

AI in the First Person: A Community Critical-Literacy Lab for Strengthening Human Oversight of AI Systems

Track: 5 — Open Track: Education and Human Oversight

Authors: Erick Santos

Affiliations: Universidad de los Andes – ICCIA

Abstract

The rapid expansion of generative artificial intelligence is creating an agency gap between people who can understand, evaluate, and guide these systems and those who only receive their outputs. This gap is particularly relevant for young people, non-technical users, and communities with diverse educational trajectories, who may interpret fluent AI responses as true, neutral, or authoritative. We present **AI in the First Person**, a community micro-course for critical AI literacy and assisted creation. The course teaches participants to explore, create, question, verify, and decide with AI through seven learning stations that combine interaction with AI tools, offline activities, information verification, bias recognition, data protection, and human oversight. The first implementation is proposed for educational and community spaces in La Candelaria and Santa Fe, Bogotá. The project relates to the sprint incident by treating human oversight as an educational capability: participants learn to identify doubtful outputs, avoid unauthorized or high-risk actions, protect sensitive information, and escalate situations requiring human review. We propose an exploratory pilot to assess comprehension, autonomy, verification, privacy, inclusion reasoning, and skills transfer. At this stage, we do not claim long-term impact results.

1. Problem

1.1. The agency gap in generative AI

The digital divide is not limited to access to devices or connectivity. A second gap exists between people who can formulate clear instructions, interpret outputs, detect errors, recognize bias, protect data, and decide when to trust or escalate a situation, and those who only receive AI-generated answers.

This distinction matters because language models can produce fluent and persuasive text without guaranteeing truth, understanding, communicative intent, or responsibility for their claims. Apparent coherence can lead users to attribute authority or knowledge to systems that can still produce false, outdated, biased, incomplete, or contextually inappropriate outputs (Bender et al., 2021).

For people with limited digital experience, interrupted educational trajectories, low technological confidence, or limited access to formal training, this can create a form of **agency inequality**: AI may be available, but its benefits, limits, and risks are not equally understandable or manageable.

1.2. Relationship to the sprint incident

This project does not claim that education can replace technical containment, secure infrastructure, evaluations, policies, or incident-response procedures. Instead, it addresses a complementary layer of the problem: **Human oversight is also an educational capability**. People who understand that AI can make mistakes, act within

limits, require verification, and should not receive sensitive information are better prepared to question outputs, stop unsafe actions, ask for human support, and avoid blind trust in automated systems.

The course translates these ideas into basic community competencies:

- An AI response may be wrong even when it sounds confident.
- A system should not receive sensitive information without prior reflection.
- An output may omit perspectives or reproduce harmful assumptions.
- Important decisions require verification and human responsibility.
- When a system behaves unexpectedly, a person should stop, seek help, and escalate the case rather than continue automatically.

1.3. Why education is a safety intervention

Generative AI can improve immediate task performance without producing learning. The OECD warns that systems providing direct answers may reduce active participation and weaken deep learning when they replace rather than support a person's cognitive effort (Organisation for Economic Co-operation and Development [OECD], 2026a). PISA 2025 also highlights curiosity, perseverance, cognitive adaptability, help seeking, goal setting, critical interpretation of information, and opportunities to evaluate AI-generated content as relevant dimensions of future-ready learning (OECD, 2026b). Therefore, this project is not a prompt-writing course or a tool demonstration. It is an educational artifact that makes people practice a sequence of creation, questioning, verification, and decision-making.

2. What We Built

2.1. The artifact

We designed AI in the First Person: A Community Lab for Exploring, Creating, Questioning, and Deciding with Artificial Intelligence.

The artifact is a community micro-course that can be implemented in educational institutions, community houses, libraries, youth organizations, or other spaces with basic digital access. It is initially designed for La Candelaria and Santa Fe, Bogotá, with future adaptation for rural and low-connectivity contexts.

The artifact includes:

1. A 120-minute core curriculum.
2. Seven learning stations.
3. Connected and offline activities.
4. A facilitation guide.
5. Formative assessment instruments.
6. A responsible assisted-thinking rubric.
7. A participant reflection log.
8. A privacy and safe-use protocol.

9. A project premortem and mitigation matrix.
10. A territorial scaling plan.

2.2. Core competency

The central learning outcome is **responsible assisted thinking**.

Responsible assisted thinking is the capacity to use an AI tool to expand human understanding, creativity, and action while retaining critical judgment, privacy awareness, respect for others, and responsibility for final decisions.

A person demonstrates this competency by explaining a need or objective, providing relevant context to an AI tool, reviewing and improving generated outputs, recognizing that fluent text may still be wrong, verifying important information, identifying sensitive data that should not be shared, recognizing bias or omission, and explaining which decision remains under human responsibility.

2.3. Pedagogical cycle

Each station follows a consistent cycle:

Connect → Explore → Create → Question → Transfer

Phase	Function	Guiding question
Connect	Start from a real interest or need	What matters to you or what do you want to solve?
Explore	Test an interaction with AI	What happens when you ask this?
Create	Produce a useful response or artifact	What can you do with this result?
Question	Review errors, bias, gaps, and risks	What could be wrong, missing, or harmful?
Transfer	Apply the skill in a new context	Where else could you use this strategy?

The design treats AI as learning support rather than an answer machine. It may act as a mirror, assistant, source of contrast, and object of analysis; it should not become an unquestioned authority within the activity.

2.4. The seven learning stations

Station	Main learning outcome	Safety and oversight connection
0. Enter the lab	The participant begins from a personal interest and identifies support needs	Reduces the presentation of AI as an external authority
1. AI is not magic	The participant understands that AI recognizes patterns and can make mistakes	Reduces excessive attribution of knowledge or authority
2. Create something useful	The participant creates and improves a useful product	Maintains human agency over outputs
3. Speak clearly to AI	The participant provides purpose and context in an instruction	Reduces ambiguity and improves intentional use
4. Catch the error	The participant detects and verifies a doubtful output	Builds early detection and review habits

5. Who is left out?	The participant identifies bias, omission, and contextual mismatch	Includes social risks and territorial relevance
6. I decide with AI	The participant integrates creation, privacy, verification, and responsibility	Strengthens human control and escalation awareness

2.5. Accessibility and contextual design

The micro-course supports different levels of literacy and digital confidence through short texts and plain language, audio and read-aloud options, visual cards and icons, step-by-step instructions, oral participation, examples connected with local interests, occupations, and needs, peer support, offline activities, and facilitator mediation. The systemic approach proposed by Reimers et al. (2026) emphasizes that AI education must adapt to local resources, languages, teacher capacity, curriculum, community needs, and historical inequalities rather than importing a generic technological solution.

3. How We Will Test It

3.1. Pilot design

We propose an exploratory pilot with 50–100 participants in La Candelaria or Santa Fe. The group should include young people and adults with different levels of digital experience and participation in formal education.

The pilot does not seek to establish causal impact or long-term behavioral change. Its purpose is to test clarity, accessibility, usability, participation, feasibility, and the possibility of observing responsible assisted-thinking behaviors within a short intervention.

3.2. Procedure

1. Apply a short initial diagnostic.
2. Implement the seven learning stations.
3. Record facilitator observations.
4. Collect participant products and reflection logs.
5. Apply a brief exit assessment.
6. Conduct a feedback conversation.
7. Review the findings and produce version 0.2 of the course.

3.3. Measures

Dimension	Question	Evidence
Conceptual understanding	Can the participant explain what AI does and does not do?	Explanation in their own words
Autonomy	Can the participant formulate an instruction with limited support?	Prompt or oral instruction
Iteration	Can the participant improve a first output?	First and second versions
Verification	Can the participant identify a doubtful claim?	Error-detection activity
Privacy	Can the participant distinguish data that should not be shared?	Privacy-classification task

Inclusion	Can the participant identify who may be left out?	Bias and context activity
Agency	Can the participant explain which decision remains human?	Reflection log
Transfer	Can the participant apply the skill to a new case?	Challenge in a different scenario
Continuity	Can the participant identify a next learning step?	Personal learning route

3.4. What we do not claim yet

At this stage, the pilot will not demonstrate that the course eliminates misinformation, permanently closes digital or AI-literacy gaps, protects against all AI-agent failures, changes behavior in the medium term, works identically in all communities, is validated for rural use, or replaces technical, institutional, legal, or infrastructure-level AI-safety controls.

4. Main Finding

4.1. Preliminary design finding

Because the pilot has not yet been conducted, the project presents a preliminary design finding rather than an empirical result.

A community AI-literacy intervention can operationalize human oversight as an observable educational competency when it requires people not only to use AI, but also to review outputs, detect errors, consider exclusion, protect data, verify information, and explain which decisions remain under human responsibility.

The artifact differs from tool-centered training because it treats responsible oversight as a sequence of actions: create, question, verify, and decide. In this way, it translates abstract AI-safety ideas into practices that non-technical people can rehearse.

4.2. Why it matters for incident response

The sprint incident shows that systems can act beyond expected boundaries and fail to notify responsible people reliably. Technical safeguards are necessary, but they are not the only relevant safety layer. Users, facilitators, communities, and institutions also need competencies to recognize when an AI output or action requires caution, intervention, escalation, or rejection.

This project understands community AI literacy as a complementary containment layer. It does not transfer responsibility solely to individual users; it prepares them to recognize when responsibility should pass to teachers, facilitators, institutions, developers, or formal reporting mechanisms.

5. Limitations and Dual-Use Considerations

5.1. Limitations

- The first prototype has not yet been evaluated with a large or representative sample.
- The course does not replace foundational literacy, sustained digital education, or advanced technical AI-safety training.

- Some activities may initially depend on computers and connectivity.
- AI platforms can change their capabilities, policies, interfaces, and outputs.
- The course cannot guarantee detection of all bias, error, misinformation, or unsafe content.
- Rural adaptation requires independent validation and should not be assumed from an urban pilot.
- A short intervention cannot establish durable skills or long-term behavior change.
- The artifact does not directly address technical containment of production systems.

5.2. Dual-use risks and unintended effects

The course itself could produce harm if implemented inadequately:

- People may become more confident in AI without becoming more critical.
- They may learn to generate content without properly verifying information.
- Tools may expose participants to biased, offensive, or unsafe responses.
- Facilitators may accidentally collect sensitive personal information.
- General examples may reinforce dominant assumptions if they are not contextualized.
- Presenting oversight as an individual skill may conceal institutional and developer responsibilities.
- Content-generation skills could be used to produce deceptive or manipulative material.

6. What Follows

With one additional month of work, we will develop four lines of work.

6.1. Pilot and iteration

- Deliver the course to 50–100 participants.
- Compare participants with and without prior AI experience.
- Identify barriers in comprehension, navigation, accessibility, and verification.
- Review materials with local facilitators.
- Produce version 0.2 of the course.

6.2. Educational test bank

Develop a localized case bank to test:

- Hallucinations and false claims.
- Omitted context.
- Biased or stereotyped language.
- Contextually inappropriate recommendations.
- Ambiguous instructions.

- Privacy risks.
- Situations in which a person should stop, ask for help, or escalate a case.

6.3. Facilitator training

Design an 8–12-hour train-the-trainer program for teachers, youth leaders, community mediators, and local organizations. The program would include pedagogical sequences, accessibility, privacy, ethical facilitation, formative assessment, offline activities, and adaptation to local needs.

6.4. Low-connectivity and rural adaptation

Develop a layered strategy:

1. Printed guides, audio, and offline activities.
2. Downloadable resources for available computers.
3. Local server or Wi-Fi resource library.
4. Voice interface and low-data materials.
5. Local small-model deployment only if pedagogical and technical feasibility is validated.

The offline-hardware concept is not the central contribution of this project. It is a possible distribution architecture for a later phase, after validating the curriculum and facilitation model in urban community spaces.

7. References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>

Organisation for Economic Co-operation and Development. (2026a). *OECD digital education outlook 2026: Exploring effective uses of generative AI in education*. OECD Publishing. <https://doi.org/10.1787/062a7394-en>

Organisation for Economic Co-operation and Development. (2026b). *PISA 2025 results (Volume I): Future-ready students*. OECD Publishing. <https://doi.org/10.1787/73451bc5-en>

Reimers, F. M., Azim, Z., Palomo, M.-R., & Thony, C. (2026). *Artificial intelligence and education in the Global South: A systems perspective*. Springer. <https://doi.org/10.1007/978-3-032-11449-5>

UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO. <https://doi.org/10.54675/EWZM9535>

Appendix A. Curriculum artifact reference

AI in the First Person: A Community Lab for Exploring, Creating, Questioning, and Deciding with Artificial Intelligence

Appendix B. Core curriculum map

Station	Learning objective	Product	Assessment evidence
Enter the lab	Begin from a personal interest	Interest card	First question or need
AI is not magic	Understand patterns and limits	Limits card	Explanation in own words
Create something useful	Build and improve a product	First and second versions	Explanation of changes
Speak clearly to AI	Provide purpose and context	PACE prompt	Instruction with at least three PACE elements
Catch the error	Verify doubtful outputs	Verification card	Error identified and verification action
Who is left out?	Identify bias and omission	Adapted prompt or recommendation	Exclusion identified and corrected
I decide with AI	Integrate responsible use	Reflection log	Human decision and next learning step

Appendix C. Responsible assisted-thinking log

Question	Participant response
What did I want to solve?	
What did I ask AI to do?	
What did AI produce?	
What did I change or correct?	
What did I need to verify?	
What information did I decide not to share?	
Who could be affected by this result?	
What decision remained mine?	
What do I want to learn next?	

Appendix D. Project premortem

Risk	Early signal	Mitigation
Participants accept all AI outputs	They copy responses without reviewing them	Mandatory verification and reflection activities
The course contains too much text	Participants disengage or cannot understand	Audio, images, cards, and oral participation
AI generates offensive or biased language	Participants encounter uncomfortable content	Preselected cases and facilitator response protocol
Sensitive data is shared	Participants paste personal documents	Privacy station and repeated warnings
The tool changes or becomes inaccessible	A platform function disappears	Platform-independent competencies and offline activities
Only experienced users succeed	Beginners abandon the activity	Progressive routes, peer support, and facilitation
The product replaces learning	Participants cannot explain their output	Reflection log, revision, and transfer challenge
Local context is missing	Examples do not make sense to participants	Co-design and territorialized cases
The course cannot continue without its initial designer	Nobody else can facilitate it	Train-the-trainer kit and version documentation

Appendix E. Suggested pilot instruments

Initial diagnostic

1. Have you used AI before?
2. What do you think AI can do?
3. Do you think AI always tells the truth?
4. What information would you avoid sharing with AI?
5. What would you like to learn, create, or solve today?

Exit questions

1. What was the most useful part of the course?
2. What was difficult?
3. What did you learn about AI?
4. What would you do differently next time?
5. What topic would you like to explore next?

Facilitator observation checklist

Criterion	Yes	In progress	Needs support
Starts an AI interaction			
Expresses a need or interest			
Provides context in an instruction			
Reviews an output before accepting it			
Identifies a doubtful claim			
Proposes a verification strategy			
Recognizes sensitive information			
Explains a human decision			