

A Combined Strategy for the Pedagogical Evaluation of Automated Feedback



**Badmavasan
Kirouchenassamy^{1,*}**



Amel Yessad¹



**Sébastien
Jolivet^{1,2}**



Sébastien Lallé¹



**Matthieu
Branthôme³**



Vanda Luengo¹

¹ Sorbonne Université, CNRS, LIP6, F-75005 Paris, France ² IUFE & TECFA, Université de Genève, Switzerland ³ University of Rennes, CNRS, IRISA, 22305 Lannion, France

* Corresponding author

28 June 2026

CONTEXT

Pedagogy is multidimensional

- Pedagogy is a **deliberate orchestration** of instructional actions toward a *learner*, a *content* and an *objective*. [Hattie & Timperley, 2007; Narciss et al., 2014]
- A feedback is **not atomic**: it involves *content*, the *timing and form* of delivery, and its appropriateness to the learner. [Shute, 2008; Hattie & Timperley, 2007]



THE PLATFORM

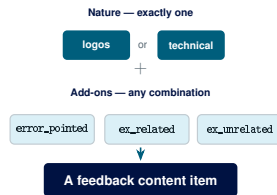
Feedback Content Characterization Framework

For high-school students in **France & Switzerland**

Characteristic	What it is
logos	Conceptual explanation: the notion and why it holds; no code, no procedure.
technical	Procedural hint: which mechanism to use or what to check; no working code.
error_pointed	Names the student's error and redirects to the notion; no correction.
ex_related	Short code example anchored in the current exercise.
ex_unrelated	Code example in a neutral context, independent of the exercise.

A Structured Combination of Characteristics

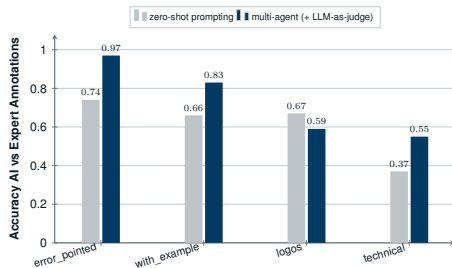
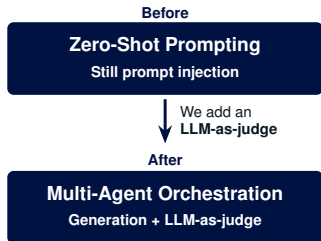
[Jolivet et al., EIAH 2025]



⇒ Content **validity is decidable**: checked *against the framework*, not by asking experts or students “*is it useful?*”

FEEDBACK AS AN ITEM

Zero-shot generation vs. LLM-as-judge



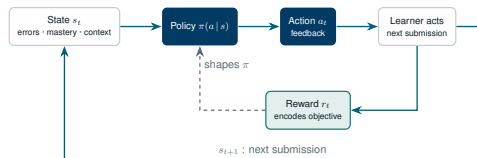
- Feedback is produced **in masses** $\Rightarrow$ per-item expert review is impossible $\Rightarrow$ evaluation must be **automated**.
- Zero-shot**: the AI's own tag is **wrong in $\approx 39\%$** of cases; and even otherwise it **aligns poorly with experts**, on technical, unseen $\sim \frac{2}{3}$ of the time.

FEEDBACK AS A DECISION

Evaluating the Decision

- To evaluate the **pedagogical value of a decision**, one must first **explicitly declare an objective**: the evaluation is then defined *relative to that objective*.
- **No objective** $\Rightarrow$ **no orchestration** $\Rightarrow$ **no pedagogy**.

Objective	Metric	Question
Task Completion	Off-policy return + stat. test (<i>offline</i>)	Are there more completions?
Efficiency	Off-policy return + stat. test (<i>offline</i>)	More completions in fewer attempts?
Error Non-Recurrence	Off-policy return + stat. test (<i>offline</i>)	Do errors reappear?
Mastery Gain	Off-policy return + stat. test (<i>offline</i>)	Did the learner gain mastery?



- **Example: Efficiency** (offline): on the **25,000+ past interaction logs** we estimated *how fast* students would solve the problem under each decision policy (the **Learned** policy vs. an **Expert-defined** and a **Randomized** one) and the efficiency-optimized policy gave the best result (**fewest submissions**), with no deployment on students.

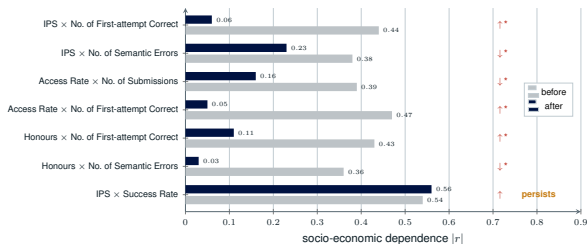
FEEDBACK FOR WHOM

Fair Impact

- Pedagogy must help **every** learner: equity is part of pedagogy, and feedback acts on learners more directly than prediction. [Baker & Hawn, 2022; Kizilcec & Lee, 2022]

Term	Meaning
IPS	Socio-economic index
Access rate	School admission rate
Honours	Honours (mention) rate
first-attempt correct	First submission passes
semantic errors	Runs but wrong result
submissions	Number of attempts
success rate	Eventually solved

- Before vs. after deployment:** does behaviour depend on the school's **socio-economic profile**?
- After deployment most dependence **collapses**, but **IPS × success rate persists** (highlighted): a remaining equity gap.



↑/↓ = as the first factor increases, the second increases / decreases; * = dependence significantly reduced after deployment.

The closer the value is to 0, the more **equitable** the models, only **IPS × Success Rate** stays high.

POSITIONING & CONCLUSION

One feedback, three evaluations, together

- We do **not** propose a new metric, but a **Layered Evaluation Pipeline** that mirrors the multidimensional nature of feedback.
- **Fail any layer** $\Rightarrow$ **Pedagogically Deficient**, even if the others look great.
- The evaluation **feeds back to improve** the feedback decision system, closing the loop.

