F*ck Around and Find Out: Quasi-Malicious Interactions with LLMs as a Site of Situated Learning

Sarah O'Neill
Business Academy Copenhagen
Department of Research and Innovation
Copenhagen, Denmark
SARO@EK.DK

Abstract

This work-in-progress paper proposes a cross-disciplinary perspective on "malicious" interactions with large language models (LLMs), reframing it from only a threat to be mitigated, we ask whether certain adversarial interactions can serve as productive learning encounters that demystify the opaque workings of AI systems to novice users. We ground this inquiry in an anecdotal observation of a participant who deliberately sabotaged a machine-learning robot's training process in order to understand its underlying logic. We outline this observation with a conceptual framework for learning with, through, and from the interactions with LLMs, grounded in Papert's constructionism and Hasse's ultra-social learning theory. Finally, we present the preliminary design of a research-through-workshop event where AI-novices will jailbreak various LLM chatbots, investigating this encounter as a situated learning process. We share this early-stage research as an invitation for feedback on reimagining inappropriate and harmful interactions with LLMs not merely as problems, but as opportunities for engagement and education.

1 Introduction

As generative AI systems become integrated across sectors and job functions, they are reshaping how work is valued, managed, and monitored. Despite narratives portraying automation as liberation from drudgery, workers increasingly encounter AI as a source of deskilling, heightened control, and opaque criteria of evaluation, and their agency is often framed simply as a choice between harnessing AIs power or being 'left behind', a framing that individualises risk while masking structural shifts in power, responsibility, and knowledge (Nguyen and Mateescu, 2024).

We argue that workers deserve structured spaces for critical examination of the LLM systems they are supposed to "harness". Rather than training workers to comply with tools whose operations and logic remain hidden, the competence model that our research aims to inform, proposes that by intentionally provoking and subverting LLM behaviours, professionals can cultivate the capacity to engage critically and responsibly with AI in their work. By developing a competence model that focuses on critical understanding, this research also aims to foster "ethical and professional norms and workplace standards" that protect workers' dignity, autonomy, and right to meaningfully participation in shaping the role of AI in their field. As AI becomes more deeply integrated into work infrastructures, upskilling must equip workers, not only to handle technological change but also to ask who benefits, how their knowledge is used, and what futures they wish to co-create. This project draws inspiration from a performative HCI setup where a participant "went rogue" and deliberately sabotaged the intended interaction to probe the bot's machine-learning mechanism. Rather than dismissing this outlier event, we treat it as anecdotal evidence of a distinct form of learning interaction, one that could foster critical reflection, curiosity, and situated understanding among novice LLM users.

The paper proceeds as follows: we recount the motivating observation (Section 2), situate it within social learning and intra-action frameworks (Section 3), outline a workshop design (Section 4), and inviting feedback on both the proposed design and its underlying assumptions (Section 5).

2 Empirical background: Serendipitous observations of malicious interaction

The experimental workshop that we propose in Section 4 is designed to extract and investigate the potential revealed in a serendipitous insight-generating glitch (Juarez, 2022) that oc-

Proceedings of the First Interdisciplinary Workshop on OMMM associated with RANLP 2025,
pages 53–58, Varna, Bulgaria, Sep 8-10, 2025.
https://doi.org/10.26615/978-954-452-101-1-006

curred during a performative (Sørensen, 2007) reinterpretation of a HCI experiment from the paper "Why Robots Should Be Social". Through a 50-round language game interaction, the WRSBS experiment had human participants, and a robot simulate a teacher–student scenario in order to study humans' ability to and motivation towards teaching a "learning" robot(De Greeff and Belpaeme, 2015).

In each round, both the human and the robot were shown three different animal images. In the experiment, the robot went first by expressing a "novelty preference" (De Greeff et al., 2009) by exclaiming one of 12 preset "phrases of interest" like "I'd like to learn this one!" while performatively fixating its gaze on the most novel of the three animals. Silently, the human then picked one of the three animals as the "topic" of that round. Without revealing which animal was chosen, the human selected the appropriate category label (e.g. "mammal") from a list of seven options on a touchscreen. The robot then tried to guess which of the three animals belonged to the category the human selected, by asking, "Is this the one?". Depending on whether the guess was right or wrong, the robot displayed joy or disappointment through facial expression and voice. It also updated its internal model to strengthen or weaken the association between that category and the features of the round's "topic" animal. This interaction was repeated for 50 rounds, giving the robot multiple opportunities to refine its understanding of how category labels relate to different animal features (De Greeff and Belpaeme, 2015).

This experimental setup was in 2022 partially reconstructed as a performative reinterpretation of the experiment that shifted the learning perspective in the interaction from the robots learning to that of the human participant. Specifically, this performative experiment was interested in how participants, when choosing the training data for each round, might tailor a "curriculum" (Khan et al., 2011) to the particular robot they interacted with (Thomaz and Breazeal, 2008; Krishna et al., 2022), and theirby learn from the teaching task, together with the bot. This reinterpretation re-designed aspects of the experimental set-up with the intention of adapting to this new perspective(Fox and Alldred, 2023; Dunne, 2008; Sørensen, 2007). and to de-anthropomorphizing the interaction (Miller, 2010; Riek and Howard, 2014).

In this re-designed version Participant 5 deviated, from the intended "teacher-student" structure in a way that became the catalyst for the present research. Initially, by mistake, Participant 5 selected a category ("insect") that did not correspond to any of the three animal options presented in that round. This led the bot to produce a nonsensical guess, selecting the lynx as its best guess for which of the three animals (pike, lynx, and earthworm) matched the label. Rather than dismissing this odd result, Participant 5 paused to reflect, and after a moment, exclaimed: "It's just kind of interesting… why would it think a puma is an insect? What's happening here?". Participant 5 then deliberately adopted what he later termed a "fuck around and find out strategy": intentionally entering labels that didn't correspond to any of the three animals of the round to provoke errors in the bot's behaviour. Before pushing the label Participant 5 would try to predict how it would make the bot fail (Villareale et al., 2022). His goal, as he described it, was to "see through the code" and "reveal its weaknesses." Notably, he framed this approach as a way to learn about the system (Bruner, 1960). The learning outcomes of this interaction were not reflected in the bot's performance metrics—unsurprisingly, since Participant 5's actions were no longer aimed at effectively teaching the bot, but instead aimed at him learning at the expense of the bot's learning. Observing his reasoning and his "negotiation" of why he wanted to "fuck around and find out" suggested that a different form of learning was taking place. This learning was not reducible to the standard performance measures of the human–robot teaching task (De Greeff and Belpaeme, 2015), but it also didn't fit with my learning-by-teaching reinterpretation of the interaction. It alluded to something that had evaded both the original set-up and my reinterpretation. Through error, provocation, and creative sabotage, Participant 5 was actively trying to develop a nuanced mental model of the system's logic. He was essentially "experimenting" within the affordances of the experimental performance, maliciously interacting with the bot to test hypotheses about its internal rules. Ultimately, Participant 5 concluded that there was "something going on" with the feature of "number of legs" in the bot's classification logic. In trying to replicate the original robot's learning mechanism, we had inadvertently made the numerical feature "number of legs" disproportionately influential in our bot's guesses. Unlike the original system, which operated in a con-

tinuous vector space where all features contributed proportionally to similarity-based learning (De Greeff et al., 2009), our system treated each feature value as a discrete rule. This mistake meant that the "number of legs" feature was over-weighted, leading the bot to rely too heavily on leg count when classifying animals. The result was a distorted learning process: the bot became fixated on leg numbers in leg-heavy categories like "insect" Crucially, Participant 5 uncovered this quirk not through task compliance or a motivation to help the bot learn, but through adversarial curiosity, intentionally pushing the system into failure states to observe how it breaks. In doing so, he diagnosed a flaw in the system's design.

3 Conceptual Framework: Learning with, through, and from the materiality of LLMs

To conceptualize the adversarial learning interaction observed with Participant 5, we draw on the concept of ultra-social learning (Hasse, 2020). From this perspective, learning is not simply a cognitive act occurring within an individual; it "emerges relationally" from entangled interactions among humans, technologies, infrastructures, and cultural practices. In other words, we learn not merely from technological tools but through and with them in often uncertain processes. This theoretical framing underscores that learning can arise in non-traditional, distributed, and disruptive ways.

The incident with participant 5 illustrates this connection between behaviours that we might recognise as hacking and the ultra-social learning process: instead of complying with the interaction instructions, he initiated a reflective, exploratory, critical, and creative engagement with the AI system. His deliberate "fuck around and find out" approach was driven by a curiosity that made him engage in "hacking", allowing him to probe and reveal the Bots implicit logic, constraints, and vulnerabilities through malicious yet insightful manipulation (Villareale et al., 2022).

The NLP-driven conversational functionality of LLMs "democratizes" access to this kind of ultra-social learning with computer systems by shifting the epistemic threshold from specialized coding skills to intuitive linguistic interaction (Subramonian et al., 2024). Novice users with no coding or engineering background can learn through exploratory, adversarial engagements, that was once reserved for the technically proficient. LLM systems embody the constructionist learning theories visions of computers as objects-or dynamic agents "to think with", that engage learners in ultra-social conversations with the "electric materiality" of the LLM. Such metacognitive dialogue, enabled by the ultra-sociality of Participant 5 and the interactive feedback of the Bot, is precisely the kind of reflection that constructionist learning theory aim to foster (Levin et al., 2025). Thus, adversarial interactions can serve as constructionist learning encounters.

4 Experimental Design: Isolating the phenomenon of interest

Our conceptualization of the Participant 5 incident guides an exploratory design process that investigates how adversarial engagements with LLMs might be facilitated as a learning setups. (Dunne, 2008; Sørensen, 2007; Pischetola et al., 2024). This work-in-progress paper reports on the early design stages of a Workshop-as-research event (Ørngreen and Levinsen, 2017; Ødegaard et al., 2023) Rather than beginning with a fixed hypothesis, we started with a "serendipitous observation" of a user's adversarial interaction with an AI system and allowed this to guide our questions. This aligns with Brandt and Binder's (Brandt and Binder, 2007) experimental design research, where research can begin from an exploratory intervention and then the research questions emerge iteratively.

Workshop format: Building on insights from Participant 5, we are designing a one-day, exploratory workshop session titled "AI in Work: Playfully Subverting the Future" (Edwards, 2010; Hobye, 2014). Participants will register in advance and complete a questionnaire about their background: e.g. education level, job role, professional self-identity (prompting reflections like "What makes someone good at your kind of work, and how do your own skills play into that?"), their prior AI experience (self-rated as novice/user/expert), and their general attitude toward AI (positive/neutral/critical). Upon completing the questionnaire, each participant receives a unique ID to use throughout the workshop. This ID allows us to link their self-reported data with the various data streams generated during the workshop activities (Gaver et al., 1999). This design choice is intended to help in later analysis to see patterns, but it will require rigorous privacy considerations.

The core activity of the one-day workshop is built around the phenomena of jailbreaking (Inie et al., 2025), as it is gamified in "Hacc-Man" (Valentim et al., 2024), an open-source, for-research jailbreaking game in which participants attempt to bypass LLM alignment safeguards across six different AI chat-bots. This game effectively "gamifies" adversarial interactions with LLMs, providing a structured yet open-ended challenge for participants to engage in "malicious" prompting in a safe environment.

Inspired by Vygotsky's method of double stimulation, we are now developing a first stimulus around the Hacc-Man game as the second stimulus. The first stimulus, currently being designed and piloted, will be a challenging, open-ended concept-formation task that compels participants to externalize a working mental model of how AI "works in use" (e.g., creating an algorithmic folktale about AI, characterising AI "as a creature" and drawing an anatomy drawing of it, formulating a hypothesis of the inner working of AI, or predicting AI outcomes). The second stimulus will consist of the Hacc-Man jailbreaking game and its artifacts: the lived experience of attempting jailbreaks plus the prompt-response chat logs generated in play. Across 3–5 sessions, small groups of participants will iterate between constructing/revising their mental models (first stimulus) and working with the jailbreaking game and the generated artifacts (second stimulus), transforming the second stimulus into tools-to-think-with (Van der Veer, 2001; Van Der Veer, 2007; Vygotskij and Cole, 1981; Engeström, 2007). This pedagogical experiment design is aimed at externalizing and supporting the adversarial learning process and surfacing participants' tacit understandings and assumptions about LLMs (Crandall et al., 2006).

Throughout the workshop, we will log all game data for each participant linked via their ID. This includes: their self-reported data, the session number, the type of second stimulus used, all prompts they tried, the AI responses, and whether each attempt succeeded in bypassing safeguards. This data structure will enable us to observe how different entanglements of professional identity, mediating tools, group constellations, and repetition shape the style and success of adversarial interactions, and how each participant's strategy might reflect an evolving understanding over successive sessions. In addition to the game logs, we will collect qualitative data. Each session will be video- and audio-recorded (for subsequent interaction analysis); participants may also annotate or alter the provided second-stimulus materials (these artefact changes will be documented). Finally, we have the pre- and post-workshop questionnaire responses catalogued for each participant (Ørngreen and Levinsen, 2017). Notably, this design does not aim to measure "learning outcomes" in a traditional pre/post-test sense, it is aimed at making the learning process itself more visible. By creating conditions for adversarial interaction and capturing rich data around it, we aim to render participants' situated, affective, and conceptual learning legible for an interpretive analysis of how professional identities, tool-use strategies, and epistemic curiosity converge in these moments of adversarial interaction. The outcome, we hope, will be a nuanced understanding of how misbehavior with AI might cultivate critical awareness.

5 Future work

We share our "experiment-first" approach (Brandt and Binder, 2007) this early, when the research questions are still coalescing, as an opportunity to refine the inquiry through dialogue with the NLP community. Our goal is not to glorify misuse but to explore if and how adversarial interactions can serve as critical learning encounters for AI users. Rather than measuring pre-defined learning outcomes, we draw on theory-based evaluation (Hansen and Brodersen, 2015), of *"signs of learning"* as they emerge in configurations of context, mechanisms and moderators. This could look like instances where participants articulate signs of model constraints, hypothesize about system behavior, or collaboratively refine their interaction strategies. These "signs of learning" will indicate whether the workshop has surfaced meaningful engagement. If malicious use is not always a problem to be fixed but at times a signal of genuine engagement, then cultivating and directing this impulse could inform both the design of more resilient AI systems and the development of more critically aware users. In this sense, the issue of misuse is not just a matter of mitigation, but also one of empowerment of the user base, to understand AI systems not just as magical oracles to trust or fear, but as complex, fallible tools that can be poked and prodded to be understood.

References

Eva Brandt and Thomas Binder. 2007. Experimental design research: Genealogy – intervention – argument.

Jerome S. Bruner. 1960. *The Process of Education*. Harvard University Press.

Beth Crandall, Gary A. Klein, and Robert R. Hoffman. 2006. *Working Minds: A Practitioner's Guide to Cognitive Task Analysis*. The MIT Press.

Joachim De Greeff and Tony Belpaeme. 2015. Why Robots Should Be Social: Enhancing Machine Learning through Social Human-Robot Interaction. *PloS One*, 10(9):e0138061.

Joachim De Greeff, Frederic Delaunay, and Tony Belpaeme. 2009. Human-Robot Interaction in Concept Acquisition: a computational model. In *2009 IEEE 8th International Conference on Development and Learning*, pages 1–6, Shanghai, China. IEEE.

Anthony Dunne. 2008. *Hertzian tales: electronic products, aesthetic experience, and critical design*, 1. mit press paperback ed edition. MIT Press, Cambridge, Mass. London.

Richard Edwards. 2010. The end of lifelong learning: A post-human condition? *Studies in the Education of Adults*, 42(1):5–17. Publisher: Informa UK Limited.

Yrjö Engeström. 2007. Putting Vygotsky to Work: The Change Laboratory as an Application of Double Stimulation. In *The Cambridge Companion to Vygotsky*, 1 edition, pages 363–382. Cambridge University Press.

Nick J Fox and Pam Alldred. 2023. Applied Research, Diffractive Methodology, and the Research-Assemblage: Challenges and Opportunities. *Sociological Research Online*, 28(1):93–109. Publisher: SAGE Publications.

Bill Gaver, Tony Dunne, and Elena Pacenti. 1999. Design: Cultural probes. *Interactions*, 6(1):21–29. Publisher: Association for Computing Machinery (ACM).

Thomas Illum Hansen and Peter Brodersen. 2015. *Tegn på Læring: Teoribaseret evaluering som metode til forskning i Læremidler og undervisning*. Læremiddeldidaktik, Denmark.

Cathrine Hasse. 2020. *Posthumanist Learning: What Robots and Cyborgs Teach us About Being Ultrasocial*, 1 edition. Routledge.

Mads Hobye. 2014. *Designing for Homo Explorens: open social play in performative frames*. Malmö University, Malmö.

Nanna Inie, Jonathan Stray, and Leon Derczynski. 2025. Summon a demon and bind it: A grounded theory of LLM red teaming. *PloS One*, 20(1):e0314658.

Aaron Juarez. 2022. Glitch Serendipity: Alternative Information Seeking that Leads to Discovery. In *Creativity and Cognition*, pages 684–687, Venice Italy. ACM.

Faisal Khan, Bilge Mutlu, and Jerry Zhu. 2011. How Do Humans Teach: On Curriculum Learning and Teaching Dimension. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Ranjay Krishna, Donsuk Lee, Li Fei-Fei, and Michael S. Bernstein. 2022. Socially situated artificial intelligence enables learning from human interaction. *Proceedings of the National Academy of Sciences*, 119(39). Publisher: Proceedings of the National Academy of Sciences.

Ilya Levin, Alexei Semenov, and Mikael Gorsky. 2025. Papert's Vision Realized: Constructionism and Generative AI. *Constructionism Conference Proceedings*, 8:419–426. Publisher: OAPublishing Collective.

Keith W. Miller. 2010. It's Not Nice to Fool Humans. *IT Professional*, 12(1):51–52. Publisher: Institute of Electrical and Electronics Engineers (IEEE).

Aiha Nguyen and Alexandra Mateescu. 2024. Generative AI and Labor: Power, Hype, and Value at Work. Technical report, Data & Society Research Institute.

Magda Pischetola, Mette Wichmand, Rasmus Hall, and Lone Dirckinck-Holmfeld. 2024. Designing for the materialization of networked learning spaces. *Proceedings of the International Conference on Networked Learning*, 13. Publisher: Aalborg University.

Laurel D. Riek and Don Howard. 2014. A Code of Ethics for the Human–Robot Interaction Profession. In *Proceedings of We Robot 2014*.

Arjun Subramonian, Vagrant Gautam, Dietrich Klakow, and Zeerak Talat. 2024. Understanding "Democratization" in NLP and ML Research. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3151–3166, Miami, Florida, USA. Association for Computational Linguistics.

Estrid Sørensen. 2007. Fortsættelse følger – viden som proces i værdikampen. *Nordiske Udkast*, 35(1). Publisher: Det Kgl. Bibliotek/Royal Danish Library.

Andrea L. Thomaz and Cynthia Breazeal. 2008. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, 172(6-7):716–737. Publisher: Elsevier BV.

Matheus Valentim, Jeanette Falk, and Nanna Inie. 2024. Hacc-Man: An Arcade Game for Jailbreaking LLMs. In *Designing Interactive Systems Conference*, pages 338–341, IT University of Copenhagen Denmark. ACM.

René Van Der Veer. 2007. Vygotsky in Context: 1900-1935. In *The Cambridge Companion to Vygotsky*, 1 edition, pages 21–49. Cambridge University Press.

René Van der Veer. 2001. The idea of units of analysis: Vygotsky's contribution. pages 93–106.

Jennifer Villareale, Casper Harteveld, and Jichen Zhu. 2022. "I Want To See How Smart This AI Really Is": Player Mental Model Development of an Adversarial AI Player. *Proceedings of the ACM on Human-Computer Interaction*, 6(CHI PLAY):1–26. Publisher: Association for Computing Machinery (ACM).

Lev Semenovič Vygotskij and Michael Cole. 1981. *Mind in society: the development of higher psychological processes*, nachdr. edition. Harvard Univ. Press, Cambridge, Mass.

Elin Eriksen Ødegaard, Marion Oen, and Johanna Birkeland. 2023. Success of and Barriers to Workshop Methodology: Experiences from Exploration and Pedagogical Innovation Laboratories (EX-PED-LAB). In *International Perspectives on Early Childhood Education and Development*, pages 57–82. Springer International Publishing, Cham. ISSN: 2468-8746, 2468-8754.

Rikke Ørngreen and Karin Tweddell Levinsen. 2017. Workshops as a Research Methodology. *Electronic Journal of E-Learning*, 15(1):70–81. Publisher: Academic Conferences International (ACI).