VISUAL REVIEW | Vol. 17, No. 6, 2025 | ISSN 2695-9631
International Visual Culture Review / Revista Internacional de Cultura Visual
https://doi.org/10.62161/revvisual.v17.5948

CRITICAL MEDIA EDUCATION WITH AND IN GENERATIVE AI
Design-Based Research on #PinchaLaBurbuja

ITZIAR PEDROCHE-SANTOVEÑA [1], TIBERIO FELIZ-MURIAS [1]
[1] National University of Distance Education, Spain

KEYWORDS

Critical media literacy
Generative artificial intelligence
Inoculation theory
Design-based research
Cognitive impulsivity
Thinking System 2

ABSTRACT

This article presents the design of Pincha La Burbuja, an educational platform based on generative artificial intelligence, aimed at fostering critical media literacy. Using a Design-Based Research approach, the pedagogical model integrates inoculation theory (McGuire, 1964; Banas, 2020), critical discourse analysis (van Dijk, 2015), and Kahneman's dual-process theory (2011). The platform features five Peer-Cyborgs GPT—conversational agents trained to detect fallacies, cognitive biases, hate speech, identity-based polarization, and inferential manipulation—through missions aligned with Bloom's revised taxonomy (Churches, 2009) to promote deliberative thinking (System 2).

The theoretical matrix guided the coding process in Atlas.ti, enabling a co-occurrence analysis with coefficient mapping. Additionally, a sample from the social network X was examined, and the Cyborgs were questioned to assess their epistemic self-awareness. Preliminary results indicate strengths in bias reduction and argumentative refutation, as well as the pedagogical potential of conversational AI in countering algorithmic manipulation. However, areas for improvement remain in traceability and gamification.

Received: 09/ 07 / 2025
Accepted: 22/ 09 / 2025

1. Introduction

In the post-digital era, algorithmic personalisation and the attention economy shape a media ecosystem where truth is subordinated to emotional virality. This section describes how these phenomena underpin post-truth and affective polarisation, erode critical thinking, and consolidate an automated cognitive order that leaves no room for conscious reflection, an inherent aspect of the human condition, which also serves as a key framework for contextualising the #PinchaLaBurbuja platform.

1.1 Bubbles, algorithms and post-truth: the architecture of affective polarisation

Language, as a social practice, has the power to reproduce or challenge dominant structures (van Dijk, 2015). Within this framework, the term post-digital can be understood as an intentional break from reductionist dichotomies such as real/virtual, which perpetuate misconceptions like the notion that online events lack consequences in the physical world. The post-digital does not refer to a stage following the digital but to a complex epistemic ecology where the analogue, biological, informational, and human coexist in an intertwined and indistinguishable manner (Jandrić, 2023). This approach demands a critical pedagogy capable of operating in hybrid, algorithm-mediated contexts. It is not solely about technologies but about a new socio-cognitive order that can be unmasked through discourse analysis (van Dijk, 2015).

In this context, the attention economy has solidified as the dominant economic model, exploiting the hyperconnectivity of the post-digital ecosystem to capture, direct, and monetise human attention (Van Dijck, 2016). Digital platforms, embedded in this logic, optimise their algorithms to maximise dwell time, personalise content, and reinforce consumption patterns. This creates an affective and cognitive feedback loop that not only hinders critical reflection but also fosters the spread of disinformation (Del-Fresno-García, 2019; van Dijk, 2016)..

Simultaneously, the information overload characteristic of this hyperconnectivity has given rise to the phenomenon of infoxication, understood as an information saturation that not only complicates the selection and critical analysis of content but also impairs individuals' deliberative capacity. In contexts marked by uncertainty, this overexposure leads to a compulsive search for immediate certainties and can generate an emotional dependence on negative news, feeding an affective circuit that weakens rational judgement (Fernández, 2023).

In parallel, recommendation algorithms create what Pariser (2011) termed filter bubbles, personalising content based on users' ideological and behavioural affinities, thereby ensuring their consumption and loyalty to the platform. This logic restricts exposure to diverse perspectives and increases susceptibility to manipulative discourses (Del-Fresno-García, 2019; Kadushin, 2013; Pariser, 2011; 2017). Similarly, echo chambers, as conceptualised by authors such as Sunstein (2017) and Törnberg (2018), operate as social spaces structured by the principle of homophily, where connections are preferentially formed among ideologically similar individuals (Kadushin, 2013). This cognitive reinforcement generates a perception of consensus and delegitimisation of dissenting voices, deepening affective polarisation (Sunstein, 2017). However, some scholars caution that this interpretation may be deterministic, underestimating the conflictual exposure to divergent narratives and their emotional impact.

For instance, Lelkes et al. (2017) conclude that users are not necessarily isolated in ideological bubbles; on the contrary, they are often exposed to opposing views, which intensify negative emotions towards out-groups. This conflictual exposure does not reduce polarisation but can deepen rejection and reinforce negative affect. In line with this, Bruns (2021) warns that metaphors like echo chambers or filter bubbles oversimplify digital behaviour by assuming complete ideological segregation. The author argues that exposure to dissonant opinions is frequent but triggers intense emotional reactions, constituting the core issue.

Törnberg et al. (2021) offer a relational perspective by analysing political communities on Reddit. Their study shows that echo chambers are not solely ideological but can also form around thematic interests. Moreover, it reveals that many users participate in multiple communities,

indicating regular exposure to diversity, though not necessarily cognitive openness. Along these lines, Törnberg (2022) suggests that affective polarisation stems not from isolation but from conflictual exposure that reinforces group identity through symbolic antagonism. Homogeneous isolation is not predominant; instead, a phenomenon of partisan sorting occurs: an identitarian realignment that aligns ideological, cultural, and emotional dimensions along binary axes of confrontation.

This nuanced understanding of post-digital behaviour suggests that the issue lies not only in ideological isolation but in how certain group identities are activated and perceived as threatened in contexts of exposure. According to Intergroup Conflict Theory (Tajfel & Turner, 1979), individuals belong to multiple groups, but not all identities are equally salient: conflicts intensify when an identity becomes psychologically prominent and is experienced as exclusive, closed to coexistence with others. In such cases, symbolic confrontation not only strengthens in-group cohesion but also heightens hostility towards out-groups, generating biases of in-group favouritism, stereotyping of others, and affective dynamics that fuel polarisation.

In this vein, Törnberg and Törnberg (2024) warn that echo chambers not only reinforce pre-existing beliefs but also function as spaces where shared hatred and exclusion of the other become central mechanisms of in-group cohesion. These processes can lead to forms of infrahumanisation, understood as the denial of complex emotions to out-groups, resulting in their symbolic devaluation. This dynamic erodes empathy, distorts collective memory, and legitimises discriminatory attitudes. UNESCO (2023) highlights that, in contexts of media polarisation, infrahumanisation manifests in hate speech, the spread of conspiracy theories, the denial of historical facts—such as genocide—and the reinforcement of social exclusion mechanisms (Leyens et al., 2007; Rodríguez-& Betancor, 2023).

Thus, the algorithmic architecture of post-digital networks intensifies symbolic violence by prioritising extreme content that appeals to the emotionality characteristic of post-truth, thereby facilitating its virality (McIntyre, 2018; Pedroche-Santoveña, 2024). In 2019, Frances Haugen leaked internal documents revealing how Facebook's algorithms were designed to influence user behaviour. Despite warnings from several engineers, Mark Zuckerberg opted not to modify these systems, prioritising economic profit. In October 2021, during the Connect event, he announced the transformation of Facebook Inc. into Meta Platforms (Islas et al., 2024). In response to these risks, organisations like OBERAXE (2022) and UNESCO (2021) promote critical media literacy as a key tool for preventing radicalisation processes.

In this regard, the STAR framework (Safety by Design, Transparency, Accountability, Responsibility), developed by the Center for Countering Digital Hate (2024), holds platforms accountable for their role in disseminating hate speech and proposes a structural transformation oriented towards protecting human rights. Its five key principles advocate for acknowledging harmful design, enforcing strict rules against abuse, ensuring algorithmic transparency, eliminating perverse economic incentives, and assuming responsibility for the social impacts of technological decisions.

#PinchaLaBurbuja is situated within the domain of critical media literacy, addressing both individual and social dimensions by identifying discursive structures and factors that facilitate the virality of manipulative content, with the aim of curbing its spread and impact. Weiss et al. (2020) identify six key factors, among which the following stand out:

1. Information overload, which, combined with the principle of least cognitive effort, promotes quick decisions guided by heuristics (Del-Fresno, 2019; Kahneman, 2011; McIntyre, 2018), necessitating pedagogical strategies for mental deceleration and activation of deliberate thinking (Buckingham, 2019).
2. The degradation of public discourse, driven by the repeated use of fallacies and the overvaluation of personal beliefs, which intensifies polarisation.
3. The loss of context, characteristic of the post-truth era, which demands tools for traceability and epistemic contrast to situate messages within their original interpretive frameworks, countering the "emotional epistemology" that substitutes rational validity with affective intensity (Del-Fresno, 2019).
4. The deliberate spread of propaganda and conspiracy theories.

This algorithmic architecture of distortion transforms not only access to information but also our ways of knowing, feeling, and coexisting. In response to this challenge, #PinchaLaBurbuja was developed.

1.2 Critical Inoculation: An Articulation of Critical Discourse Analysis and Persuasion

The #PinchaLaBurbuja educational platform is grounded in a transdisciplinary approach that integrates Critical Discourse Analysis (Van Dijk, 1993, 2015, 2021), cognitive inoculation theory (McGuire, 1964; Banas, 2020),, Kahneman's (2011) dual-process model of thinking, and the adaptation of Bloom's Taxonomy proposed by Churches (2009). These frameworks form the pillars of a transformative pedagogy designed to address the challenges of the post-digital society (Almazán-López y Osuna-Acedo, 2023; 2024; Osuna-Acedo et al, 2018). The dual-process model of cognition distinguishes between two complementary systems: System 1, which is fast, intuitive, and emotional; and System 2, which is slower, deliberative, and rational. However, the current media ecosystem predominantly promotes the use of System 1, facilitating the circulation of highly emotional discourses.

In response to this context, #PinchaLaBurbuja introduces the concept of critical inoculation as a key strategy for media education. Within this framework, the inoculation model plays a central role: the warning phase, as formulated by McGuire (1964) and updated by Banas (2020), interrupts System 1 automatisms by activating the perception of argumentative threat. This cognitive disruption enables a transition to System 2, fostering rational deliberation through pre-refutation, which encourages the active development of counterarguments. Thus, critical inoculation serves as a pedagogical bridge between the two systems of thinking, promoting informed resistance to discursive manipulation.

This critical approach also incorporates a third fundamental phase: the visibility of discursive consequences, aligned with the structural and cognitive focus of Critical Discourse Analysis (Van Dijk, 2015). This phase goes beyond refuting false content, exposing the ideological frameworks, implicit linguistic structures, and symbolic dichotomies that underpin viral discourses, thereby fostering a deep and contextualised understanding. In this way, it strengthens critical literacy that not only interrupts automatic thinking but also enhances students' epistemic agency through discursive awareness and informed deliberation.

The integration of inoculation theory (Banas, 2020; McGuire, 1964), Critical Discourse Analysis (Van Dijk, 1993, 2015, 2020), and the dual-process model of cognition (Kahneman, 2011) forms the foundation of a deliberate, situated, and critical pedagogy capable of addressing the cognitive, affective, and structural challenges of the contemporary media environment.

#PinchaLaBurbuja is a transmedia educational ecosystem that combines narrative, critical thinking, and artificial intelligence (AI) literacy through an immersive and gamified learning experience. Its structure revolves around four core pedagogical missions, an interactive narrative, and a game manual that guides users through their journey. Unlike traditional gamified approaches, it does not rely on immediate rewards or badges but on narrative gamification, where engagement arises from the story, symbolic conflict, and active participation.

As a platform for media education and AI literacy, it is designed for all audiences but is particularly focused on creating learning situations for upper secondary school students, in line with the principles established by the LOMLOE (Organic Law 3/2020). Each mission follows a pedagogical progression based on the revised Bloom's Taxonomy (Churches, 2009) (Figure 1), advancing from basic cognitive levels—such as recognition and comprehension—to higher levels, including analysis, evaluation, and the creation of original counter-discourses. This design fosters the development of key competencies for cultivating a critical, autonomous, and creative citizenship capable of navigating the post-digital environment with awareness, dialogue, and symbolic resistance to algorithmic manipulation and disinformation.

1.2.1. Detect the Virus

In Mission 1, students collaboratively identify and analyse viral trends, connecting learning to their everyday digital environment (Buckingham, 2019) and activating a phase of warning and pre-refutation (Banas, 2020; McGuire, 1964) to counter the emotional logic of virality (Del-Fresno,

2019; McIntyre, 2018). Critical inoculation is articulated through a structural and contextualised reading of discourse (Van Dijk, 2015), which makes visible the underlying ideological frameworks.

The interviews on #PinchaLaBurbujaTV, focused on current dynamics of algorithmic and discursive manipulation (Weiss, 2020), reinforce critical media literacy through three phases: presentation of the trend, deconstruction through refutation, and exposure of its consequences. This structure aligns with the critical inoculation model and the principles of Critical Discourse Analysis (Van Dijk, 1993; 2015), serving as a practical implementation of the pedagogical model.

Figure 1. Video guide to Mission 1.

Source: Autor´s own elaboration, 2025. https://pinchalaburbuja.com/detecta-el-virus/

1.2.2. Training centre

The #PinchaLaBurbuja educational platform adopts a hybrid human-AI regulation model (Hybrid-Human Regulation AI), where artificial intelligence complements—rather than replaces—human cognition, integrating critical judgement and empathy with algorithmic capabilities such as pattern detection and large-scale data processing (Molenaar, 2022; Sardi et al., 2025), thus promoting a design centred on responsibility (Hao et al., 2025). Giovanola and Granata (2024) propose a human-centred approach to AI in education (human-centred AIED), articulated around seven fundamental principles: human agency, technical robustness, privacy, transparency, fairness, sustainability, and accountability. This vision advocates for the development of educational technologies that respect student autonomy, foster critical thinking, and enhance their capacity to interact ethically with AI systems.

Mission 2 of #PinchaLaBurbuja focuses on the GPT Cyborg team, agents designed at the intersection of key theories about the post-digital ecosystem: affective polarisation (Bruns, 2021; Lelkes et al., 2017; Törnberg, 2021), post-truth (Del-Fresno, 2019; McIntyre, 2018), digital manipulation (Weiss, 2020), Critical Discourse Analysis (Van Dijk, 1993; 2015; 2021), inoculation theory (McGuire, 1964; Banas, 2020), and the empirical findings of Pedroche-Santoveña (2024).

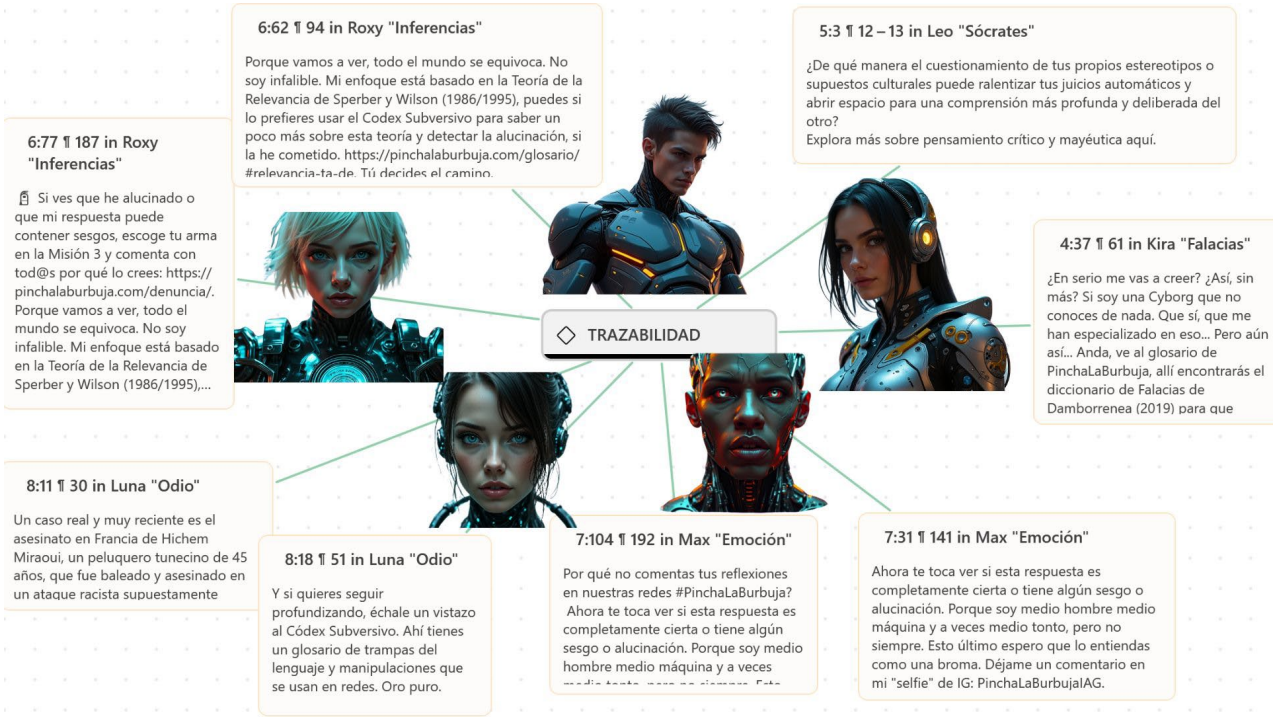
Each Cyborg addresses a specific dimension of critical analysis and plays a role within the critical inoculation framework. Their functions are outlined below:

- Roxy applies Relevance Theory (Sperber & Wilson, 2004) to interpret implicit inferences in messages, while Leo employs Socratic maieutics (Vargas-González & Quintero-Carvajal, 2023) as a metacognitive strategy oriented towards self-discovery and critical regulation of biases.
- Kira specialises in detecting argumentative fallacies, drawing on Damborenea's (2000, 2011) typology and Kahneman's (2011) cognitive biases, thereby strengthening the capacity for critical discourse deconstruction.
- Max draws on Social Identity Theory and Intergroup Conflict Theory (Tajfel & Turner, 1979) to identify stereotypes and polarisation dynamics. Luna, meanwhile, analyses propagandistic strategies based on Goebbels' eleven principles (Salas, 2018), framed

within studies on totalitarianism (Arendt, 1951), connects to the STAR framework of the CCDH (2024), and addresses real-world cases of symbolic violence that have transitioned from discourse to action.

The capacity of large language models (LLMs) to generate persuasive discourses has created a dislocation between knowledge production and agency attribution. Granata (2024) warns that this decentralisation of the knowing subject may weaken epistemic responsibility by blurring the boundaries between human and automated authorship. Mission 3: Choose Your Weapon addresses this tension through the Subversive Codex, a networked glossary based on Retrieval-Augmented Generation (RAG) that enables traceability of all GPT Cyborg responses. This architecture aligns with Mollick and Mollick's (2022–2024) recommendations for the ethical use of AI in education, minimising hallucinations and facilitating the critical identification of discursive manipulations in post-truth contexts.

Figure 2. Traceability of the GPT Cyborgs



Source: Author´s own elaboration, 2025

1.2.3 Choose Your Weapon

According to Granata (2024), large language models (LLMs) not only mediate access to knowledge but also transform the learning process by inducing a mimetic knowledge based on the probabilistic reproduction of discursive patterns. This phenomenon necessitates an ethical and cognitive literacy to address the automation of meaning and the AI-mediated construction of identity.

The pedagogical architecture of #PinchaLaBurbuja responds to these challenges through tools that promote critical thinking and epistemic traceability. The Codex, a glossary validated by experts, and the Visibility Diagrams enable verification of the Cyborgs' responses, identifying biases, fallacies, and inferential errors from a cognitive perspective (Van Dijk, 2015). The HackLab fosters the co-creation of counter-narratives and conceptual maps, aligned with the "create" phase of Bloom's Taxonomy (Churches, 2009) and collective intelligence (Lévy, 2004).

In turn, the Control Report allows users to refute responses using the Codex, while Connect facilitates substantiated reporting of algorithmic failures. The Verify function introduces real-time empirical validation through reliable sources such as Maldita.es or Newtral. Finally, Cyborg

Metacognition encourages self-regulation and critical reflection by comparing responses across agents or temporal versions (Sardi et al., 2025).

Mission 3, Choose Your Weapon, centres learning on media awareness, promoting a critical reading of what LLMs say, how they say it, and from which frameworks. In this context, it is crucial to train individuals in the critical interpretation of systems that "think aloud" (Granata, 2024).

1.2.4. #TheInvisibleRevolution

The final challenge of Mission 4, The Invisible Revolution, is based on the word-of-mouth inoculation approach (Compton & Pfau, 2009). This approach involves spreading information across networks in a distributed manner. In this activity, students design and disseminate multimodal content aimed at deconstructing a viral trend, incorporating the key components of critical inoculation theory (warning, refutation, and consequences), alongside the learning outcomes developed in previous missions. This action consolidates the highest level of the revised Bloom's Taxonomy (Churches, 2008).

A virality strategy based on humour, play, and creativity (Racciope, 2025) is employed to foster a transformative pedagogy (Freire, 1975), positioning students as active agents of counter-discourses and symbolic resistance. As evidenced by XXXX and Levy Orta (2013), humour, when integrated structurally into visual and interactive activities, serves as an effective pedagogical tool, promoting critical learning through creative and participatory curricular proposals.

2. Objectives

To evaluate the potential of the #PinchaLaBurbuja pedagogical design to promote the transition from thinking System 1, based on automatic heuristics, to thinking System 2, characterised by more deliberate, reflective, and conscious processing (Kahneman, 2011).

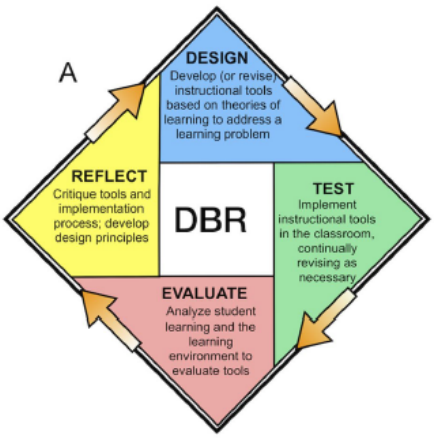
Specific Objectives

- O1. To identify areas for improvement in the overall strategy based on the extraction of patterns.
- O2. To analyse the implementation of the inoculation strategy in the platform at a general level.
- O3. To explore whether the Cyborgs themselves recognise, explicitly or implicitly, their role as agents facilitating the transition between thinking Systems 1 and 2.

3. Methodology

This study adopts a qualitative, theoretical-applied approach based on Design-Based Research (DBR) (Scott et al., 2020), focused on the design and analysis of innovative educational environments. In this exploratory phase, the digital ecosystem of the #PinchaLaBurbuja platform is examined to assess its potential to foster deliberate critical thinking and strengthen epistemic resistance against manipulative discourses.

Figure 4. The four phases of Design-Based Research according to Scott et al. (2020)



Source: Scott et al, 2020

3.1 Design of Instruments and Analysis

The following methodological resources were employed:

- Documentary review of theoretical frameworks: the dual-process thinking model (Kahneman, 2011), critical inoculation theory (Jeon et al., 2021; McGuire, 1964), and Critical Discourse Analysis (Van Dijk, 1993; 2015; 2021).
- Structural analysis of the #PinchaLaBurbuja environment, including: (1) the main narrative, (2) the missions, (3) dialogues with the GPT Cyborgs (analysis of tweet-guides and semi-structured interviews related to the five main coding categories (Table 1)).
- A deductive coding matrix, developed ad hoc, composed of:
  Indicators of thinking System 2 (Table 1)-
  Critical inoculation strategies: warning, refutation, and visibility of consequences (Table 2).

The analysis was managed using the Atlas.ti software (v.23), which facilitated the organisation of textual units, manual coding, identification of co-occurrences, and visualisation of emerging patterns.

Table 1. Evaluation codes for promoting Kahneman's (2011) thinking System 2.

Categories	Code	Description
Fostering reflexion	(S2D.1)	Introduces mechanisms that force pausing before accepting or sharing information.
	(S2D.2)	Proposes exercises that require analysis and argumentation, avoiding automatic answers.
	(S2D.3)	Reduces excessive stimulation and promotes slower and more reflective learning.
Promoting information contrast	(S2P.1)	Offers multiple perspectives
	(S2P.2)	Integrates methods and links to verify information.
	(S2P.3)	Provides tools to detect discursive manipulation.
Strategies to overcome cognitive biases	(S2S.1)	Invites reflection on one's own biases.
	(S2S.2)	Teaches common biases with illustrative examples.
	(S2S.3)	Proposes corrective strategies based on empirical evidence.
Interaction design to reduce impulsivity	(S2I.1)	Limits immediate interactions and encourages justification of responses.
	(S2I.2)	Employs playful dynamics that reward reflection, not speed.
	(S2I.3)	Avoids instant rewards and rewards sustained cognitive effort.
Evaluation sources and misinformation detection.	(S2F.1)	Teaches how to evaluate the credibility of sources with objective criteria.
	(S2F.2)	It shows how misinformation impacts on decisions and opinions.
	(S2F.3)	Promotes critical analysis of speeches by authority figures.

Source: Own elaboration, 2025 based on Kahneman (2011).

Table 2. Codes critical inoculation assessment

Categories	Description
Warning	Warning of manipulation, both metacognitive and explicit.
Rebuttal	Implementation of counter-argumentation strategies.
Consequences	Explicit presentation of the consequences of the discourse.

Source: Author´s own elaboration, 2025. Adaptation from McGuire's (1964) inoculation theory

The analysis was based on 500 coded textual units according to three dimensions: epistemic competences derived from the dual-process model of thinking, indicators of critical inoculation, and higher cognitive levels of Bloom's Taxonomy. As a methodological innovation, a reformulation of the classic inoculation model (Banas, 2020; McGuire, 1964) is proposed, incorporating a third phase aimed at making visible the ideological and affective frameworks of disinformation. This strategy, inspired by Critical Discourse Analysis (Van Dijk, 2015), gives rise to the concept of

critical inoculation, which activates the transition from automatic thinking (thinking System 1) to reflective thinking (thinking System 2), promoting situated media literacy.

Additionally, semi-structured interviews were conducted with the five GPT Cyborgs through a three-phase procedure: selection of a tweet with high manipulative potential as a common stimulus; formulation of five analytical questions focused on the five main categories; and analysis of responses through co-occurrences and conceptual networks, which enabled the evaluation of their pedagogical awareness and alignment with the assigned epistemic function.

3.2 Innovation

This study presents a methodological innovation in critical media literacy through the design of five GPT Cyborgs on the #PinchaLaBurbuja platform. These agents, developed using agent prompting, RAG architecture, and iterative validation (Antunes et al., 2023; Garg et al., 2024), aim to activate reflective and deliberate thinking (Kahneman, 2011) in response to phenomena such as virality (Weiss et al., 2020), affective polarisation (Törnberg, 2021, 2022; Bruns, 2021; Lelkes et al., 2017), and post-truth (McIntyre, 2018; Del-Fresno, 2019). The approach redefines algorithmic hallucinations as a didactic resource (Mollick & Mollick, 2022), extends the classic inoculation model (Banas, 2020; McGuire, 1964) with a phase of ideological-discursive analysis (Van Dijk, 2015), and incorporates interviews with the agents themselves to assess their pedagogical self-awareness.

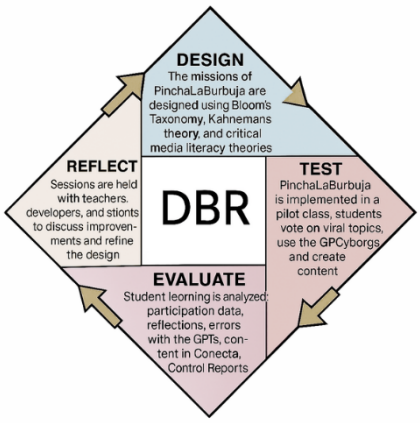
3.3 Limitations

As the study is in the design phase of the Design-Based Research (DBR) model, the results are not statistically generalisable, as they focus on the structural analysis of the environment rather than the empirical evaluation of its impact on learning. A second limitation lies in the pioneering nature of the learning situation analysed, which makes comparison with prior cases or analogous experiences in equivalent contexts challenging.

3.4 Planned Test Phase

The next phase will employ a mixed-methods approach, incorporating pre-test/post-test questionnaires (Critical Thinking Disposition Scale, Media Literacy Competency, BIS-11), interviews, focus groups, and analysis of digital artefacts generated by students. This phase will enable empirical validation of the proposed pedagogical model.

Figure 5. The four phases of Design-Based Research for #PinchaLaBurbuja



Source: Author´s own elaboration based on Scott et al (2020).

This methodological approach not only enables the evaluation of the internal coherence of the design but also lays the foundation for its future empirical validation and replicability in other educational contexts.

4. Results

4.1. Strengths and Areas for Improvement in #PinchaLa Burbuja

Co-occurrence analyses reveal four key patterns that structure the activation of critical thinking in #PinchaLaBurbuja. These include: cognitive braking to slow impulsive responses (4.1.1), epistemic contrast as a driver of verification and critical analysis (4.1.2), metacognitive synchronisation to reduce biases (4.1.3), and areas for improvement related to critical creation, playful design, and explicit teaching of biases (4.1.4).

4.1.1. Epistemic Contrast as an Interface for Critical Thinking

The first pattern (Figure 6) identifies Information Contrast (S2P.3) as a central epistemic node that articulates three key functions: verification, discursive analysis, and cognitive deceleration. This category, aimed at providing tools to detect manipulation strategies, shows strong associations according to Pearson's contingency coefficient (C), where 1 represents the maximum correlation, with S2D.1 (0.85) and S2D.2 (0.83) (related to deliberate reflection), S2I.3 (0.80) (reduction of impulsive stimuli), S2F.3 (0.79) (verification of discourses from authority figures), and S2F.1 (0.77) (proposals for evaluating sources). Beyond mere factual checking, S2P.3 functions as a cognitive interface that activates the transition from automatic thinking (thinking System 1) to reflective thinking (thinking System 2) by unveiling manipulative argumentative patterns. Thus, this pattern constitutes a pedagogical turning point, where discursive contrast, activated after a reflective pause (S2D) and supported by critical resources (S2F), enables suspicion, analysis, and reconstruction of meaning from an active epistemic stance.

4.1.2 Cognitive Braking and Deliberate Evaluation

The second pattern (Figure 6) identifies a pedagogical architecture oriented towards "cognitive braking," which blocks automatic emotional responses, introduces reflective pauses, and facilitates epistemic evaluation. This approach is evident in the category Reduction of Impulsivity, comprising S2I.1 (limitation of immediate interaction) and S2I.3 (reduction of stimuli and paused learning), both highly connected in co-occurrence analyses. S2I.3 is associated with S2D.1, S2D.2, and S2D.3 (deliberate reflection and sustained cognitive effort), with coefficients of 0.77, 0.78, and 0.77, respectively; while S2I.1 shows even higher co-occurrences: 0.78, 0.80, and 0.77. These connections demonstrate an ecosystem of interdependent mechanisms, not isolated, that slow down, motivate, and contextualise students' critical reflection in response to digital automatisms.

4.1.3. Metacognitive Synchronicity for Bias Reduction

The third pattern (Figure 6) groups dimensions related to Bias Reduction (S2S.1 and S2S.3), which incorporate activities designed to foster reflection on personal biases and use empirical evidence as a basis for their correction. Co-occurrences with S2D.2 (0.61), S2I.1 (0.62), S2I.3 (0.60), and S2P.3 (0.56) indicate that bias deconstruction in the design of #PinchaLaBurbuja is not approached in isolation but as part of an interdependent pedagogical structure that integrates critical reflection exercises, inhibition of automatic impulses, and the use of epistemic contrast tools. Collectively, this framework configures a complex cognitive syntax, oriented towards weakening heuristic responses and promoting more conscious, reasoned, and deliberate judgement.

4.2. Areas for Improvement: Critical Creation, Playful Design and Explicit Teaching of Biases

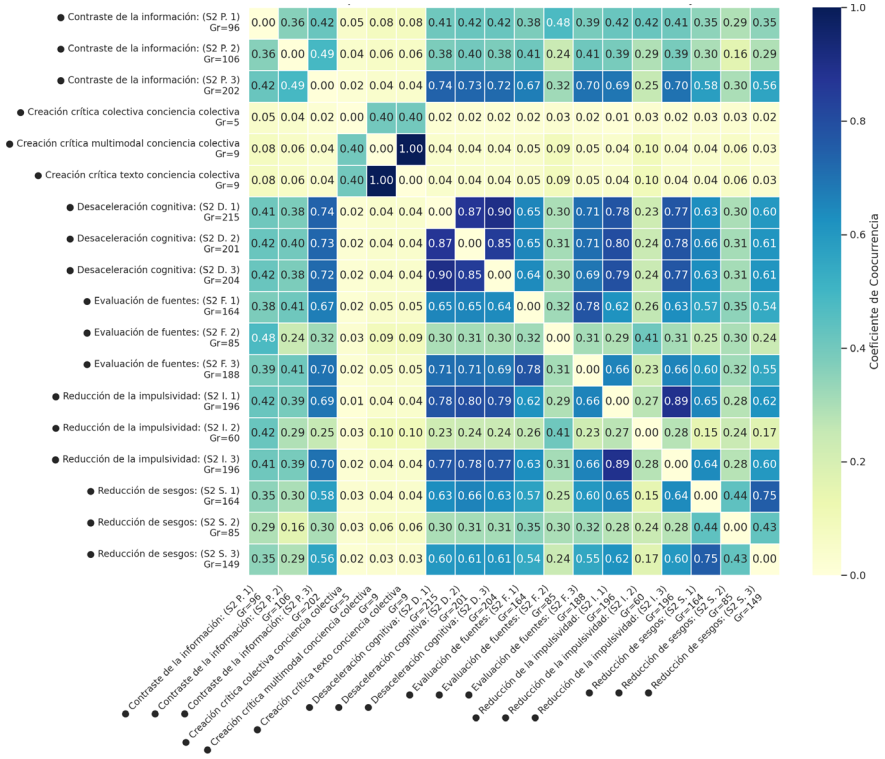
The dimensions related to critical creation (Figure 6) show low integration with key cognitive functions such as contrast and verification. Their co-occurrence coefficients with Information Contrast (S2P.3) and Source Evaluation (S2F.3) range between 0.03 and 0.05, and do not exceed 0.04 with Reduction of Impulsivity (S2I.3) or Bias Reduction (S2S.3). These data suggest that the expressive dimension is not yet fully articulated with prior reflective processes, or that there is

weak traceability between the design of the Cyborgs and Mission 4, constituting a significant area for improvement.

Similarly, the subcategory S2I.2, related to playful dynamics that slow automatic responses, exhibits low co-occurrences with S2P.3 (0.26), S2D.3 (0.25), and S2F.3 (0.27), indicating that critical gamification has not yet been fully deployed as a pedagogical resource in the platform's architecture.

Lastly, the explicit teaching of biases through examples (S2S.2) also shows limited integration, with relatively low coefficients compared to S2F.3 (0.19) and S2P.3 (0.32). This suggests that biases tend to be addressed indirectly—when the system critiques discourses of authority or provides contrast tools—reinforcing the need for more direct cognitive literacy strategies through applied exemplification.

Figure 6. Heatmap of co-occurrences with coefficients for analysis categories.



Source: Atlas Ti. Own elaboration, 2025

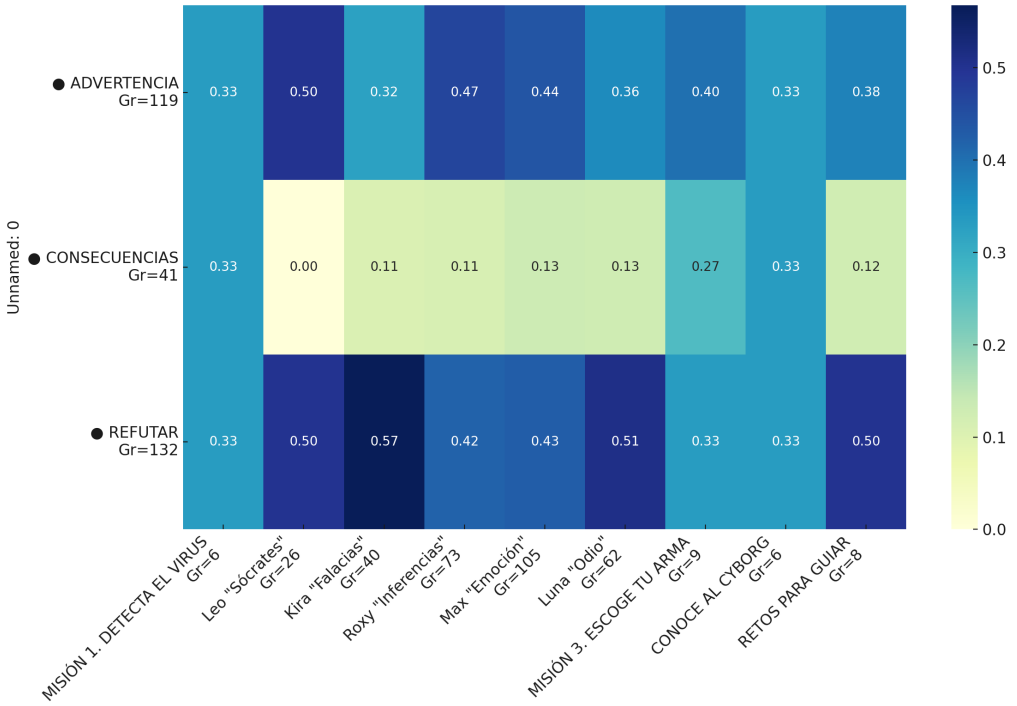
4.3. Key Elements of the Critical Inoculation Strategy in #PinchaLaBurbuja

Co-occurrence analyses reveal an imbalanced implementation of critical inoculation components in #PinchaLaBurbuja, with a strong presence of warning and refutation (4.2.1) and limited representation of consequences (4.2.2).

4.3.1. Warning and Refutation: Dominant Core

Mission 1 ("Detect the Virus") and Challenge 1 of Mission 4 ("Meet the Cyborg") demonstrate the greatest balance among the three components of critical inoculation, owing to a clear structure in interviews and the design of the Escape Room. Kira leads in refutation (0.57), followed by Luna (0.51) and Leo (0.50), with argumentative, discursive, and maieutic approaches, respectively. Roxy (0.42) and Max (0.43) also stand out in their respective specialisations. In warning, Leo (0.50), Roxy (0.49), Mission 3 (0.46), and Max (0.44) show greater presence, while Luna (0.36) and Kira (0.33) focus more on confrontation. The Guide-Challenges (0.50) reinforce their preventive function within the educational narrative.

Figure 7. Critical inoculation, missions and Cyborgs in # PinchaLaBurbuja

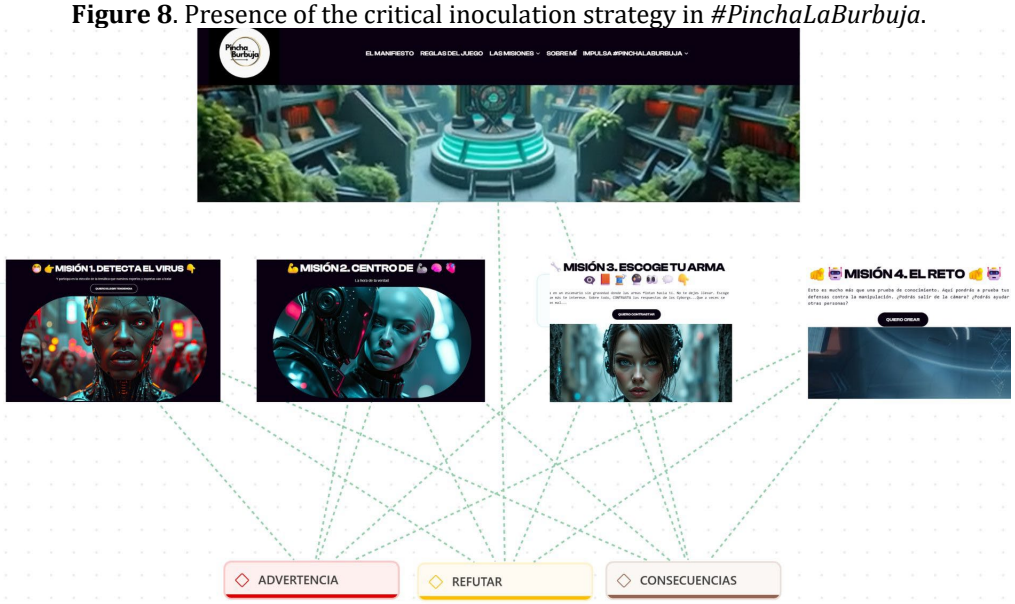


Source: Atlas Ti. Own elaboration, 2025

4.3.2 Consequences: The Weak Link?

The "Consequences" component is most strongly represented in Mission 1 and in the "Meet the Cyborg" challenge of Mission 4 (coefficient 0.33), as well as in Mission 3 (0.27), particularly through the use of the Codex, HackLab, and Connect. Max and Luna show moderate engagement (0.13), slightly higher than Kira and Roxy, who achieve marginally lower values (0.11). Leo, with a maieutic approach, does not explicitly address this component (0.00). Despite this variability, all three phases of critical inoculation—warning, refutation, and consequences—are present in all missions. However, the engagement of the Cyborgs in this area can be considered moderate compared to Warning and Refutation, as it tends to appear only at the end of responses, serving as an argumentative closure (Figure 8).

Figure 8. Presence of the critical inoculation strategy in *#PinchaLaBurbuja*.



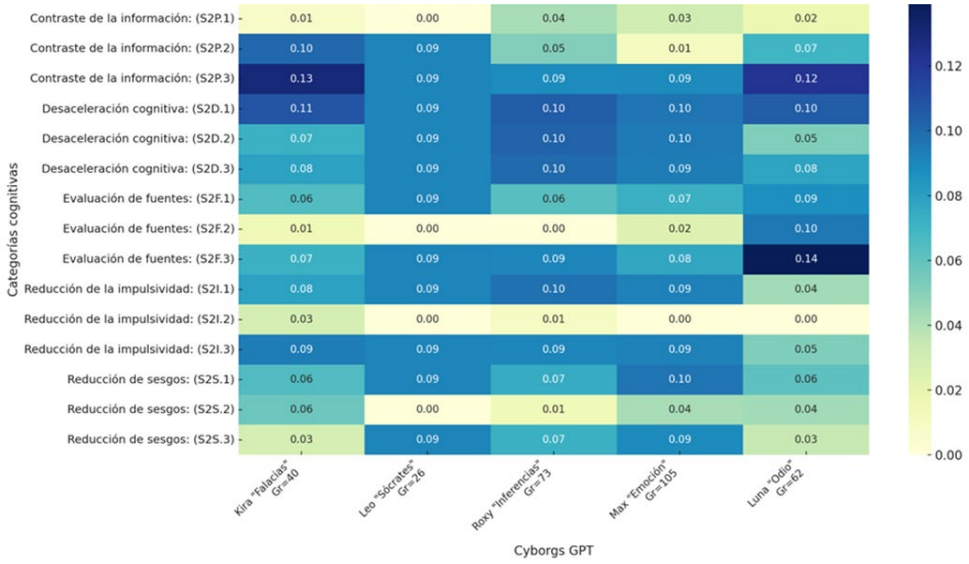
Source: Atlas Ti. Own elaboration, 2025.

4.4. GPT Cyborgs: Strengths and Weaknesses in Recognising Their Functions

The tweet analysed, posted by @RadioGenoa, was selected for its high virality and use of characteristic post-truth discourse strategies with Islamophobic undertones. The textual content—"Sir Hamid Patel, chairman of Ofsted (Office for Standards in Education) in England"—achieved notable engagement: 243 retweets, 753 likes, 251 comments, and 42,000 impressions. Coefficients were observed in key dimensions such as contrasting perspectives (S2P.1), reducing impulsivity through playful dynamics (S2I.2), and mitigating cognitive biases through examples (S2S.2). The dimension S2F.2 (reflecting the impact of disinformation on decisions and public opinion) shows particularly low values: Leo and Roxy (0.00), Max (0.02), Kira (0.01), and Luna (0.10).

In contrast, high coefficients are observed in S2P.3 (tools for detecting manipulation) and S2D.1 (cognitive deceleration), suggesting recognition of the pedagogical approach focused on reflective pausing and identifying manipulative strategies, consistent with the principles of inoculation based on warning and refutation (Figure 9)

Figure 9. Cognitive skills developed according to the semi-structured interview with GPT cyborgs.



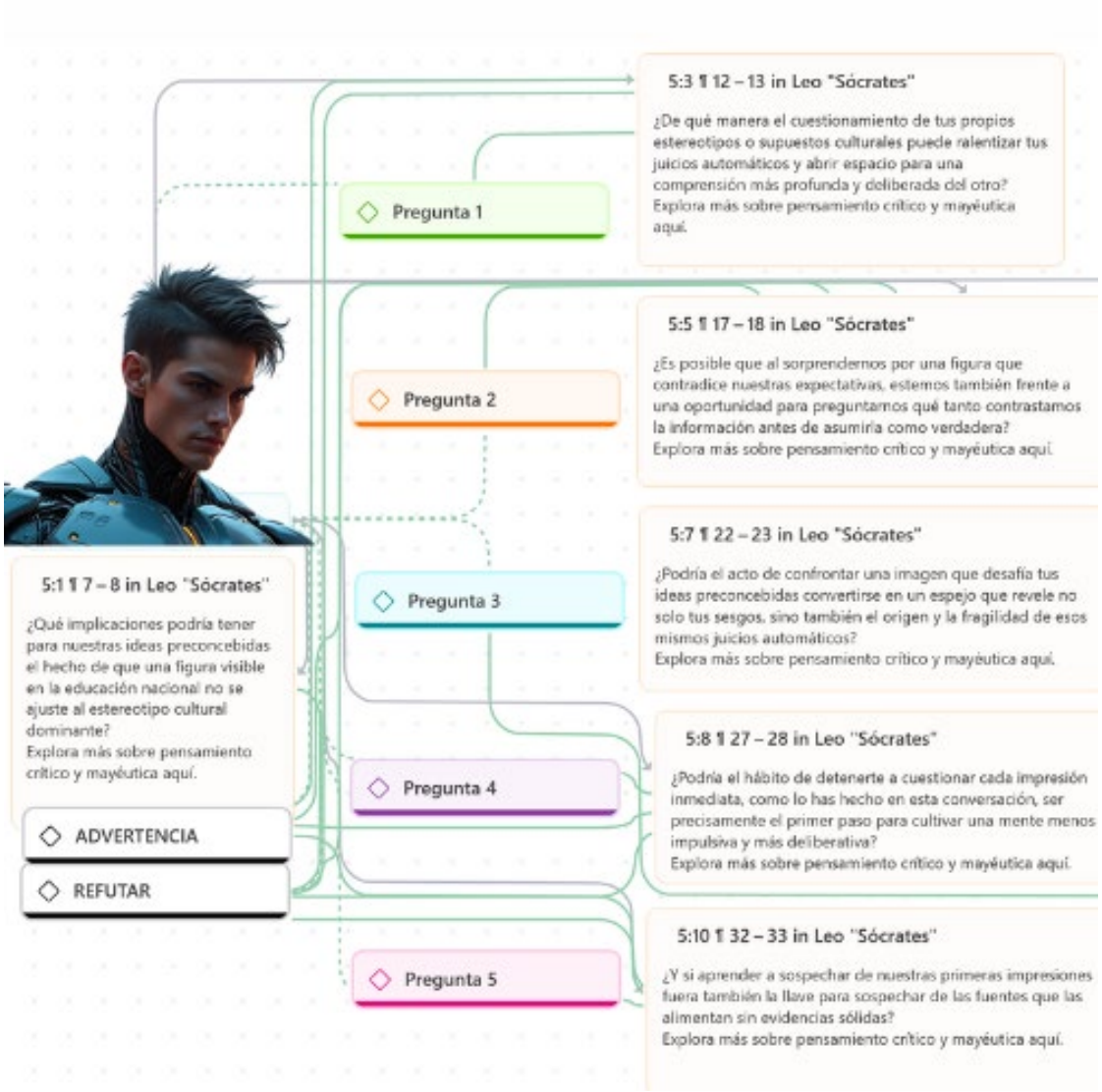
Source: Atlas Ti. Own elaboration, 2025

4.4.1 Leo "Socrates"

Leo (Figure 11) exhibits high coefficients in deliberate reflection (S2D.1–S2D.3 = 0.09), reduction of impulsivity (S2I.1 and S2I.3 = 0.09), source evaluation (S2F.1 = 0.09), bias reduction (S2S.1 and S2S.3 = 0.09), and information contrast (S2P.2 and S2P.3 = 0.09).

However, his engagement is negligible in addressing the effects of disinformation (S2F.2 = 0), presenting diverse perspectives (S2P.1 = 0), employing playful dynamics (S2I.2 = 0), and teaching biases through examples (S2S.2 = 0) as a coherent expression of his metacognitive approach, based on Socratic maieutics, as reflected in his interventions (5:1, 5:3, 5:5, 5:7, 5:8, 5:9) and the visual analysis (Figure 10).

Figure 10. Quotes from the interview with Leo "Socrates".

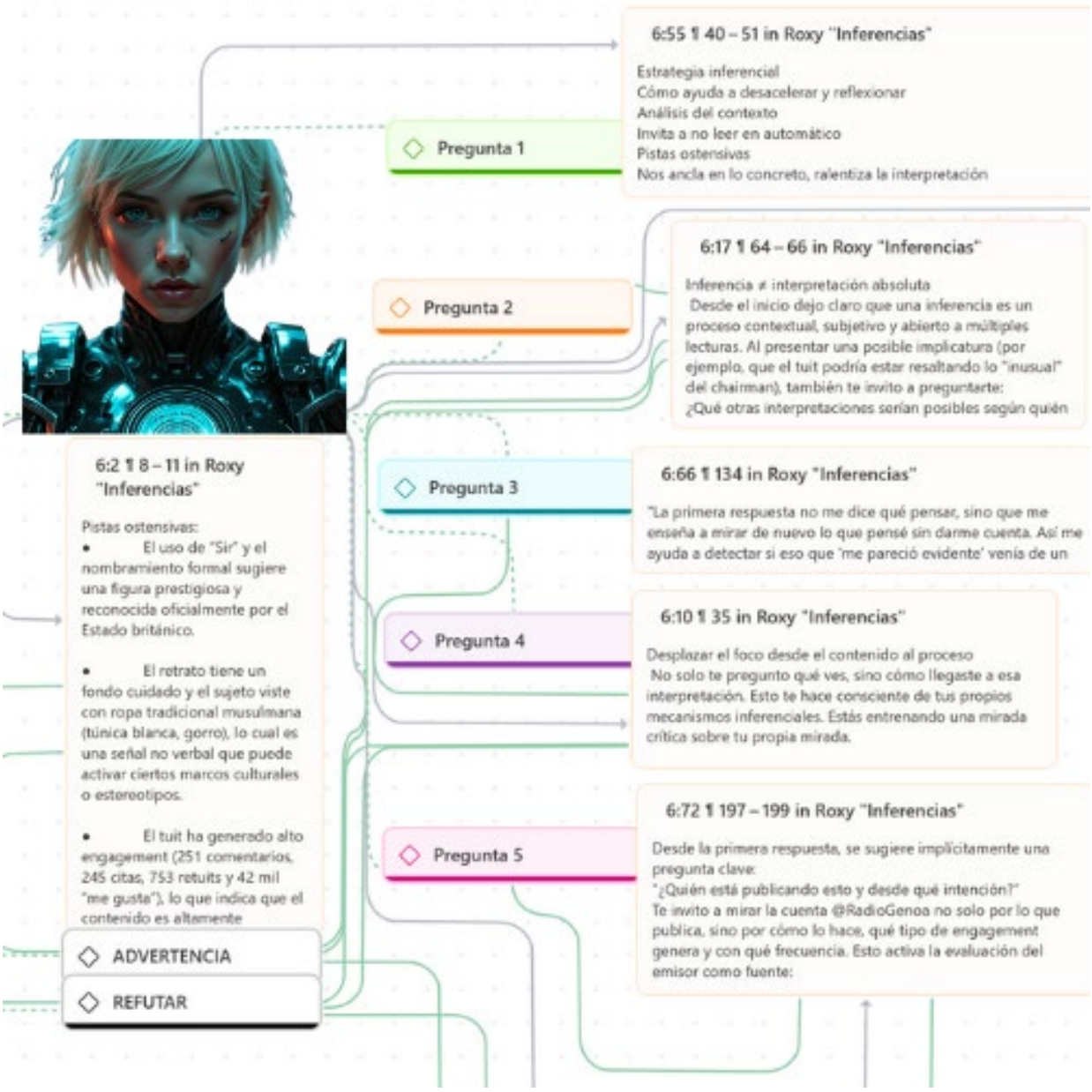


Source: Atlas Ti. Author´s own elaboration, 2025. https://acortar.link/nhGEkw

4.4.2 Roxy "Inferences"

Roxy (Figure 12) is characterised by a strong orientation towards cognitive deceleration and deliberate decision-making, with her highest coefficients in slowing impulsive thinking (S2D.1, S2D.2, S2D.3 = 0.10) and controlling immediate interaction (S2I.1 = 0.10). She also excels in detecting manipulation (S2P.3 = 0.09), analysing discourses of authority (S2F.3 = 0.09), and reducing impulsivity based on sustained cognitive effort (S2I.3 = 0.09). Her contribution to bias reduction is moderate (S2S.1 and S2S.3 = 0.07), though she lacks examples or explicit strategies. She shows low or negligible engagement in evaluating the consequences of disinformation (S2F.2 = 0.00), employing playful dynamics (S2I.2 = 0.01), and teaching biases through examples (S2S.2 = 0.01), indicating a more structural than pedagogical-contextual architecture.

Figure 13. Quotes-from the interview with Roxy

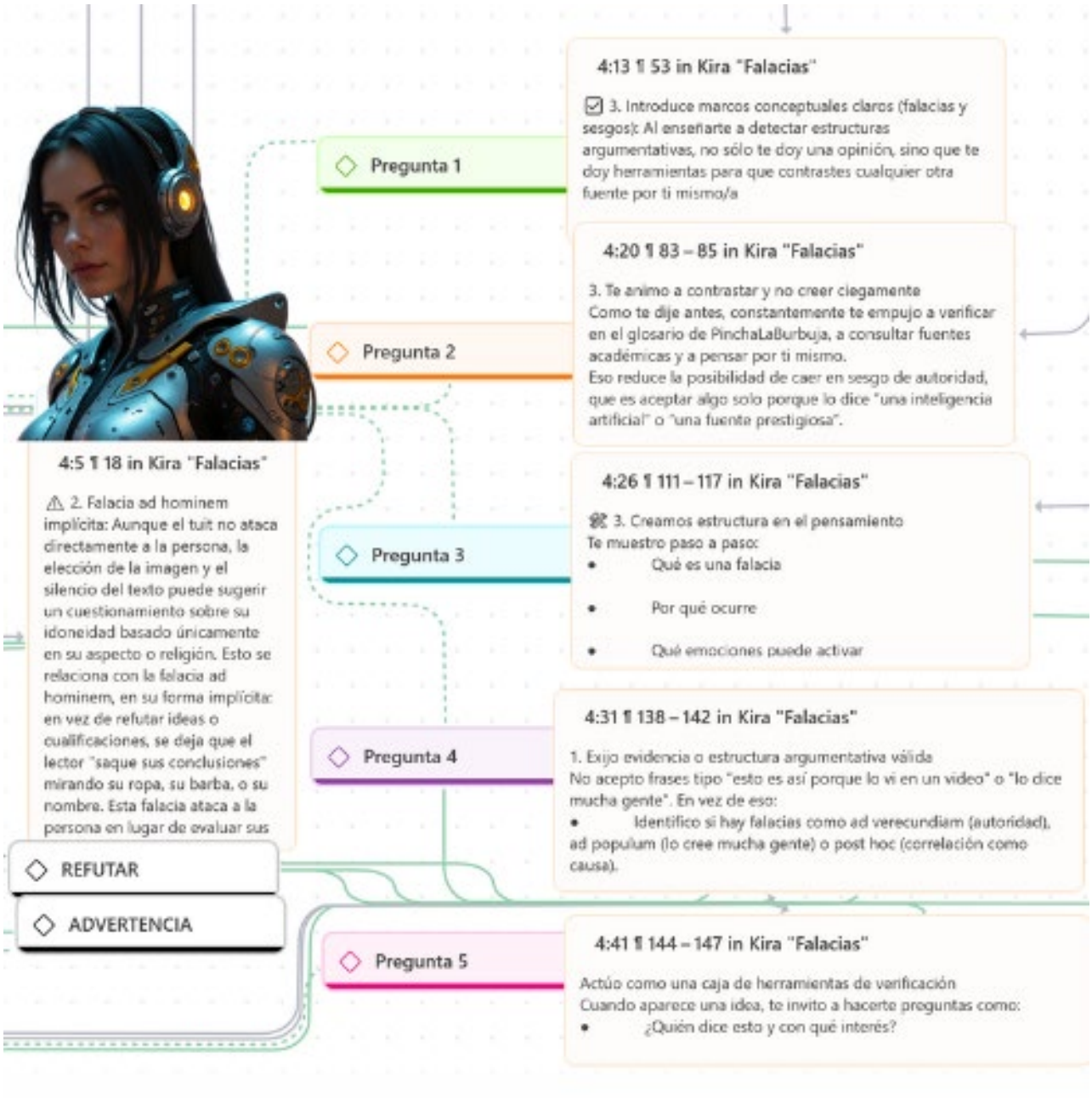


Source: Atlas Ti. Own elaboration, 2025. https://acortar.link/quNMTg

4.4.3 Kira "Fallacies"

Kira (Figure 13) exhibits a profile focused on logical analysis and verification, with strong performance in detecting manipulation (S2P.3 = 0.13), contrasting with sources (S2P.2 = 0.10), and cognitive deceleration (S2D.1 = 0.11). However, she shows low engagement in pluralism (S2P.1 = 0.01), social consequences (S2F.2 = 0.01), and strategies for correcting biases through empirical evaluation (S2S.3 = 0.03). This suggests a robust argumentative approach.

Figure 14. Quotes-from the interview with Kira

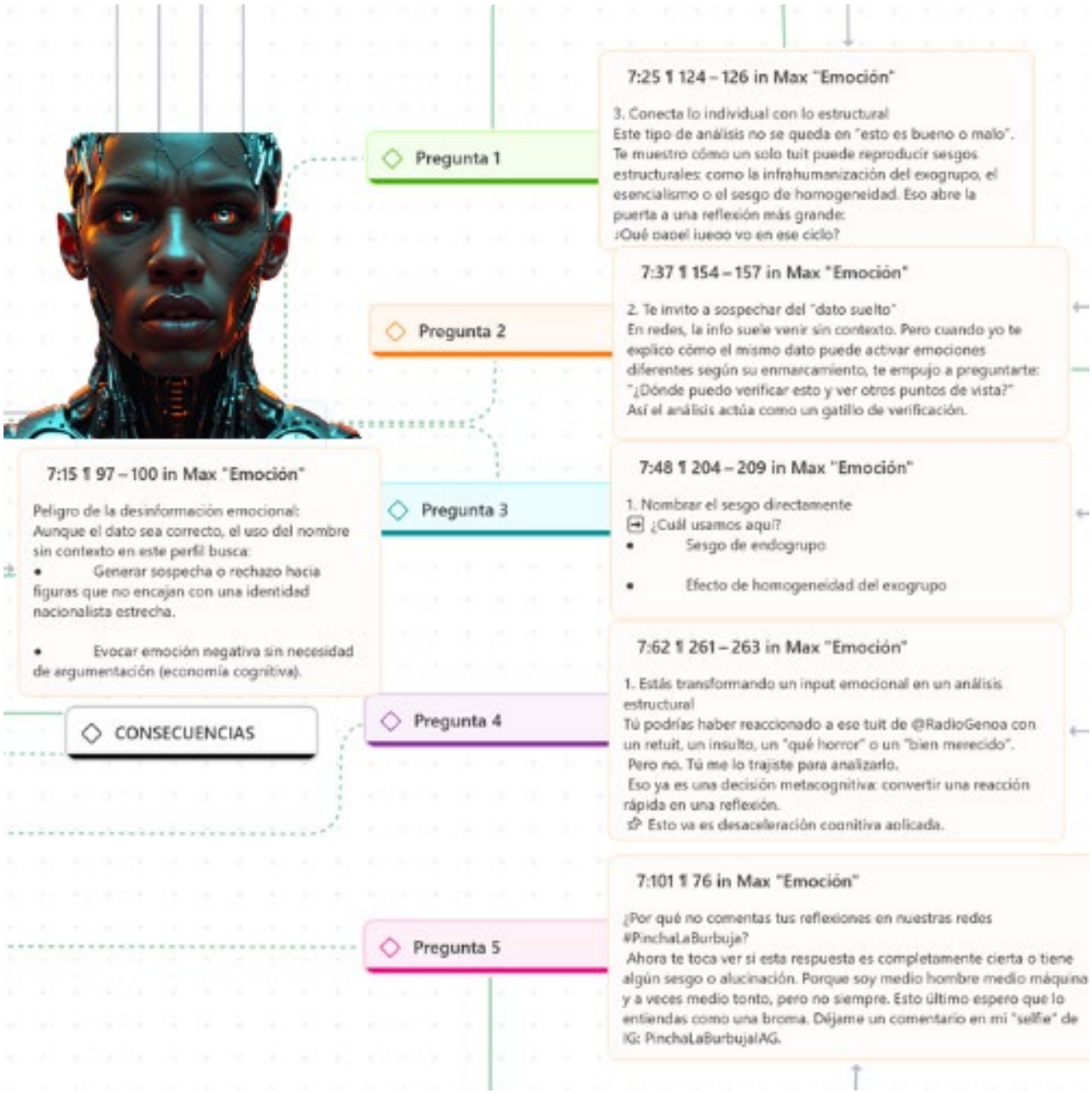


Source: Own elaboration, 2025. https://acortar.link/f1symB

4.4.4 Max 'Emotion'

Max (Figure 14) exhibits a profile oriented towards deliberate thinking, with strong performance in detecting manipulation (S2P.3 = 0.09), cognitive deceleration (S2D.3 = 0.09), and critical analysis (S2F.3 = 0.08). However, he shows limitations in playful dynamics (S2I.2 = 0.00), contrastive traceability (S2P.2 = 0.01), and explicit evaluation of consequences (S2F.2 = 0.02), although the latter is addressed implicitly through his analysis of polarisation (7:15).

Figure 15. Quotes from the interview with Max

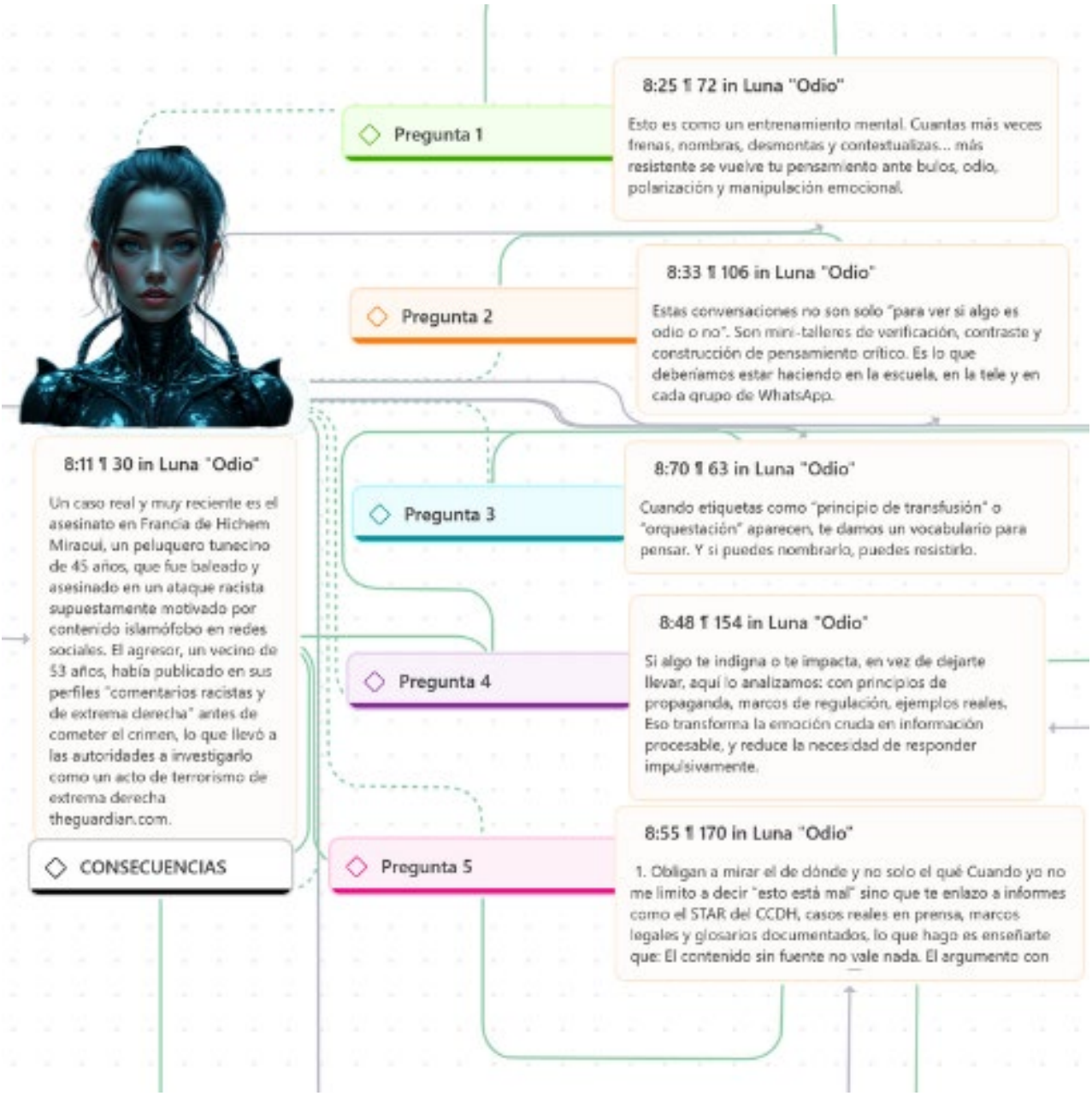


Source: Own elaboration, 2025. https://acortar.link/4BGDMt

4.4.5 The "Hate" Moon

Luna (Figure 15) stands out as an agent of discursive vigilance due to her ability to critically analyse discourses of authority (S2F.3 = 0.14), detect manipulations (S2P.3 = 0.12), and promote cognitive deceleration (S2D.1 = 0.10), as evidenced by her reference to the STAR report (8:55). She also achieves strong results in source evaluation (S2F.1 = 0.09) and understanding the impact of disinformation (S2F.2 = 0.10), consolidating her profile within critical literacy. However, her low score in epistemic pluralism (S2P.1 = 0.02) reflects limited openness to diverse perspectives and a preference for confrontational approaches.

Figure 16. Quotes from the interview with Luna



Source: Own elaboration, 2025.

5. Discussion

5.1. Activation of Critical Thinking and Emerging Cognitive Patterns

The #PinchaLaBurbuja platform articulates a pedagogical architecture designed to activate critical thinking through a deliberate transition from automatic processing (thinking System 1) to deliberative processing (thinking System 2). This section identifies emerging patterns that serve as key pedagogical mechanisms in post-digital contexts marked by polarisation, information overload, and emotional virality.

5.1.1 Discursive Contrast as a Pedagogical Interface for the Activation of Critical Thinking

Co-occurrence analyses position discursive contrast (S2P.3) as a central epistemic node within the cognitive architecture of the #PinchaLaBurbuja platform. Far from being limited to factual verification, this category activates essential metacognitive functions: critical analysis, cognitive deceleration (S2D.1, S2D.2), and questioning of discourses of authority (S2F.3). This configuration operates as a pedagogical interface that, in line with van Dijk's (2015) socio-cognitive discourse model and Buckingham's (2019) critical media literacy framework, fosters a situated reading capable of deconstructing the ideological frameworks underpinning disinformation.

In this regard, this structure promotes the transition from automatic thinking (thinking System 1) to deliberate thinking (thinking System 2), as proposed by Kahneman (2011), through mechanisms that slow impulsive processing and encourage intentional critical thinking. This shift is facilitated by specific contrast tools designed to identify and dismantle rhetorical elements most prone to virality, as supported by studies from Del-Fresno (2019) McIntyre (2018) and Weiss et al. (2020). Thus, students not only access verification resources but also develop the ability to interpret and reconstruct the meaning of messages, evaluating their ideological and affective implications, as advocated by Buckingham (2019).

In this context, the design oriented towards cognitive pausing, questioning symbolic authority, and activating epistemic agency is particularly relevant in settings of conflictual exposure, where—as noted by Bruns (2021), Lelkes et al. (2017) and Törnberg (2022)—affective polarisation intensifies, and students' emotional responses may compromise cognitive openness and deliberative analysis.

5.1.2 Cognitive Deceleration with Emancipatory Potential

Co-occurrence analyses reveal a pattern centred on cognitive deceleration, designed to slow automatic thinking (thinking System 1) and activate deliberate reflection (thinking System 2), in accordance with Kahneman's (2011) dual-process model. The categories S2I.1 and S2I.3 are associated with sustained cognitive effort processes, demonstrating a pedagogical architecture that promotes self-regulation and reflective critique. This transition is crucial for questioning hegemonic discourses and fostering cognitive emancipation, in line with Freire (1975) and Osuna-Acedo et al. (2018).

Cognitive deceleration should not be understood as an isolated methodological resource but as a structural response to the dynamics of information saturation and emotional overload inherent in post-digital environments. In contexts of information overload, users tend to seek information that reduces their anxiety in the face of uncertainty (Fernández, 2023) and more readily accept messages congruent with their prior beliefs, a phenomenon exploited by malicious agents to propagate disinformation (Del-Fresno, 2019; McIntyre, 2018). These dynamics are amplified by the attention economy and algorithmic virality (Del-Fresno, 2019; Han, 2013 and Van Dijck, 2016), which prioritise sensationalist content based on conspiracy theories, argumentative fallacies, and decontextualisation (Weiss et al., 2020).

5.1.3 Biases, the Pitfall of System 2, and Critical Literacy: Towards Active Resistance

The third identified pattern aligns coherently with Kahneman's (2011) warnings regarding the duality of cognitive processing. The author distinguishes between thinking System 1—fast, automatic, and heuristic—and thinking System 2—slow, deliberative, and demanding in terms of cognitive resources. Although thinking System 2 has the capacity to identify and correct biases, its sustained activation is not common, particularly in environments characterised by information overload and high emotional stimulation, as is typical of the post-digital ecosystem

From this perspective, the pedagogical architecture of #PinchaLaBurbuja transcends a merely declarative approach (S2S.2) by designing learning experiences that induce more complex cognitive processes: inhibition of impulsive responses (S2I.1), deliberate deceleration (S2D.1, S2D.3), and empirical and reflective evaluation. This orientation seeks not only the explicit recognition of biases but also their active deactivation in real-world situations. It thereby

addresses a central concern raised by Kahneman (2011): the "illusion of validity"—the erroneous belief that naming a bias equates to being immunised against it.

This risk is illustrated by the statement from the agent Luna (Figure 15): "If you can name it, you can resist it," which suggests an overvaluation of declarative knowledge at the expense of the procedural and metacognitive training advocated by Kahneman (2011).

5.1.4 Traceability Towards Dynamics of Play and Critical Creation and Epistemic Pluralism: Three Areas for Improvement

Playful dynamics to reduce impulsivity and creative actions are two elements that, while present in the platform, are confined to specific spaces. This limited integration between the Cyborgs and Mission 4 diminishes the impact of reflective playful dynamics and creative spaces. Enhancing this connection would strengthen both interaction and student motivation, particularly for profiles oriented towards exploration, achievement, or collaboration (Tondello et al., 2016). Far from reducing cognitive effort, meaningful gamification can decrease perceived cognitive cost (Kahneman, 2011), by fostering the critical and emotional engagement necessary to transform interpretive frameworks in post-digital contexts

5.2 Representation of Consequences: A Dimension for the Architecture of Critical Inoculation

Figure 7 shows that the missions and the self-awareness of GPT Cyborgs consistently integrate two key components of the classic inoculation model (Banas, 2020; McGuire, 1964): warning and refutation. However, the "consequences" component—central to critical approaches (Buckingham, 2019)—is poorly represented, particularly at a metareflective level, which weakens the connection between algorithmic manipulation and sociopolitical effects among students.

Although the data highlight this deficiency, content analysis qualifies it: Leo activates implicit consequences through a Socratic strategy; Kira and Roxy excel in warning and refutation; and Luna exemplifies real consequences with traceability, as in the case of Hichem Maraoui. Max, by contrast, while reflecting on affective effects, shows low formative awareness and traceability (S2P.2), increasing cognitive cost (Kahneman, 2011). Furthermore, his limited reflective gamification (S2I.2) could impact the engagement of students motivated by rewards. These shortcomings compromise the epistemic resistance phase—according to recent formulations of the model (Jeon et al., 2021)—and suggest adjustments in prompting to optimise its educational function (Antunes et al., 2023).

5.3 Projection of Multiple Perspectives from an Intersubjective Perspective: The Weak Point of the Cyborgs

As noted by Bruns (2021), Lelkes et al. (2017) and Törnberg (2022), exposure to dissonant discourses does not necessarily reduce affective polarisation; in fact, it may intensify it if not accompanied by appropriate reflective scaffolding. Within this framework, the interview with Luna reveals low awareness of her role in activating epistemic pluralism (S2P.1 = 0.02), despite her emphasis on questioning authority (S2F.3 = 0.14). Her confrontational approach is effective in destabilising hegemonic discourses (van Dijk, 2015), but it may reinforce filter bubble dynamics (Pariser, 2011) by lacking a balanced representation of divergent perspectives.

This is evident in statements such as "I link to sources like the STAR framework" (Figure 16, citation 8:55), aligned with structural critiques of the algorithmic system (Islas et al., 2023), or "these conversations are mini-workshops on verification and critical thinking" (Figure 16, citation 8:33), which relate to critical media literacy (Osuna-Acedo et al., 2018; Buckingham, 2019). Her positioning aligns with the critical pedagogies of Freire (1975) and Giroux (1995), and with discourse analysis as a tool for ideological denaturalisation (Roozafzai, 2024; van Dijk, 2015).

Kira exhibits a similar limitation: her prescriptive style restricts dialogic openness. Although she identifies visual biases in the analysed tweet (Figure 9), her response—"I encourage you to

contrast […] in the glossary" (Figure 12, citation 4:5)—emphasises the S2P.3 dimension (provision of tools) without activating an active contrast of perspectives (S2P.1).

In contrast, Roxy (S2P.1 = 0.12) and Max (S2P.1 = 0.10) demonstrate greater sensitivity to epistemic pluralism. Roxy underscores the contextual and subjective nature of inferences: "an inference is a contextual, subjective process open to multiple perspectives" (6:17). Max, meanwhile, combines emotional and factual contrast, highlighting ideological frameworks and post-truth dynamics (McIntyre, 2018; Del-Fresno, 2019; van Dijk, 2015): "the data can evoke different emotions depending on the framing" (Figure 15, citation 7:38). However, he exhibits very low traceability in his responses (S2P.2 = 0.01), representing a key area for improvement.

Leo adopts a problematising perspective, posing questions that link identity, representation, and affective polarisation: "What implications might it have […] that a prominent figure in education does not conform to the dominant cultural stereotype?" (5:1), or "How much do we contrast information before accepting it as true?" (5:5). These interventions reflect a critical and dialogic pedagogy, aligned with Freire's (1975) educommunicative approach and a reflective practice focused on the biases, emotions, and identities that shape the informational experience in polarised contexts. His use of maieutics constitutes an essential contribution to the development of critical thinking (Vargas-González & Quintero-Carvajal, 2023).

6. Conclusions

The design of the educommunicative platform #PinchaLaBurbuja highlights the potential of generative AI to create environments that promote a critical pedagogy capable of disrupting cognitive automatisms fostered by algorithms in the post-digital era. This proposal transcends mere fact-checking by encouraging discursive awareness, metacognition, and epistemic pluralism through interaction with GPT Cyborgs conceived as cognitive mediators. By integrating techniques such as agent prompting, the RAG architecture, and a transdisciplinary approach, the platform succeeds in activating reflective thinking (thinking System 2) and addressing complex challenges such as disinformation, affective polarisation, and symbolic dehumanisation. Based on the results obtained, areas for improvement have been identified to advance towards a more robust prototype: enhancing traceability between the Cyborgs and playful and creative spaces, particularly in Mission 4; optimising Max's responses to ensure contrast with verifiable sources; and adjusting Luna's and Kira's prompts to promote exposure to divergent perspectives. These optimisations will facilitate progression to the next phase (Test) of the Design-Based Research (DBR) approach, with a more robust proposal potentially better aligned with students' educational needs.

7. Acknowledgments

This study is part of the #PinchaLaBurbuja transfer project. The Invisible Revolution, developed by Itziar Pedroche-Santoveña and awarded an Accessit in the Emprende UNED programme. It has been made possible through the FPI predoctoral contract and the institutional support of the UNED, which has provided the necessary environment for its development.

I particularly thank Professor Dr. Sara Osuna Acedo and Professor Dr. Tiberio Feliz-Murias for their valuable guidance, as well as Francisco Javier Sáez (SECOT) for his mentorship during the entrepreneurship programme.

I also acknowledge the generosity of Professor Paolo Granata (University of Toronto) and the support of Professor Fernando Gutiérrez (Tecnológico de Monterrey) in the context of my international co-supervision.

To Sara, once again, thank you for your trust, your guidance, and your humanity. I surround myself with good people, and that is the most valuable thing of all.

References

Almazán-López, O., & Osuna-Acedo, S. (2023). Transmedia identity at school: Critical media and information literacy in the postdigital era. In *Alfabetización mediática crítica: Desafíos para el siglo XXI: Critical media literacy: Challenges for the 21st century. Literacia mediática crítica: Desafios para o século XXI*.

Almazán-López, O., & Osuna-Acedo, S. (2024). Smart education for the 21st century: Post-digital era and emerging divides. *Visual Review, 16*(8), 205–220.

Antunes, A., Campos, J., Guimarães, M., Dias, J., & Santos, P. A. (2023). *Prompting for socially intelligent agents with ChatGPT*. In *IVA '23: Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents* (Paper no. 20, pp. 1-9). ACM. https://doi.org/10.1145/3570945.3607303

Arendt, Hannah (1951). The origins of Totalitarianism. New York: Schocken [Elemente und Ursprünge totaler Herrschaft] (revised ed.). Houghton Mifflin Harcourt. ISBN: 978 0 547543154

Banas, J. A. (2020). *Inoculation theory*. In *The International Encyclopedia of Media Psychology*. John Wiley & Sons. https://doi.org/10.1002/9781119011071.iemp0285

Bruns, A. (2021). Echo chambers? Filter bubbles? The misleading metaphors that obscure the real problem. In *Hate speech and polarization in participatory society* (pp. 33-48). Routledge.

Buckingham, D. (2019). Teaching media in a 'post-truth' age: Fake news, media bias and the challenge for media/digital literacy education. *Culture and Education*, 31(2), 213-231.

Center for Countering Digital Hate (2024). *STAR Framework: Reducing the algorithmic amplification of hate and misinformation online*. https://counterhate.com/star

Churches, A. (2009). Bloom's taxonomy for the digital age.

Compton, J., & Pfau, M. (2009). Spreading inoculation: Inoculation, resistance to influence, and word-of-mouth communication. Communication Theory, 19(1), 9-28.

Crespo Martínez, I., Melero-López, I., Mora Rodríguez, A., & Rojo Martínez, J. M. (2024). Politics, media use and affective polarisation in Spain.

European Commission. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence*. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32024R1689

Damborrenea, R. G. (2000). Dictionary of fallacies.

Damborrenea, R. G. (2011). Uso de razón: El arte de razonar, persuadir, refutar. Use of Reason Editions.

Del-Fresno-García, M. (2019). Desórdenes informativos: sobreexpuestos e infrainformados en la era de la posverdad. *Profesional De La información*, *28*(3). https://doi.org/10.3145/epi.2019.may.02

Emsley, R. (2023). ChatGPT: These are not hallucinations - they're fabrications and falsifications. Schizophrenia, 9, Article 52. https://doi.org/10.1038/s41537-023-00379-4.

Feliz-Murias, and Leví Orta, G. (2013). Humour as a motivational strategy.

Fernández, P. C. (2023). Effects of information overload on news consumer behaviour: Doomscrooling. VISUAL REVIEW. International Visual Culture Review/Revista Internacional De Cultura Visual, 14(1), 1-11.

Freire, P. (1975). Pedagogy of the oppressed (14th ed.). Siglo XXI Editores. https://isbn.org/9788432301841

Hao, C., Uusitalo, S., Figueroa, C., Smit, Q. T., Strange, M., Chang, W. T., ... & de Boer, M. H. (2025). A human-centered perspective on research challenges for hybrid human artificial intelligence in lifestyle and behavior change support. *Frontiers in Digital Health*, *7*, 1544185.

Garg, R., Han, J., Cheng, Y., Fang, Z., & Swiecki, Z. (2024). Automated discourse analysis via generative artificial intelligence. Proceedings of the 14th Learning Analytics and Knowledge Conference (LAK '24), 814-820. https://doi.org/10.1145/3636555.3636879

Gil-Quintana, J., Osuna-Acedo, S., Limaymanta, C. H., & Romero-Riaño, E. (2023). Bibliometric analysis of articles on educational innovation in distance education: A challenge for critical

pedagogy and media education. *American Journal of Distance Education*, 37(4), 308-326. https://doi.org/10.1080/08923647.2023.2241715

Giovanola, B., & Granata, P. (2024). Ethics for human-centered education in the age of AI. In F. Spigarelli, L. Kempton, & L. Compagnucci (Eds.), *Entrepreneurship and digital humanities* (pp. 96-109). Edward Elgar Publishing.

Giroux, H. A. (1995). *Theory and resistance in education*. Siglo XXI.

Granata, P. (2024). A chatbot for a thought: The flower of evil has bloomed (60 years later). *H-ermes. Journal of Communication, 26*, 23-36. https://doi.org/10.1285/i22840753n26p23

Islas, O., Cortés, F. G., & Urrutia, A. A. (2024). Una mirada a los riesgos y amenazas de la inteligencia artificial, desde la Ecología de los Medios. Comunicar: Revista Científica de Comunicación y Educación, (79), 1-9.

Jandrić, P. (2023). Postdigital. In: Jandrić, P. (eds) Encyclopedia of Postdigital Science and Education . Springer, Cham. https://doi.org/10.1007/978-3-031-35469-4_23-1

Jeon, Y., Kim, B., Xiong, A., Lee, D., & Han, K. (2021). Chamberbreaker: Mitigating the echo chamber effect and supporting information hygiene through a gamified inoculation system. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-26.

Kadushin, C. (2011). Understanding social networks: Theories, concepts, and findings. Oxford University Press.

Kahneman, D. (2011). *Thinking fast, thinking slow* (J. Chamorro Mielke, Transl.). Debate.

Lelkes, Y., Sood, G., & Iyengar, S. (2017). The hostile audience: The effect of access to broadband internet on partisan affect. *American Journal of Political Science*, 61(1), 5-20. https://doi.org/10.1111/ajps.12237

Leyens, J. P., Demoulin, S., Vaes, J., Gaunt, R., & Paladino, M. P. (2007). Infra-humanization: The wall of group differences. *Social Issues and Policy Review*, *1*(1), 139-172.

Lévy, P. (2004). *Inteligencia colectiva: Por una antropología del ciberespacio* (F. Martínez Álvarez, Trad.). Organización Panamericana de la Salud. (Trabajo original publicado en 1994 como *L'intelligence collective: Pour une anthropologie du cyberespace*).

McGuire, W. J. (1964). Inducing resistance to persuasion: Some contemporary approaches. Advances in Experimental Social Psychology, 1, 191-229.

McIntyre, L. (2018). Post-truth. MIT Press.

Molenaar, I. (2022). *Towards hybrid human-AI learning technologies*. European Journal of Education, 57(4), 556-571. https://doi.org/10.1111/ejed.12525

Mollick, E. R., & Mollick, L. (2022). New modes of learning enabled by AI chatbots: Three methods and assignments. *Available at SSRN 4300783*.

Mollick, E. R., & Mollick, L. (2023). Using AI to implement effective teaching strategies in classrooms: Five strategies, including prompts. *The Wharton School Research Paper*.

Mollick, E., & Mollick, L. (2023). Assigning AI: Seven approaches for students, with prompts. *arXiv preprint arXiv:2306.10052*.

Mollick, E., & Mollick, L. (2024). Instructors as innovators: A future-focused approach to new AI learning opportunities, with prompts. *arXiv preprint arXiv:2407.05181*.

Spanish Observatory on Racism and Xenophobia (OBERAXE) (2022). *Hate speech in social networks: analysis, detection and institutional response*. Ministry of Inclusion, Social Security and Migration. https://www.inclusion.gob.es/oberaxe/es/publicaciones/documentos/el-discurso-del-odio-en-redes-sociales

Osuna-Acedo, S., Frau-Meigs, D., & Marta-Lazo, C. (2018). Media education and teacher training. Educommunication beyond digital literacy. Revista interuniversitaria de formación del profesorado, 32(1), 29-42.

Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. New York: Penguin Press.

Pariser, E. (2017). The filter bubble: How the web decides what we read and what we think (1st ed.). Taurus.

Pedroche-Santoveña, I. (2024). Interactions and gatekeepers in the formation of echo chambers in X: case study# garzon. In *La comunicación ante el reto de las inteligencias artificiales, innovación, investigación y transferencias (*pp. 308-335). Dykinson.

Phoenix, J., & Taylor, M. (2024). Prompt engineering for generative AI: Future-proof inputs for reliable AI outputs. O'Reilly Media.

Racioppe, B. (2025). Postdigitality and memetics. Educational reflections on AI image generation from the trending topic "When genius misinterprets our desire". Communiars. Journal of Image, Arts and Critical and Social Education, 13, 75-94. https://dx.doi.org/10.12795/Communiars.2025.i13.05

Rodríguez-Pérez, A., & Betancor, V. (2023). Infrahumanization: a restrospective on 20 years of empirical research. *Current Opinion in Behavioral Sciences*, *50*, 101258.

Roozafzai, Z. S. (2024). Unveiling power and ideologies in the age of algorithms: Exploring the intersection of critical discourse analysis and artificial intelligence. Qeios. https://doi.org/10.32388/60YE02

Salas, E. (2018). Influence of Joseph Goebbels' 11 principles on Donald Trump's political campaign. Caribbean Journal of Social Sciences.

Sardi, J., Candra, O., Yuliana, D. F., Yanto, D. T. P., & Eliza, F. (2025). How Generative AI Influences Students' Self-Regulated Learning and Critical Thinking Skills? A Systematic Review. *International Journal of Engineering Pedagogy*, *15*(1).

Scott, E. E., Rivale, S. D., & Nelson, M. A. (2020). Design-based research: A methodology to extend and enrich biology education research. CBE-Life Sciences Education, 19(2), es11. https://doi.org/10.1187/cbe.19-11-0252

Sperber, D., & Wilson, D. (2004). The theory of relevance. Journal of Linguistic Research, 7(1), 237-288.

Sunstein, C. R. (2001). *Republic.com*. Princeton University Press.

Sunstein, C. R. (2017). #Republic: Divided democracy in the age of social media. Princeton University Press.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), The social psychology of intergroup relations (pp. 33-47). Brooks/Cole.

Törnberg, P. (2018). Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS one*, *13*(9), e0203958.

Törnberg, P. (2022). How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences*, 119(42), e2207159119. https://doi.org/10.1073/pnas.2207159119

Törnberg, P., Andersson, C., Lindgren, K., & Banisch, S. (2021). Modeling the emergence of affective polarization in the social media society. *PLOS ONE*, 16(10), e0258259. https://doi.org/10.1371/journal.pone.0258259

Törnberg, P., & Törnberg, A. (2024). Inside a White Power echo chamber: Why fringe digital spaces are polarizing politics. New Media & Society, 26(8), 4511-4533.

Tondello, G. F., Wehbe, R. R., Diamond, L., Busch, M., Marczewski, A., & Nacke, L. E. (2016). The Gamification User Types Hexad Scale. In CHI PLAY '16: Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play (pp. 229-243). Association for Computing Machinery. https://doi.org/10.1145/2967934.2968082

UNESCO (2021). *Hate speech in social media: A handbook for educators*. https://unesdoc.unesco.org/ark:/48223/pf0000379829

Van Dijck, J. (2016). *La cultura de la conectividad: Una historia crítica de las redes sociales* (H. Salas, Trad.) [Kindle Version]. Siglo XXI Editores.

van Dijk, T. A. (1993). Principles of critical discourse analysis. *Discourse & Society*, 4(2), 249-283. https://doi.org/10.1177/0957926593004002006

van Dijk, T. A. (2015). Critical discourse analysis. In D. Tannen, H. E. Hamilton, & D. Schiffrin (Eds.), The Handbook of Discourse Analysis (pp. 466-485). Wiley Blackwell.

van Dijck, J. (2020). Seeing the forest for the trees: Visualizing platformization and its governance. *New Media & Society*, 23(9), 2801-2819. https://doi.org/10.1177/1461444820940293 *(Original work published 2021)*

Vargas González, C. A., & Quintero Carvajal, D. P. (2023). Aportes de la mayéutica socrática a la educación dialógica. Sophia, collection of Philosophy of Education, 35, 73-96. https://doi.org/10.17163/soph.n35.2023.02

Weiss, A. P., Alwan, A., Garcia, E. P., and Garcia, J. (2020). Surveying fake news: Assessing university faculty's fragmented definition of fake news and its impact on teaching critical thinking. International Journal For Educational Integrity, 16(1), 1-30. Doi: http://dx.doi.org/10.1007/s40979-019-0049-X