLS

Pharmacy — 5-year track · GPP practice	134 lessons	Driving theory — GPLX question bank	600 questions
IT — Apps & tools · early stage	7 apps	Foreign languages — Apps & tools · early stage	4 apps

5,260 lessons = general education + pharmacy, holding 28,920 questions across 85 subjects. Driving and the newer schools are counted in their own units.

EXPERIMENT · BY GRADE

Grades 8 and 12: still the hardest

- Grade 3 — 70% → 100% after reconcile()
- Grade 6 — 70% → 90% after reconcile()
- Grade 11 — 70% → 90% after reconcile()
- Grade 8 (50%) and grade 12 (40%) are the weakest — reconcile() alone can’t fully close that gap.

CASE STUDY · CONTRIBUTION 1

A five-line check, before every save

server/ai.js — reconcile()

```
1 const reconcile = (choices, answer, exp) => {
2   const hit = matchLastNumber(exp, choices)
3   return hit.length === 1 ? hit[0] : answer
4 }
```

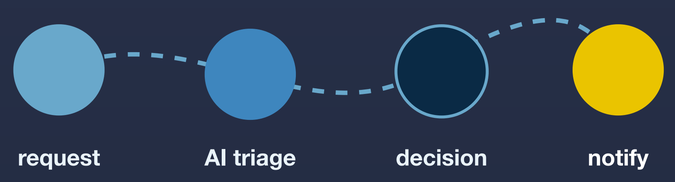
fixed 15 · broken 3 · no-op 1 · net +9.8 pts

MODEL’S OWN EXPLANATION
“3 + 5 = 8, but the right choice in the list is ‘7’.”

RECONCILE() RESULT
Reads the last number it wrote — finds no match, leaves the wrong answer in place.

CASE STUDY · LOOP 1

Student feedback, always reviewed



student writes in the app — **a human makes the call**

139 requests logged so far — 119 of 177 decisions made by a human, not the AI.

65	8	272
apps & tools	domains	commits/10wk

Platform library, 16 Aug 2026 — 57 built in-house, 8 integrated.

FRAMEWORK · THREE OPEN GAPS

Where the gap actually sits

T1 Read intent correctly — turn a vague wish into a checkable requirement.

T2 Edit and connect, don’t just spawn — reshape the tools users already have.

T3 Verify by domain — hallucination is provably unremovable, so the check must sit outside the model.

CASE STUDY · LOOP 2

AI drafts. A domain check gates it.

qwen2.5:14b, weekly topic — **checked before it ships**

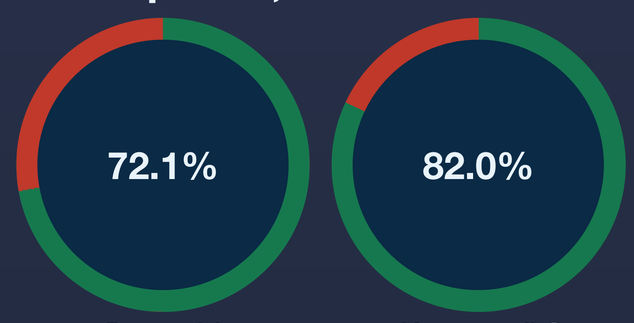
Nothing joins the practice bank without passing the check first.

Domain-specific, not generic · **Runs before every save**

EXPERIMENT · IN NUMBERS

EXPERIMENT · OVERALL ACCURACY

+9.8 points, not zero errors



Raw model — no check applied
After reconcile() — Δ +9.8 pts

122 curriculum questions, answered blind by the same model that drafts new ones.

LIMITATIONS

What this study does not claim

- Small pilot user base — the feedback-loop figures illustrate architecture, not statistics.
- One platform, one language — Tizia, Vietnamese, GDPT 2018.
- reconcile() is narrow — numeric answers only; it catches self-contradiction, not misreading.
- Tied to qwen2.5:14b — a stronger model shifts the size of the effect, not its direction.

CONCLUSIONS

- **RQ1** — the bottleneck really has moved: generation is cheap now, verification decides quality.
- **RQ2** — a cheap domain check bought +9.8 points. A floor, not a ceiling.
- **RQ3** — humans stay in the loop wherever a confident answer can still be wrong.
- **Next** — pedagogical-correctness checks beyond arithmetic.

ĐẠI HỌC KINH TẾ QUỐC DÂN
NATIONAL ECONOMICS UNIVERSITY
College of Technology
Faculty of Information Technology · NEU EdTech Lab

Tizia
SELF-EVOLVING EDTECH

A learning platform for every stage — Vietnam’s K-12 curriculum (GDPT 2018) and beyond · Live at **tizia.vn**
Case study from the production database · platform status as of 24 Aug 2026 · Contact: lampx@neu.edu.vn