

# Designing with Imperfection: An Educational Paradox Framework for LLM-driven NPCs in Serious Games

Giammarco Tosi

Aurel Vlaicu University of Arad, Romania

[gmtosi@gmail.com](mailto:gmtosi@gmail.com)

**Abstract:** Large language model (LLM)-driven non-player characters (NPCs) enable open dialogue, situated role-play, and adaptive support in serious games, but they can also fabricate information, lose state, respond slowly, violate game constraints, and create operational risks. In education, these imperfections can affect knowledge, cognitive load, trust, and assessment validity. This article reports a structured integrative review and conceptual synthesis of work published from 2020 through July 2026. Six searches were run in Crossref and OpenAlex. The 12 capped executions returned 1,200 query-level items; because pre-deduplication payloads were not retained, row-level audit begins with 993 unique records. The preserved dataset includes 1,191 record-query memberships, 993 rule-based decisions, 218 eligible records, and a named 24-source core corpus. The resulting Educational Paradox Framework distinguishes observed imperfection, pedagogical reframing, containment and transparency, learner appraisal, and learning or harm outcomes. It specifies five conditions under which an imperfection may be reframed as a learning mechanic and advances five testable propositions. Imperfection is educationally defensible only when it serves the objective, is legible and bounded, includes authoritative recovery, and is excluded from unsupported high-stakes judgment.

**Keywords:** Large language models, Non-player characters, Serious games, Game-based learning, Scaffolding, Calibrated trust, Hallucination

## 1. Introduction

LLMs allow NPCs to interpret free-form input, vary dialogue, and generate feedback at runtime. In serious games, these affordances support inquiry, role-play, language practice, and adaptive scaffolding. Chen and Hou (2024), for example, embedded ChatGPT-based scaffolding in employee ethics training, whereas Ngu et al. (2025) used multiple generative-AI scaffolds to develop workplace competencies. Other studies have examined elementary AI education through question-driven game support (Gong et al., 2026a) and digital-literacy training through an AI-driven NPC (Chernbumroong et al., 2026). Generative characters have also enabled interaction with a virtual Confucius (Luo et al., 2024), second-language practice (Zhao et al., 2024), and job-interview rehearsal (Mongkoljaturong et al., 2025). Further applications include role-based misinformation-awareness training (Tang et al., 2025) and educational role-playing games that provide immediate, situated feedback (Chien et al., 2025).

Educational use also raises the stakes of familiar game-AI failures. A fabricated statement from a fictional suspect may be material for inquiry; the same statement from a virtual tutor can become misinformation. Delay may create reflection in a turn-based task but undermine pronunciation or emergency practice. Context loss can appear as characterization while invalidating evidence gathering or feedback. Open-endedness is therefore an affordance, not evidence of educational effectiveness. Studies report learning benefits from some LLM-based scaffolds, but also inaccuracies, latency, repetition, prompting difficulty, and concerns about transparency and quality (Chen & Hou, 2024; Chernbumroong et al., 2026; Goldshtein et al., 2026). Players have also preferred human-authored dialogue where it provided stronger grounding and logical flow (Akoury et al., 2023).

Game research has begun to treat hallucination as possible design material through narrative boundaries or deliberately irrational characters (Vi et al., 2025). Educational use requires a stricter test: fictional coherence does not show that a mechanic supports learning or prevents misconception. Productive failure is a structured learner attempt followed by consolidation, not exposure to uncontrolled system error (Kapur, 2008). The relevant question is therefore which imperfection, for which objective and learner, under which safeguards, and with what evidence of benefit or harm.

This article asks:

*RQ1. Which LLM-related limitations recur in NPC and educational-game research, and how can they affect learning?*

*RQ2. Under what conditions can a limitation be reframed as a useful learning mechanic?*

*RQ3. What design and evaluation requirements would make such a framework testable?*

The contribution is the Educational Paradox Framework (EPF), a conceptual design and evaluation framework that links recurrent imperfections to learning mechanisms, safeguards, and outcomes. Its intended scope is low- to moderate-stakes formative learning in secondary, higher, and professional education, especially dialogue-rich inquiry and role-play. Unsupervised diagnosis, certification, consequential grading, and other high-stakes uses fall outside its safe default scope.

## **2. Educational Foundations**

Game-based learning depends on alignment among activity, instructional support, and outcomes, not on game elements alone (Plass et al., 2015; Wouters et al., 2013). Generative NPCs can act as simulated stakeholders, coaches, feedback sources, or objects of critique. Yet effective feedback must clarify the goal, current progress, and next action (Hattie & Timperley, 2007). Fluent but ungrounded dialogue may satisfy none of these functions.

LLMs can produce plausible output that is unfaithful to external facts or supplied context (Ji et al., 2023). Such content errors can create misconceptions, while contradictions, state loss, and poorly timed feedback create interaction errors. Both may impose extraneous cognitive load: effort caused by system behavior rather than the target knowledge (Sweller, 1988). The risk is greatest for novices who cannot independently detect fluent errors.

Trust is consequently both a mechanism and an outcome. Appropriate reliance requires perceived trust to track actual capability (Lee & See, 2004). Overreliance can transmit false feedback; disuse can neutralize a reliable scaffold. A label stating that content is AI-generated does not by itself create calibrated trust. Learners need an intelligible NPC role, accessible evidence, opportunities to challenge output, and authoritative resolution. Playful settings may offer low-consequence practice in those judgments, provided human expertise remains central (Huber et al., 2024).

## **3. Method**

### **3.1 Review Design and Search**

The study used a structured integrative review with concept-directed synthesis because the emerging literature combines educational evaluations, prototypes, system studies, reviews, and conceptual work. It does not claim systematic exhaustiveness or meta-analysis.

On July 30, 2026, Crossref and OpenAlex were searched for records dated January 1, 2020, to July 30, 2026. The exact strings were: “large language model” “educational game”; “large language model” “serious game”; “large language model” NPC education game; “large language model” role-playing learning; “generative AI” “educational game” character; and “LLM” NPC learning game. Each of the 12 query-index executions requested the first 100 results in the API’s default order. The saved count log therefore totals 1,200 returned items, but the individual pre-deduplication rows were not retained.

The auditable dataset begins with 993 records deduplicated by DOI and then normalized title. Aggregated query fields preserve 1,191 unique record-query memberships. A rule-based title-and-abstract screen retained records containing (a) an LLM or generative-AI term, (b) a game, simulation, role-play, metaverse, or virtual-patient term, and (c) either an education/learning or NPC/dialogue term. All 993 decisions are preserved: 218 records met the rule and 775 did not.

ACM Digital Library, IEEE Xplore, SpringerLink, ScienceDirect, DiGRA Digital Library, and ACL Anthology were consulted to verify versions and locate game-specific proceedings. These supplementary routes were not logged comprehensively. The 24-source core corpus comprises 18 records that passed automated screening, Vi et al. (2025), which was in the 993-record set but missed the lexical rule, and five supplementary sources (Chelli & Batista, 2025; Gallotta et al., 2024; Wang et al., 2025; Wevelsiep et al., 2025; Xu et al., 2024). Aster et al. (2025) provides adjacent non-LLM evidence; seven purposive sources support the educational theory. No unsupported intermediate count of manually examined records is reported.

Table 1: Auditable review trail

| Stage             | Result                             | Evidence and decision boundary                                                          |
|-------------------|------------------------------------|-----------------------------------------------------------------------------------------|
| API retrieval     | 1,200 query-level items            | Twelve saved counts; individual payload rows unavailable                                |
| Deduplication     | 993 unique records                 | DOI, then normalized title; full roster retained                                        |
| Query aggregation | 1,191 memberships                  | Unique record-query associations after deduplication                                    |
| Automated screen  | 218 eligible; 775 excluded         | Complete 993-decision rule-based dataset                                                |
| Core corpus       | 24 named sources                   | 18 eligible, one from the 993 set, five supplementary; no complete manual exclusion log |
| Interpretive set  | One adjacent; seven theory sources | Identified separately from the core corpus                                              |

### 3.2 Synthesis and Data Boundary

Sources were included when they were English-language, peer-reviewed, and directly relevant to LLM integration in games, LLM-driven NPC dialogue, or conversational scaffolding in serious games. Generic classroom chatbots, nonlanguage agents, marketing claims, anecdotes, and superseded preprints were excluded from the core corpus.

The educational evidence covered contextual scaffolding for employee-ethics training (Chen & Hou, 2024), generative role-play with immediate feedback (Chien et al., 2025), and multi-scaffold support for workplace competencies (Ngu et al., 2025). Studies of AI and digital literacy evaluated question-driven scaffolding, AI-driven NPCs, contextual adaptation, and game-based chatbots (Chernbumroong et al., 2026; Gong et al., 2026a; Wang et al., 2025; Xu et al., 2024). Other contributions examined stakeholder requirements and learner-plan generation (Goldshtein et al., 2026; Goslen et al., 2025), historical or cultural interaction and second-language learning (Luo et al., 2024; Zhao et al., 2024), and misinformation-awareness play (Tang et al., 2025).

Prototype studies demonstrated LLM hacking as puzzle material, metaverse-based language learning, and tangible fact-checking, but were treated as feasibility evidence rather than proof of learning effectiveness (Gallotta et al., 2025; Plupattanakit et al., 2024; Tang & Singha, 2024). Game and NPC studies informed risks involving player judgments and dialogue grounding (Akoury et al., 2023), immersion and emotional responses (Chelli & Batista, 2025; Marincioni et al., 2024), multimodal interview rehearsal (Mongkoljaturong et al., 2025), and voice-driven dialogue architectures (Wevelsiep et al., 2025). Finally, reviews and conceptual studies mapped the wider design space, its educational implications, and proposals for treating hallucination as a game feature (Gallotta et al., 2024; Gong et al., 2026b; Huber et al., 2024; Vi et al., 2025).

The synthesis consolidated recurrent problems, mapped them to educational mechanisms and outcomes, evaluated candidate reframings against necessary safeguards, and derived propositions and measures. The retained audit trail documents aggregate counts for the 12 query–index combinations, 993 deduplicated records, 1,191 record–query memberships, 993 automated screening decisions, the 218-record eligible set, the 24-source core corpus, and artifact hashes. It does not include—and cannot reconstruct—the original pre-deduplication API payloads, repository paths, manual exclusion reasons, or a source-level coding matrix.

## 4. Findings and the Educational Paradox Framework

### 4.1 Recurrent Imperfections

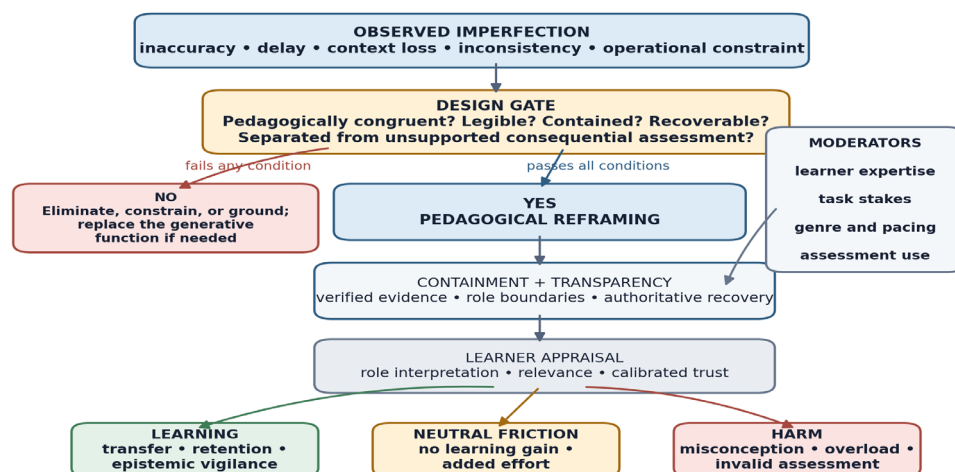
Five limitation families answer RQ1. Inaccuracy includes factual error, unfaithfulness to supplied evidence, fabricated sources, and claims about unavailable game actions. Delay arises across inference, retrieval, speech, animation, and orchestration; its effect depends on whether timing is part of the task. Context loss includes forgetting earlier evidence, goals, or learner attempts. Narrative or mechanical inconsistency occurs when dialogue contradicts roles, rules, resources, or executable state. Operational dependence covers cost, privacy, accessibility, network failure, and model-version drift.

The literature shows why these categories cannot be judged through engagement alone. Guarded NPC designs have produced learning gains while still reporting inaccuracies and latency (Chernbumroong et al., 2026). Open scaffolding can improve learning without outperforming conventional support on flow or motivation and can shift effort to question formulation (Chen & Hou, 2024). Hybrid architectures restrict generation to preserve coherence (Wevelsiep et al., 2025), while multimodal systems introduce additional delay and integration points (Mongkoljaturong et al., 2025). Generative freedom may support emotional or social interaction (Marincioni et al., 2024; Chelli & Batista, 2025), but instructional validity requires stable facts, state, and assistance. Operational failures rarely offer a learning mechanism and should normally be treated as deployment gates.

### 4.2 Framework Constructs and Design Gate

The EPF is conceptual, not a validated predictive model. It contains five constructs: observed imperfection, a reproducible departure from intended behavior; pedagogical reframing, a deliberate mapping to an activity such as evidence evaluation or reflection; containment and transparency, controls that bound uncertainty and enable recovery; learner appraisal, interpretation of the NPC's role and reliability; and learning or harm outcome, including achievement, transfer, trust, load, misconception, inequity, or invalid assessment. Task stakes, learner expertise, genre, and assessment use moderate the pathway.

The causal logic is deliberately indirect. Imperfection alone predicts no educational benefit. Reframing selects the learner activity; containment governs exposure to uncertainty; learner appraisal shapes reliance or challenge; and measured outcomes determine whether the design produced learning, neutral friction, or harm. A fluent falsehood may be narratively coherent yet educationally defective, while an explicitly contestable statement from a biased witness may activate source comparison. Reframing is therefore more than explaining an error through character lore: it must change what learners are asked to notice, do, and justify. The framework is also asymmetric. Content, state, or pacing failures enter the paradox branch only after justification, whereas privacy, accessibility, and infrastructure failures remain deployment constraints. This places the burden of evidence on the creative reframe rather than on learners who would otherwise absorb its risk.



**Figure 1: Educational Paradox Framework.** An observed imperfection enters a design gate. Unsafe or irrelevant imperfections are eliminated or contained; a candidate reframe must specify a learning mechanism, safeguards, learner appraisal, and measurable outcomes

A candidate reframe must satisfy five conditions. Congruence: handling the imperfection directly serves a stated learning objective. Legibility: learners understand the NPC’s epistemic role and when verification is expected. Containment: authoritative facts, game state, and prohibited outputs are bounded. Recoverability: correction, comparison, or debriefing occurs before error stabilizes. Assessment separation: stochastic dialogue is not unexamined ground truth or the sole basis for consequential scoring. Failure of any condition sends the design to elimination or containment, especially for novices and high-stakes domains.

**Table 2: Limitation-to-mechanism map**

| Imperfection and default risk                           | Possible reframe and mechanism                                              | Minimum safeguards                                                          | Validation                                                                              |
|---------------------------------------------------------|-----------------------------------------------------------------------------|-----------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|
| <b>Inaccuracy: misconception or unsafe feedback</b>     | Contestable witness or synthetic claim; source evaluation and argumentation | Disclosed role, verified evidence, bounded topic, correction and debrief    | Fact-check accuracy, justification, transfer, residual misconception, trust calibration |
| <b>Delay: broken flow or timing unfairness</b>          | Signaled deliberation interval; prediction or retrieval                     | Bounded wait, relevant task during delay; exclude speed-critical practice   | Reflection, retention, load, abandonment, pacing                                        |
| <b>Context loss: contradiction or repeated feedback</b> | Evidence reconstruction; summarization and monitoring                       | Inspectable authoritative log; NPC cannot overwrite state; final correction | Recall, repair strategy, frustration, state consistency                                 |
| <b>Perspective or mechanical inconsistency</b>          | Biased role or multiple perspectives; comparison and bias recognition       | Role boundaries, fact-perspective distinction, cross-source check, debrief  | Causal explanation, perspective differentiation, misconception                          |
| <b>Cost, privacy, access, network failure</b>           | Normally no educational reframe                                             | Budget, privacy, accessibility, fallback, and uptime gates                  | Cost/session, completion parity, defects, failures, compliance                          |

Table 2 answers RQ2: an imperfection has no inherent pedagogical value. It becomes a candidate mechanic only when the learning activity makes its detection or management meaningful and the safeguards make error inspectable and recoverable.

## 5. Contrastive Cases and Propositions

Four educational cases locate existing work along the EPF continuum. They are illustrations from the corpus, not proof that the framework is validated.

**Table 3: Contrastive educational cases**

| Case                                                             | Evidence                                                                                                                         | EPF interpretation                                                                                                        |
|------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------|
| <b>Chen and Hou (2024), employee ethics, n=61</b>                | Both ChatGPT-NPC and document conditions improved; no clear advantage in flow, motivation, or anxiety; some prompting difficulty | Open dialogue is not automatically superior; measure prompting burden and scaffold quality                                |
| <b>Chernbumroong et al. (2026), digital literacy, n=60</b>       | Higher credibility-evaluation gains than traditional instruction; latency and occasional inaccuracies remained                   | Containment aligns role prompts, filters, context checks, and regeneration with the objective; accuracy still needs audit |
| <b>Tang et al. (2025), misinformation awareness, n=47</b>        | Mixed-method within-subject evaluation reported gains in identification and awareness                                            | Variable persuasive output becomes an object of critique rather than an authoritative answer                              |
| <b>Gallotta et al. (2025), prompt-security serious-game demo</b> | LLM vulnerabilities form the puzzle material; no outcome trial                                                                   | Strong congruence, but feasibility does not establish learning or safety                                                  |

Other work demonstrates adaptive plans, multiple scaffolds, AI chatbots, metaverse language practice, fact-checking, and cultural interaction (Goslen et al., 2025; Plupattanakit et al., 2024; Tang & Singha, 2024; Xu et al., 2024). These prototypes and evaluations expand the design space, but feasibility or learner interest cannot substitute for factual accuracy, transfer, delayed retention, and misconception checks.

The synthesis yields five propositions:

P1 - Congruence and legibility. A clearly framed imperfection that directly serves the objective will improve learning and trust calibration relative to the same undisclosed imperfection; without congruence or legibility it will create friction or harm.

P2 - Containment. Grounding, rules, authoritative logs, and recovery feedback will preserve more benefit from open interaction while reducing error, state drift, and residual misconception.

P3 - Delay. A bounded delay will aid reflection or retrieval only when signaled, paired with a relevant cognitive task, and placed where deliberation advances the objective.

P4 - Contradiction. Inconsistency will support epistemic vigilance or perspective taking only when learners have sufficient knowledge or evidence and receive an authoritative debrief; novices will otherwise be more vulnerable.

P5 - Trust and assessment. Hidden, uncorrected error and unlogged variation will reduce calibrated trust and assessment validity; separating formative dialogue from consequential scoring and retaining auditable logs will reduce this effect.

## **6. Design and Evaluation Requirements**

RQ3 requires the framework to be operational rather than rhetorical. Designers should first specify the learner, prior knowledge, objective, transfer target, and stakes, then assign the NPC an explicit epistemic role: authority, coach, stakeholder, peer, adversary, witness, or object of critique. Before release, they should reproduce failure modes and report accuracy by task, latency distributions, state consistency, model and prompt versions, and moderation failures. Each imperfection must then be eliminated, contained, or justified against the five gate conditions. Authoritative recovery can use evidence cards, source excerpts, executable state checks, scripted checkpoints, teacher oversight, or debriefing.

Evaluation should compare meaningful alternatives rather than merely LLM versus no LLM. A strong initial study would include scripted, unrestricted-LLM, grounded-LLM, and EPF conditions using equivalent curricular content. Primary outcomes should include achievement, delayed retention, transfer, residual misconceptions, verification behavior, calibrated trust, cognitive load, and abandonment. System outcomes should include faithfulness, state consistency, latency, refusals, regeneration, cost per completed session, accessibility, and failure recovery. Interaction logs permit independent error coding and reveal unequal assistance. Unvalidated model output should never be the sole scorer or source of truth in consequential assessment.

## **7. Discussion**

The central result is a role-task distinction. The same generated statement can be defective tutor feedback, a credible but biased character perspective, or evidence in a fact-checking task. The EPF therefore narrows hallucination-as-feature arguments (Vi et al., 2025): narrative coherence is insufficient when a coherent fiction can teach a falsehood. Educational legitimacy depends on an explicit learning mechanism, technical and instructional containment, and measurement of benefit and harm.

The framework also separates learner productive failure from system failure. Learner struggle becomes productive through comparison and consolidation; uncontrolled model error has no such entitlement. A system imperfection can enter the learning design only when learners are meant to inspect it and receive authoritative recovery. This makes calibrated trust a teachable competence: serious games can provide repeated, low-consequence decisions about when to accept, verify, or reject AI advice.

For practice, designers should use the smallest generative surface that serves the objective. Deterministic systems should retain canonical facts, safety rules, rewards, and executable state, while LLMs vary expression, interpret free language, or simulate bounded perspectives. A character who may be wrong must be visibly distinct from a system that is silently wrong. Debriefing is part of the mechanic, not an optional add-on. Privacy, accessibility, cost, and fallback are validity conditions because unreliable or exclusionary delivery changes who can complete the learning task.

## 8. Limitations and Future Research

The review is structured but not exhaustive. Searches were English-language, each index-query pair was capped at 100 default-ranked results, and the field changes rapidly. The 1,200 pre-deduplication rows were not retained; row-level audit begins with 993 unique records. The route from 218 automated-eligible records to the 24-source corpus is only partly auditable: five sources came from supplementary checking without preserved path-level logs, and no complete manual exclusion log or coding matrix exists. The heterogeneous corpus combines outcome studies, prototypes, and conceptual work, preventing meta-analysis and requiring cautious inference. Finally, the EPF has not been prospectively validated.

Research should manipulate one imperfection and mechanism at a time. An initial experiment could compare unrestricted, grounded, and grounded-but-deliberately-contestable NPCs in a media-literacy game, measuring transfer, misconceptions, verification, trust, and delayed retention. A separate study could manipulate a signaled reflective interval independently of technical latency. Further work should test learner expertise, domain stakes, teacher orchestration, game-state faithfulness, and equalization of dynamic assistance as moderators.

## 9. Conclusion

LLM-driven NPCs expand the conversational and adaptive possibilities of serious games, but fluency, responsiveness, and novelty are not educational outcomes. The EPF reframes model imperfection as a conditional pedagogical resource rather than a defect to tolerate or a feature to celebrate. Fallibility should be deliberately designed only when it advances a specified learning objective, remains legible and bounded, can be resolved through authoritative evidence or debriefing, and is excluded from unsupported high-stakes judgment; otherwise, it should be designed out or tightly constrained. By translating this principle into constructs, boundary conditions, propositions, cases, and an empirical evaluation design, the framework provides a testable basis for distinguishing productive fallibility from avoidable risk. Its central claim is simple: an educational NPC need not always be infallible, but its fallibility must never be accidental, opaque, unresolved, or consequential beyond the learning task.

## References

- Akoury, N., Yang, Q., & Iyyer, M. (2023). A framework for exploring player perceptions of LLM-generated dialogue in commercial video games. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 2295-2311). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-emnlp.151>
- Aster, A., Lotz, A., & Raupach, T. (2025). Theoretical background of the game design element “chatbot” in serious games for medical education. *Advances in Simulation*, 10, Article 10. <https://doi.org/10.1186/s41077-025-00341-7>
- Chelli, V. C., & Batista, A. L. F. (2025). Sophia Project: Enhancing player immersion with intelligent autonomous NPCs in 2D RPGs. In *Proceedings of the First Workshop on Artificial Intelligence for Human-Game Interaction (AI4HGI 2025)* (CEUR Workshop Proceedings, Vol. 4097). <https://ceur-ws.org/Vol-4097/paper1.pdf>
- Chen, Y.-C., & Hou, H.-T. (2024). A mobile contextualized educational game framework with ChatGPT interactive scaffolding for employee ethics training. *Journal of Educational Computing Research*, 62(7), 1737-1762. <https://doi.org/10.1177/07356331241268505>
- Chien, C.-C., Chan, H.-Y., & Hou, H.-T. (2025). Learning by playing with generative AI: Design and evaluation of a role-playing educational game with generative AI as scaffolding for instant feedback interaction. *Journal of Research on Technology in Education*, 57(4), 894-913. <https://doi.org/10.1080/15391523.2024.2338085>
- Chernbumroong, S., Intawong, K., Asawimalkit, U., Puritat, K., & Julrode, P. (2026). Design and evaluation of a generative AI-enhanced serious game for digital literacy: An AI-driven NPC approach. *Informatics*, 13(1), Article 16. <https://doi.org/10.3390/informatics13010016>
- Gallotta, R., Liapis, A., & Yannakakis, G. N. (2025). Prompt Override: LLM hacking as serious game. In *2025 IEEE Conference on Games (CoG)*. IEEE. <https://doi.org/10.1109/CoG64752.2025.11114120>
- Gallotta, R., Todd, G., Zammit, M., Earle, S., Liapis, A., Togelius, J., & Yannakakis, G. N. (2024). Large language models and games: A survey and roadmap. *IEEE Transactions on Games*. Advance online publication. <https://doi.org/10.1109/TG.2024.3461510>
- Goldshstein, M., Ahmed, I., Yu, F., Verma, V., McNamara, D. S., & Arner, T. (2026). Characterizing user needs for GenAI incorporation in educational games. *Education Sciences*, 16(2), Article 300. <https://doi.org/10.3390/educsci16020300>
- Gong, Y., Wang, M., He, L., Xu, C., & Yu, Y. (2026a). Asking, playing, learning: Investigating large language model-based scaffolding in digital game-based learning for elementary artificial intelligence education. *Journal of Educational Computing Research*, 64(2), 311-343. <https://doi.org/10.1177/07356331251396354>
- Gong, Y., Wang, M., Tu, Y.-F., Huang, C., & Zhang, D. (2026b). Beyond pre-scripted interactions: Mapping the integration of LLMs in digital game-based learning - A scoping review. *Entertainment Computing*, 56, Article 101082. <https://doi.org/10.1016/j.entcom.2026.101082>

- Goslen, A., Kim, Y. J., Rowe, J., & Lester, J. (2025). LLM-based student plan generation for adaptive scaffolding in game-based learning environments. *International Journal of Artificial Intelligence in Education*, 35(2), 533-558. <https://doi.org/10.1007/s40593-024-00421-1>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81-112. <https://doi.org/10.3102/003465430298487>
- Huber, S. E., Kiili, K., Nebel, S., Ryan, R. M., Sailer, M., & Ninaus, M. (2024). Leveraging the potential of large language models in education through playful and game-based learning. *Educational Psychology Review*, 36, Article 25. <https://doi.org/10.1007/s10648-024-09868-z>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article 248. <https://doi.org/10.1145/3571730>
- Kapur, M. (2008). Productive failure. *Cognition and Instruction*, 26(3), 379-424. <https://doi.org/10.1080/07370000802212669>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50-80. [https://doi.org/10.1518/hfes.46.1.50\\_30392](https://doi.org/10.1518/hfes.46.1.50_30392)
- Luo, H., Gao, F., Fang, K., Liu, D., Lin, Z., & Chan, W. K. (2024). Study with Confucius: An AI-based immersive educational game with multiple educational modes. In *SIGGRAPH Asia 2024 Educator's Forum*. ACM. <https://doi.org/10.1145/3680533.3697066>
- Marincioni, A., Miltiadous, M., Zacharia, K., Heemskerk, R., Doukeris, G., Preuss, M., & Barbero, G. (2024). The effect of LLM-based NPC emotional states on player emotions: An analysis of interactive game play. In *2024 IEEE Conference on Games (CoG)*. IEEE. <https://doi.org/10.1109/CoG60054.2024.10645631>
- Mongkoljaturong, K., Manitsakulwong, M., & Moodleah, S. (2025). AI-powered MetaHuman interviewer: Serious game for student job interview skills. *IEEE Access*, 13, 205505-205520. <https://doi.org/10.1109/ACCESS.2025.3640017>
- Ngu, P.-C., Chien, C.-C., Ho, Y.-T., & Hou, H.-T. (2025). A generative AI educational game framework with multi-scaffolding supports workplace competency development. *Computers & Education*, 239, Article 105421. <https://doi.org/10.1016/j.compedu.2025.105421>
- Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist*, 50(4), 258-283. <https://doi.org/10.1080/00461520.2015.1122533>
- Plupattanakit, K., Suntiichaiikul, P., Taveekitworachai, P., Thawonmas, R., White, J., Sookhanaphibarn, K., & Choensawat, W. (2024). LLMs in Eduverse: LLM-integrated English educational game in metaverse. In *2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)*. IEEE. <https://doi.org/10.1109/GCCE62371.2024.10760474>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257-285. [https://doi.org/10.1207/s15516709cog1202\\_4](https://doi.org/10.1207/s15516709cog1202_4)
- Tang, H., & Singha, M. F. (2024). A Mystery for You: A fact-checking game enhanced by large language models and a tangible interface. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM. <https://doi.org/10.1145/3613905.3648110>
- Tang, H., Sun, S., Nie, K., Li, A., Sergeeva, A., & Lc, R. (2025). Breaking the News: Taking the roles of influencer vs. journalist in an LLM-based game for raising misinformation awareness. *Proceedings of the ACM on Human-Computer Interaction*, 9(6), Article GAMES005, 131-171. <https://doi.org/10.1145/3748600>
- Vi, K., Chen, M., Sun, Y., Ming, Y., & Wang, F. (2025). Reconceptualizing LLM-induced hallucinations as game features. In *Proceedings of DiGRA 2025*. DiGRA Digital Library. <https://doi.org/10.26503/dl.v2025i2.2426>
- Wang, M., Zhang, D., Zhu, J., & Gu, H. (2025). Effects of incorporating a large language model-based adaptive mechanism into contextual games on students' academic performance, flow experience, cognitive load and behavioral patterns. *Journal of Educational Computing Research*, 63(3), 662-694. <https://doi.org/10.1177/07356331251321719>
- Wevelsiep, M., Walker, N. T., Wagner, N., & Ultes, S. (2025). A voice-controlled dialogue system for NPC interaction using large language models. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology* (pp. 29-38). Association for Computational Linguistics. <https://aclanthology.org/2025.iwds-1.4/>
- Wouters, P., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology*, 105(2), 249-265. <https://doi.org/10.1037/a0031311>
- Xu, Y., Zhu, J., Wang, M., Qian, F., Yang, Y., & Zhang, J. (2024). The impact of a digital game-based AI chatbot on students' academic performance, higher-order thinking, and behavioral patterns in an information technology curriculum. *Applied Sciences*, 14(15), Article 6418. <https://doi.org/10.3390/app14156418>
- Zhao, Y., Pan, J., Dong, Y., Dong, T., Wang, G., Ying, F., Shen, Q., & Cao, J. (2024). Language Urban Odyssey: A serious game for enhancing second language acquisition through large language models. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM. <https://doi.org/10.1145/3613905.3651112>