# LLM-based NPCs in Video Games: A Systematic Mapping Study

Dilan Mian, Jeremy Bradbury, Yann-Gael Guéhéneuc, Fabio Petrillo, and Cristiano Politowski

***Abstract*—For decades, non-player characters (NPCs) have been driven by rule-based systems whose personalities and interactions are scripted before release. Large language models (LLMs) promise NPCs that converse openly, adapt to new situations, and behave in ways developers did not fully author, prompting a wave of research and commercial experimentation. Yet the field lacks a unified picture of what has been built, how well it works, and at what cost. We address this gap with a systematic mapping study of 377 papers from the academic literature, yielding 42 papers analysed against a main research question and ten sub-questions across four thematic axes. We provide the first quantitative characterisation of component adoption, latency, safety coverage, and reproducibility: short-term memory and world-state grounding are near-universal (60% and 69%), whereas long-term memory (31%), reflection (38%), and retrieval-augmented generation (14%) remain rare, and response times span 2.6 to 15 seconds. Finally, we identify six structural barriers to commercial adoption and offer evidence-based recommendations for developers and researchers.**

***Index Terms*—non-player character, large language models, generative models, video game, mapping study**

## I. Introduction

Non-player characters (NPCs) shape player experience by delivering narrative, providing challenge, populating worlds, and mediating agency. Traditionally, explicitly scripted systems control their behaviour (dialogue trees, behaviour trees, and goal-oriented action planners), which brings consistency but is ultimately limited by what designers author in advance [1].

Large Language models (LLMs) introduce the idea of NPCs that converse openly, adapt to new situations, and act in ways that are not pre-programmed. The growing accessibility of LLMs has led researchers and the industry to experiment with LLM-based NPCs, from video game mods[2] like Skyrim to commercial platforms like InWorld AI.

However, this subject lacks a unified understanding of what has been built, how well it works, and at what cost. Existing studies use incompatible evaluation methods, customised environments, and undisclosed prompts, which hinder cross-comparison. No shared benchmarks for believability or dialogue coherence exist. Also, key issues such as latency, hallucination, and long-term consistency are rarely quantified, and few papers implement runtime guardrails.

D. Mian, J. Bradbury, and C. Politowski are with Ontario Tech University, Oshawa, Canada (e-mail: dilan.mian@ontariotechu.ca; jeremy.bradbury@ontariotechu.ca; cristiano.politowski@ontariotechu.ca). Corresponding author: D. Mian.

Y.-G. Guéhéneuc is with Concordia University, Montreal, Canada, (e-mail: yann-gael.gueheneuc@concordia.ca).

F. Petrillo is with École de Technologie Supérieure, Montreal, Canada, (e-mail: fabio.petrillo@etsmtl.ca).

We fill this gap by systematically mapping LLM-based NPC dialogue and behaviour across 377 records from ACM Digital Library, IEEE Xplore, and Scopus, yielding 42 papers for full analysis after applying inclusion, exclusion, and quality criteria. The review is structured around a main research question, *"what LLM-based approaches exist for creating lifelike NPC dialogue and behaviour, and how are they evaluated?"*. To answer this, we derived *ten sub-questions* across four thematic axes with 36 extracted variables.

This review makes three primary contributions. First, we present the first quantitative characterisation of component adoption, latency, safety coverage, and reproducibility across the corpus, showing that short-term memory and world-state grounding are near-universal (60% and 69%, respectively) while long-term memory (31%), reflection (38%), and RAG (14%) remain minority practices, and that end-to-end response times range from 2.6 s to 15 seconds. Second, we identify and explain six structural barriers to commercial adoption (latency, quality-assurance breakdown, cost at player scale, multiplayer synchronisation, content-rating compliance, and localisation) that existing literature mentions in isolation, if at all. Finally, we offer evidence-based recommendations specifying which architecture patterns show the strongest signs of practical viability, what a minimum evaluation reporting standard should contain, and what the field's highest-priority open problems are.

The remainder of this paper is structured as follows. Section II describes the systematic methodology. Section III presents the results across all ten research questions. Section IV discusses the implications of the findings for developers and researchers. Section V concludes the paper.

## II. Methodology

We crafted a systematic mapping study following the guidelines of Sweetser [3] to identify, evaluate, and synthesize research on LLM-based approaches to NPC behaviour and dialogue in interactive environments. We utilized the Prisma diagram in Figure 1 to show the full screening flow of the selected corpus. The process comprises the following five phases.

### A. Mapping Protocol and PICOC Framework

We grounded the mapping scope in a PICOC (Population, Intervention, Comparison, Outcome, Context) specification, which defines the boundaries of the primary studies' inclusion decisions [3]. Table I presents the five elements.

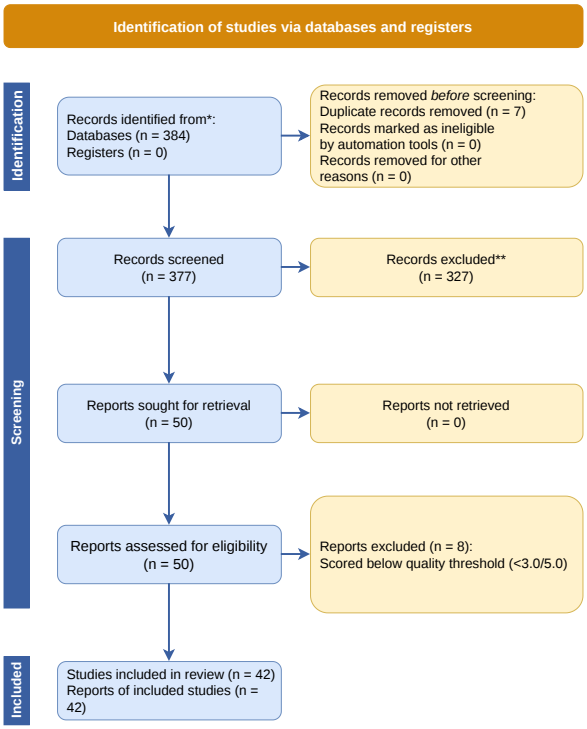



Fig. 1. PRISMA flow diagram illustrating the systematic identification, screening, eligibility assessment, and final inclusion of studies across the review process.

TABLE I
PICOC FRAMEWORK FOR THIS SYSTEMATIC MAPPING STUDY.

| Element | Definition |
|---|---|
| Population | Interactive LLM-based NPCs in games or virtual environments |
| Intervention | LLM-based tools and architectures for generating realistic dialogue and autonomous behaviour, including systems with memory, retrieval grounding, planning, and tool use |
| Comparison | Scripted NPC AI (behaviour trees, FSMs, GOAP, utility AI), non-LLM dialogue systems, or ablated LLM variants without memory, RAG, or planning components |
| Outcome | Believability and human-likeness of NPC dialogue and behaviour; dialogue coherence and consistency; behaviour quality and agency; engineering feasibility (latency and cost); safety; and designer controllability |
| Context | Interactive games and virtual worlds with real or simulated player-agent interaction |

We defined *LLM-based NPCs* broadly to include autonomous agents in simulated virtual environments (e.g., Minecraft-based research agents) where the agent fulfils a role similar to an NPC, in which they respond to player or user input, inhabiting a persistent world, and exhibiting goal-directed behaviour, even when the paper does not use the term "NPC". In these contexts, the end user is typically a researcher or an automated training process rather than a human player, but the underlying architectures still work for player-facing NPC design. This reflects the convergence between game LLM-based NPC research and autonomous agent research and captures architectures (e.g., Generative Agents [4], DEPS [5]) directly applicable to NPC design despite originating outside it.

### B. Research Questions

We organised the mapping around one main research question and ten sub-questions that branch from it along four thematic branches: architecture and design, scope and evaluation, engineering and control, and quality and transparency.

The main research question of *"what LLM-based approaches exist for creating lifelike NPC dialogue and behaviour, and how are they evaluated?"* is the baseline question that this review answers in Section IV. It branches into ten sub-questions grouped along four thematic axes:

1) **Architecture & Design**

RQ1 What LLM-based tools, architectures, and design techniques have been proposed to enable lifelike NPC dialogue and behaviour in interactive environments?

RQ2 Which system components are used to ground and control NPC outputs (e.g., memory, RAG, world-state grounding, planning, tool/function calling), and how are they combined?

1. **Scope & Evaluation**

RQ3 What kinds of NPC capabilities are targeted (dialogue-only, behaviour-only, dialogue + action), and what interactions are supported (text, voice, embodied action)?

RQ4 In what contexts are these NPCs deployed (game genres, virtual worlds, interVR/simulators), and what assumptions do authors make about the environment and world state access?

RQ5 What baselines and comparisons are used (scripted dialogue trees, behavior trees/FSM/GOAP/utility AI, non-LLM systems, or ablations), and what improvements are reported?

2) **Engineering & Control**

RQ6 What trade-offs are reported between realism and engineering constraints (latency, cost, reliability), and what strategies are used to meet real-time requirements?

RQ7 How is designer control handled (personality consistency, authorial intent, constraints, "lore compliance," controllable behaviors), and what control interfaces are provided?

RQ8 What resources are released (code, prompts, datasets, benchmarks, game environments), and how reproducible are the results?

3) **Quality & Transparency axis**

RQ9 How do studies address safety and alignment issues (toxicity, policy compliance, hallucinations, jailbreaks), especially in player-facing settings?

RQ10 What limitations are repeatedly reported, and what open research problems remain (evaluation validity, long-term consistency, multi-agent dynamics, personalization, etc.)?

### C. Search Strategy

We searched three digital libraries: ACM Digital Library, IEEE Xplore, and Scopus. We chose these sources because they collectively organise the most popular venues for human-computer interaction, game AI, and machine learning research, including ACM CHI, UIST, SIGGRAPH, NeurIPS, ECCV, and IEEE CoG. We set the publication date range to January 2022 to February 2026 to capture the period following the public release of instruction-tuned frontier LLMs (GPT-3.5, ChatGPT), which marks the practical starting point of LLM-based NPC research.

We made a Boolean search string using two complementary sub-queries joined by `OR`. The first sub-query targets papers that clearly combine LLM-based NPC or game concepts with LLMs and dialogue or behaviour terms. The second sub-query targets papers on generative or autonomous agents in virtual or open-world environments, which captures agent architectures applicable to LLM-based NPCs even when authors do not use the term "NPC." The full search string is as shown below.

```
(( ( ``non-player character'' OR npc )
AND ( ``video game'' OR videogame OR
gameplay ) AND ( ``large language
model'' OR ``generative ai'' OR llm )
AND ( dialogue OR dialog OR behavior OR
behaviour ) )      OR ( ( ``large language
model'' OR ``generative ai'' OR llm ) AND
( ``generative agent'' OR ``generative
agents'' OR ``autonomous agent'' OR
``autonomous agents'' OR ``embodied
agent'' OR ``embodied agents'' OR
``simulated agent'' OR ``simulated
agents'' ) AND ( ``virtual world'' OR
sandbox OR ``open world'' ) ))
```

We applied the search string to the title, abstract, and keyword fields in all three libraries. The search retrieved 323 records from ACM Digital Library, 8 from IEEE Xplore, and 46 from Scopus (384 records total). We removed 7 records duplicated across databases, yielding 377 records for screening.

### D. Study Selection

*a) Inclusion criteria.:* We included a paper if it met both of the following conditions: (1) it focuses on the use of LLMs or generative AI models for LLM-based NPC behaviour or autonomous agent behaviour in an interactive or simulated environment; and (2) it is written in English.

*b) Exclusion criteria.:* We excluded a paper if it met any of the following conditions: (1) it does not mention LLMs or AI agents in a simulated or interactive environment; (2) it does not address LLM-based NPCs or autonomous agent behaviour; or (3) it is a short abstract, poster, or workshop summary without a full technical or empirical contribution (fewer than 4 pages).

*c) Selection process.:* Two reviewers independently screened all 377 records by title and abstract as detailed in Figure 1. From this, 327 were excluded due to their lack of conformity with the inclusion criteria as well as the presence of the exclusion criteria, resulting in 50 papers. We resolved disagreements through discussion and, when consensus was not reached, by reading the full text of the disputed record. We applied the inclusion and exclusion criteria to all records that passed title and abstract screening, reading full texts where necessary to assess eligibility. The process yielded 42 papers for full data extraction.

### E. Quality Assessment

We assessed each included paper against five quality criteria derived from the PICOC outcomes (Table I) and the specific requirements of LLM-based NPC research. Two reviewers scored each criterion independently using a three-point scale (Yes / Partially / No), and we resolved disagreements through discussion. Table II lists the criteria and the distribution of scores across the final 42 included papers. Each criterion was scored Yes = 1, Partial = 0.5, No = 0, and summed to a 0-5 aggregate per paper; we excluded papers that scored below 3.0. To clarify, the quality assessment phase lowered the candidate set from 50 to 42 (8 of them scored below 3.0), as shown in Figure 1.

TABLE II
QUALITY ASSESSMENT CRITERIA AND SCORE DISTRIBUTION ($n = 42$).

| Criterion | Yes | Partial | No |
|---|---|---|---|
| QA1: Does the paper clearly describe the LLM(s) or generative model(s) used? | 23 | 11 | 8 |
| QA2: Is the LLM applied to NPCs or autonomous agents in an interactive or simulated environment? | 42 | 0 | 0 |
| QA3: Does the paper provide sufficient technical detail to understand or reproduce the approach? | 12 | 19 | 11 |
| QA4: Does the paper evaluate NPC behaviour or dialogue realism using qualitative or quantitative methods? | 35 | 5 | 2 |
| QA5: Is the study empirically validated (experiments, user studies, benchmarks, or simulations)? | 27 | 11 | 4 |

### F. Data Extraction

We extracted data from each included paper into a structured spreadsheet with 36 fields organised into seven thematic groups. One reviewer performed the initial extraction for each paper; a second reviewer verified all entries and flagged disagreements for discussion.

We populated categorical fields (e.g., architecture pattern, grounding components, interaction modalities) using controlled vocabularies agreed to before extraction began. We permitted multiple values per field where a paper combined more than one pattern or component. We recorded free-text evidence for each quality assessment criterion to support claims made during synthesis. For fields where we could not determine a value from the paper text alone (e.g., exact model version for papers using commercial platforms), we recorded the field as "unclear" and included an explanation in the evidence column rather than imputing a value. Included here

is the link to the file with the described usage of each variable used for extracting data[1].

## III. Results

We ran a systematic search across ACM Digital Library (323 candidates), IEEE Xplore (8), and Scopus (46) and obtained a final corpus of 42 papers published between 2023 and 2026 after we applied the inclusion and exclusion criteria and the quality assessment checklist. The corpus includes conference proceedings and journals (40) and arXiv preprints articles (2). The following subsections address each research question in turn.

*RQ1: What LLM-based tools, architectures, and design techniques have been proposed to enable lifelike NPC dialogue and behaviour in interactive environments?*

Several distinct architecture patterns appear across the corpus, ranging from single-prompt pipelines to multi-agent cognitive systems with learned components (Table III).

TABLE III
Architecture patterns identified in the corpus ($n = 42$). Papers may be coded under more than one pattern.

| Architecture Pattern | # Papers | Paper List |
|---|---|---|
| Single-agent (prompt pipeline) | 24 | [6, 7, 8, 2, 9, 10, 11] |
| Multi-agent system | 15 | [4, 12, 13, 14, 15, 16] |
| Planner + LLM | 17 | [5, 17, 18, 19, 20, 21, 22] |
| GOAP + LLM | 1 | [14] |
| Tool-using agent | 20 | [23, 24, 25, 18, 26, 27] |
| BT/FSM + LLM hybrid | 1 | [28, 14] |
| Multi-LLM modular pipeline | 2 | [17, 29] |
| RL + LLM hybrid | 2 | [30, 13] |
| Runtime code generation | 1 | [23] |
| Offline pre-generation | 3 | [31, 7, 28] |
| Text-conditioned simulation (diffusion) | 1 | [32] |

Researchers most often use the *single-agent prompt pipeline*. In this setup, one LLM receives a composed prompt containing persona, short-term conversation history, world state, and rules, then returns a dialogue response or action description. This pattern dominates dialogue-only LLM-based NPC research [6, 11, 8] because developers can integrate it into commercial engines (Unity, Unreal) via API calls.

*Planner + LLM* architectures break relatively extended tasks into sub-goal sequences that a separate controller or skill layer executes. DEPS [5] introduces a four-module loop (Descriptor, Explainer, Planner, Selector) in which the LLM replans after each failed sub-goal. STEVE [17] extends this with four specialised sub-agents (Planner, Critic, Curriculum, Describer) and a vision encoder. ODYSSEY [21] uses a planner-actor-critic triad with a skill library. These architectures target open-world task completion rather than LLM-based NPC conversation.

[1]https://doi.org/10.5281/zenodo.20490290

STEVE and Immersive Tailoring [29] exemplify the *multi-LLM modular pipeline*. Immersive Tailoring decomposes LLM-based NPC behaviour into seven specialised LLM modules (Detector, Scene, Speech, Pose & Animation, Role & Goal, Personality, Appearance) to balance capability against per-module latency.

Killer on Board [28] uses Unity's AI Navigation for movement and a behaviour tree for state transitions, while a local LLM (Hermes 3 3B) handles dialogue. MetaAgents [14] integrates Goal-Oriented Action Planning (GOAP) with an LLM for multi-agent job-fair simulations.

BPST [31] uses *offline pre-generation with adaptive runtime selection* and removes LLM inference from the game loop. The LLM generates multiple emotional tone variants of intent-based dialogue branches before launch, and a Pursuit Learning Automaton selects among them at runtime based on conversation-time feedback.

Across the corpus, authors most often use *system prompt / persona injection* (37/42 papers), *structured or JSON output* (33/42), *few-shot examples* (16/42), and *chain-of-thought* (12/42). About one-third of papers (12/42) use fine-tuning, primarily via LoRA adapters on open-weight models such as LLaMA-2 [17], LLaMA-3 [12, 21, 27], and RWKV [13].

*RQ2: Which system components are used to ground and control NPC outputs (e.g., memory, RAG, world-state grounding, planning, tool/function calling), and how are they combined?*

Table IV summarises how often each grounding or control component appears in the corpus. Short-term memory and world-state injection appear in most papers, while long-term memory, reflection, and RAG appear less often. Most papers use *sequential prompt concatenation* instead of learned integration: they put perception outputs into only one prompt, feed it to the LLM, and parse the response. Across the corpus, three integration patterns recur.

TABLE IV
Frequency of grounding and control components ($n = 42$).

| Component | # Papers | % |
|---|---|---|
| Short-term memory (prompt history) | 25 | 60 |
| World-state grounding | 29 | 69 |
| Rules / output constraints | 29 | 69 |
| Emotion / personality model | 20 | 48 |
| Planning module | 23 | 55 |
| Long-term memory (vector / summary) | 13 | 31 |
| Reflection / self-critique loop | 16 | 38 |
| Knowledge base / lore store | 18 | 43 |
| Tool / function calling | 7 | 17 |
| RAG / skill retrieval | 6 | 14 |
| Safety filter / guardrails | 2 | 5 |

*Lightweight single-pass grounding (dialogue-only papers):* Several player-facing dialogue systems assemble a single prompt per turn from a world-state summary, persona, short-term history, and rule constraints, emitting a response in one pass with no planning module. Rule lines constrain vocabulary, enforce lore, or block out-of-character responses, as seen across architecturally simple single-agent systems[8, 30, 33].

*Mid-tier looped grounding (social simulations):* Authors enhance the single-pass pattern with a long-term memory store, an emotion model that modulates next-turn behaviour, and optional reflection. Generative Agents [4] and MetaAgents [14] are the representative examples; the Orchestrating Adaptive AI system [11] extends this to player-facing RPG NPCs with trust vectors and event-driven memory decay, while CogniSociety [16] adds a shared vector database and emotion classification.

*Full cognitive loop (open-world task agents):* Agents run a repeated perception cycle: memory retrieval, reflection, planning, execution, and memory update cycle, with each step conditioning the next. ADAM [18], STEVE [17], and ODYSSEY [21] exemplify this pattern; LARM [30] replaces prompt chaining with end-to-end RL via PPO, and GROMIT [23] inverts it entirely by operating statelessly, reconstructing a fresh Semantic Scene Graph on every invocation to avoid memory drift.

*RQ3: What kinds of NPC capabilities are targeted (dialogue-only, behaviour-only, dialogue + action), and what interactions are supported (text, voice, embodied action)?*

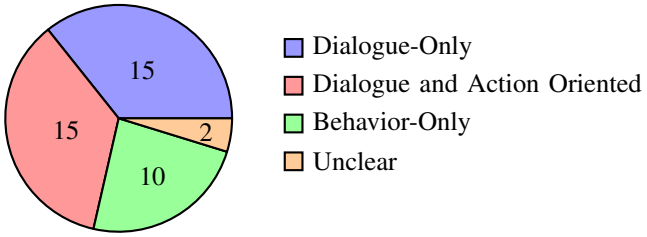



Fig. 2. The methods of behavior the LLM-based NPCs/Agents were programmed to perform.

The corpus is divided into three groups (Figure 2). *Dialogue-only* systems (15/42) restrict the LLM to generating speech and do not give it control over agent movement or game state [6, 7, 31, 2, 9]. *Dialogue and action integrated* systems (15/42) let the LLM emit speech and action commands, and the game engine parses and executes those commands [4, 24, 20, 29, 28, 34, 14]. *Behaviour-only* systems (10/42) target autonomous task completion without conversational dialogue as the primary objective [23, 5, 17, 18, 32, 19, 30, 21].

*Text* input and output appear 36 times across the corpus. *Voice* (speech-to-text and text-to-speech) appears in eight papers [6, 29, 35, 36, 37, 38], primarily in VR or accessibility-oriented contexts. *Embodied action* (the LLM-based NPC's physical presence or movement in a 3D world) appears in 19 papers, almost exclusively in Minecraft-based research or VR [5, 17, 18, 30, 21, 14, 29, 37]. *Multimodal perception* (processing visual game frames) appears in 10 papers [17, 18, 30, 26, 29, 39] and remains largely confined to research agents rather than deployed LLM-based NPC systems. *Gesture and animation* output appears rarely (4 papers: [40, 37, 29, 33]), and a separate animation system drives it rather than the LLM.

*RQ4: In what contexts are these NPCs deployed (game genres, virtual worlds, interVR/simulators), and what assumptions do authors make about the environment and world state access?*

The corpus spans six deployment contexts. *Open-world survival games*, primarily Minecraft, dominate the behaviour-only sub-corpus (DEPS [5], STEVE [17], ADAM [18], LARM [30], ODYSSEY [21], Talking-to-Build [20], Creative Agents [22]). *RPG and interactive fiction* settings dominate LLM-based NPC dialogue work, spanning commercial titles via modding (Skyrim + Mantella [2]), custom mystery games (Vaudeville [6], Killer on Board [28], Dejaboom! [24]), and purpose-built prototypes [31, 9, 11]. *Social simulations* form a growing category [4, 14, 15, 16]. *Serious games and training simulators* target educational or therapeutic goals [12, 19, 35, 36, 34]. *VR / XR environments* appear in three papers [35, 37, 29]. Three papers [19, 41, 25] apply their architectures to non-game simulation domains (education, energy, indoor navigation).

Thirty-three papers assume *partial* world-state access and inject a structured text summary of inventory, scene objects, or LLM-based NPC vicinity into the prompt. Only five papers grant the LLM *full* scene access via a live scene graph [23, 32, 11]. Four papers assume *no formal world-state access* and rely on conversational history and persona [7, 35, 31]. Eleven papers operate in *simulated-only* settings with no interactive game environment [41, 13, 16, 4].

Player-facing dialogue systems reduce the game world to flat object inventories, discarding geometric relationships that open-world task agents partially recover through visual token encoders [17] and multimodal goal representations [21].

This distinction supports the need for a sensitivity analysis comparing the two subsets across architecture, latency, safety, and reproducibility (Section IV-E).

*RQ5: What baselines and comparisons are used (scripted dialogue trees, behavior trees/FSM/GOAP/utility AI, non-LLM systems, or ablations), and what improvements are reported?*

Authors most often compare against ablations (19/42 papers), followed by non-LLM ML baselines such as behaviour-cloning policies or classical RL (9/42), scripted dialogue trees or static NPC conditions (7/42), and human-expert baselines (5/42). Notably, 16 papers include *no baseline* of any kind [8, 7, 23, 30, 39, 41] and rely exclusively on descriptive or qualitative evaluation.

Where baselines exist, authors report gains. For example, Generative Agents [4] demonstrates via ablation that the full architecture exhibits the most believable behaviour, with agent conduct rated more believable than both reduced variants and human crowdworker descriptions. ADAM [18] reaches diamond acquisition in 2.2× fewer exploration steps than the prior SOTA agent and achieves a 4.6× speedup on iron acquisition in a modified environment. The Classroom Simulacra TIR module [19] yields statistically significant improvement via paired t-test over baseline performance. For player-facing LLM-based NPC dialogue, Killer on Board [28] reports a GUESS composite score of 40.21 with a local 3B model with procedurally generated NPC identities in a Unity IDS prototype. These results should be noted as reflecting

individual study conditions and have not been independently replicated.

*RQ6: What trade-offs are reported between realism and engineering constraints (latency, cost, reliability), and what strategies are used to meet real-time requirements?*

32 of 42 papers treat real-time feasibility as a concern, but only seven report numerical latency measurements (Table V). End-to-end response times for cloud-hosted frontier models (GPT-4, GPT-4o) range from roughly 2.6 s to 15 s, well above the sub-second threshold that game UX literature commonly cites. Player-Driven Emergence [24] reports 15 s as detrimental to player experience, and 40% of participants identify latency as a pain point. The Immersive Tailoring system [29] reports 7.3 s for a tailoring interaction and 5.2 s for a standard interaction despite using GPT-4o-mini, and it attributes the dominant bottleneck to the Scene LLM's visual processing step (3.2 s mean).

TABLE V
LATENCY MEASUREMENTS REPORTED IN THE CORPUS. PAPERS THAT ACKNOWLEDGE LATENCY QUALITATIVELY, BUT PROVIDE NO FIGURES ARE EXCLUDED.

| Paper | Condition | Latency |
|---|---|---|
| GROMIT [23] | GPT-4, Sandbox | 10,600 ms |
| | GPT-4o, Sandbox | 5,200 ms |
| | GPT-4, Escape Room | 5,600 ms |
| | GPT-4o, Escape Room | 2,600 ms |
| Player-Driven Emergence [24] | GPT-4 (game loop) | 15,000 ms |
| Immersive Tailoring [29] | Tailoring path | 7,300 ms |
| | Non-tailoring path | 5,200 ms |
| Enhancing Game AI [27] | BTree generation | 147,000 ms |
| Text-Crowd [32] | Map generation | <6,000 ms |
| | Runtime (per step) | <1 ms |
| LARM [30] | Inference (embodied) | 580 ms |
| BPST [31] | Runtime selection | <1,000 ms |

Authors use several strategies to meet real-time or resource constraints. The most prevalent is switching to smaller or lighter models (14/42): replacing GPT-4 with GPT-4o, using LLaMA or RWKV variants, or applying LoRA fine-tuning [12, 29, 28, 27, 30]. Local model deployment (4/42) eliminates API round-trip latency entirely [28, 12, 31], and fallback scripted behaviour (4/42) defaults to a static response or idle animation on timeout [6, 23, 33]. Offline pre-generation (3/42) removes LLM calls from the game loop altogether [31, 42, 28]. The remaining strategies, such as structured output with retry loops, parallel modular processing, retrieval prefetch and caching, and memory compression, each appear in 2–4 papers [5, 18, 17, 24, 29, 43] and address invalid generation, critical-path latency, repeated inference, and context window limits, respectively.

Several papers delve into cost constraints qualitatively, but none quantify them. Authors repeatedly flag dependence on closed APIs (GPT-4/Codex) as a scalability barrier and position local fine-tuned models and offline pre-generation as additional mitigations.

*RQ7: How is designer control handled (personality consistency, authorial intent, constraints, "lore compliance," controllable behaviors), and what control interfaces are provided?*

Table VI summarizes the designer control mechanisms identified in the corpus. Authors use personality profiles and prompt templates in most papers, while authoring tools that provide a UI abstraction for non-technical designers remain rare.

TABLE VI
DESIGNER CONTROL MECHANISMS AND THEIR FREQUENCY ($n = 42$).

| Control Mechanism | # Papers | % |
|---|---|---|
| Personality profiles / persona | 26 | 62 |
| Prompt templates | 38 | 90 |
| Output constraints / rules | 27 | 64 |
| Controllable parameters (sliders, weights, budgets) | 25 | 60 |
| Lore knowledge base | 11 | 26 |
| Authoring tool / designer UI | 5 | 12 |
| Behaviour budgets (token caps, action limits) | 3 | 7 |

Several papers develop notable control approaches beyond generic persona prompting. GROMIT [23] introduces *SSG filtering* as a designer control lever: designers scope the Semantic Scene Graph passed to the LLM at design time to prevent generated code from accessing unintended scene elements. BPST [31] inverts the control problem by making the LLM generate a finite set of pre-approved dialogue variants offline and leaving runtime selection to a deterministic algorithm with no LLM agency. Immersive Tailoring [29] has a *player-facing* tailoring interface where users modify LLM-based NPC personality, appearance, and goals in natural language at runtime, giving players leverage over the designers. Killer on Board [28] gives authors direct control over who the LLM-based NPC is, its relationships, schedules, observations, and "protective scenes" from a fixed template.

Authors often treat prompt templates as sufficient for control, yet these templates are weaker against adversarial input and do not work in complex game worlds.

*RQ8: What resources are released (code, prompts, datasets, benchmarks, game environments), and how reproducible are the results?*

As shown in Table VII, fewer than one-third of papers release source code. Papers releasing full repositories include Generative Agents [4], OpenMap [25], DEPS [5], Text-Crowd [32], ADAM [18], Classroom Simulacra [19], Training Language Models [13], Killer on Board [28], Enhancing Game AI [27], the Effect of LLM-Based NPC Emotional States [44], Player-Driven Emergence [24], Creative Agents [22], and ODYSSEY [21]. Five categories of reproducibility barrier recur across the corpus:

Deprecated model APIs that prevent exact replication [5, 18, 17]; withheld system prompts and few-shot templates (32 papers) [23, 8, 20, 41, 34, 43]; commercial platform opacity (6 papers) from Inworld AI, Convai, or Worqhat [6, 44, 37,

TABLE VII
ARTIFACT AVAILABILITY ACROSS THE CORPUS ($n = 42$).

| Artifact Type | # Papers | % |
|---|---|---|
| Source code / repository | 13 | 31 |
| Prompts (full text) | 10 | 24 |
| Dataset | 13 | 31 |
| Benchmark / environment | 6 | 14 |
| Demo video | 10 | 24 |
| Nothing released | 18 | 43 |

38, 11]; unnamed or underspecified LLMs (8 papers) [31, 9, 34, 16, 15]; and paywalled game or dataset requirements such as Minecraft's commercial licence and Gameloft's unreleased behaviour-tree datasets [27]. Fewer than 30% of papers have sufficient transparency for independent replication, a systemic reproducibility gap in the field.

*RQ9: How do studies address safety and alignment issues (toxicity, policy compliance, hallucinations, jailbreaks), especially in player-facing settings?*

Authors report hallucination as the most critical safety concern (13/42), followed by bias (11)[7, 41, 14, 21, 20, 40, 29, 15, 16, 43, 6], toxicity (4)[20, 14, 15, 34], privacy (5)[19, 41, 29, 21, 38], player jailbreaking (2)[6, 23], and therapeutic safety (1)[34]. The most common remedy is persona and role constraint via system prompt (37/42), which is cheap to implement but vulnerable to adversarial input. Stronger measures are rare: constrained decoding appears in two papers [18, 28], human-in-the-loop design review in two others [7, 34], and a moderation API filter in one [20]. Only two papers implement runtime guardrail components as actual system elements [12, 28]; the rest either treat safety as a known limitation or depend only on prompt engineering.

*RQ10: What limitations are repeatedly reported, and what open research problems remain (evaluation validity, long-term consistency, multi-agent dynamics, personalization, etc.)?*

Table VIII lists the ten most frequently cited open problems, and hallucination, latency, and evaluation validity form a triad that appears in the limitations sections of most papers.

TABLE VIII
OPEN PROBLEMS IDENTIFIED BY AUTHORS ACROSS THE CORPUS, RANKED BY FREQUENCY OF MENTION ($n = 42$).

| Open Problem | # Papers |
|---|---|
| Evaluation validity (no benchmark, small $n$) | 33 |
| Hallucination and factual inconsistency | 26 |
| Real-time latency | 24 |
| Long-term memory consistency | 19 |
| Controllability vs. emergent behaviour | 23 |
| Dataset and benchmark gaps | 16 |
| Personalization across sessions | 22 |
| Safety in player-facing deployment | 8 |
| Multi-agent dynamics and inter-NPC coordination | 10 |
| Memory scaling to large agent populations | 8 |

*Evaluation validity.:* Researchers lack a shared LLM-based NPC believability benchmark, and papers use incompatible scales, including: GUESS-18 [28], PXI Mini [38], GEQ [20], SUS [40], MSE, NMMSP [37], success rate [5], BLEU/ROUGE [12]; this prevents direct comparison across studies. User studies typically draw small, demographically narrow samples (medians of $n \approx 25$–$30$, often university students in a single locale), which limits generalisation.

*Long-term consistency.:* LLM-based NPCs routinely contradict themselves within a session and fail to carry narrative state across sessions, despite attempts using memory compression [4, 17], dual memory architectures[11, 43, 34], and retrieval-augmented grounding [17, 27], but no solution achieves sustained cross-session coherence.

*Controllability versus emergence.:* Scripted approaches guarantee authorial consistency but eliminate emergent behaviour, while fully open LLMs produce novel interactions but remain unpredictable; BPST [31], and Killer on Board [28] offer partial mitigations, but neither provides a general-purpose resolution. The Open-Ended NPC Dialogue study [9] shows the optimal design point depends on player type.

*Multi-agent dynamics.:* Most reviewed architectures do not let LLM-based NPCs track other NPCs' states or dialogue history, with Generative Agents [4] and MetaAgents [14] being standouts here. Social emergence in multi-NPC systems (information propagation, coalition formation, emergent narrative conflict) remains largely unexplored in player-facing settings.

*Personalization.:* Robust persona prompts are the dominant individuation tactic, but do not adapt to player behaviour across sessions; BPST's Pursuit Learning Automaton [31] is the only reviewed system with dynamic player-feedback-driven adaptation, and it operates only over pre-generated dialogue. Twenty-two papers identify cross-session adaptive personality modelling as an open problem, and no reviewed system addresses it.

*Safety infrastructure.:* Hallucination is the most frequently reported reliability issue, yet only two papers implement runtime guardrail components [12, 28]. Thirteen of 42 papers report hallucination as the primary safety concern, while a separate 26 identify it as an open research problem(Table VIII). Deployments in child-facing [12] or therapeutic contexts [34] raise the stakes, yet neither has been stress-tested against adversarial input.

## IV. DISCUSSION

### A. Research Gaps and Open Directions

*Evaluation infrastructure:* The most noticeable gap is the absence of a unified benchmark for LLM-based NPC believability. A purpose-built benchmark could include (1) a publicly accessible game environment, (2) a fixed set of player-NPC scenarios with annotated ground-truth lore/dialogues, and (3) a standardised player-study protocol with pre-registered hypotheses.

*Long-term and cross-session memory:* Every paper that attempts cross-turn memory acknowledges that context windows fill, summarisation loses detail, and LLM-based NPCs eventually contradict themselves. Therefore, the main bottleneck for believable LLM-based NPCs may no longer be

dialogue generation itself, but memory persistence and identity maintenance over time. One step towards this is the Classroom Simulacra's Transferable Iterative Reflection mechanism [19], which generalises learned behavioural patterns across new groups. A related limitation is spatial encoding: flat text lists of objects are main world-state format, without geometric relationships that vision-language pipelines in STEVE [17] and Chain-of-Imagination [26] show are recoverable in game-like projects, but neither has conversational NPCs.

*Multi-agent social dynamics:* Only two reviewed architectures have NPCs communicating with each other in a game or game-like context [4, 14], yet emergent NPC-to-NPC behaviour is what open-world players expect. Extending social deduction work [13] to narrative RPG contexts (e.g., gossip, alliances, and player-driven plot) remains unexplored.

*Player-adaptive personalisation:* BPST [31] shows that an LLM-based NPC dialogue tone can shift over time, even operating on a fixed vocabulary of pre-generated types. Genuine personalisation, like updating conversational style and relationship stance from collected session data, is technically within reach of retrieval-augmented architectures, but no reviewed paper attempts it.

*Comparing LLM Implementation Results:* Comparing results across LLM-based NPC implementations is difficult because the corpus uses incompatible evaluation instruments (success rates, BLEU/ROUGE scores, GUESS-18, PXI Mini, GEQ, SUS, MSE, and NMMSP) often for single-study designs with no shared baseline. A system with 85% task success and one reporting a mean GEQ score of 3.8 cannot be meaningfully ranked, even when targeting the same NPC capability. Until a purpose-built believability benchmark exists, unified reporting standards (end-to-end latency, sample size, and at least one validated player-experience scale with task metrics) would let future reviews make cross-study comparisons that this one cannot.

### B. Barriers to Commercial Adoption

*[Barrier#1] Latency.:* The corpus treats latency as an engineering inconvenience, but not as a disqualifying issue. *Papers reporting per-turn response times of 7, 10, and 15 seconds frame these figures as manageable, but response times of this magnitude could be incompatible with real-time commercial game genres, though tolerance could vary substantially by genre and interaction design*. Of all the noticed patterns, latency is the biggest practical hurdle in the corpus.

Most user studies run under conditions tolerant of delay: short sessions, turn-based or text-only interaction, and no competitive pressure.

The papers' suggested fix, substituting smaller models, has a quality tradeoff that the corpus acknowledges but does not systematically measure. Switching to smaller or lighter models is the most common latency mitigation in the corpus, adopted by 14 of 42 papers, yet this substitution introduces a quality tradeoff that no reviewed paper measures directly. Killer on Board [28] explicitly acknowledges that its locally-deployed 3B model proved insufficient for distinct NPC personality and role-playing quality relative to larger alternatives. STEVE similarly documents that its fine-tuned 13B variant required more task iterations than its GPT-4 counterpart [17], and LARM reports performance degradation when smaller base models replace larger ones [30]. These cases suggest that *reducing model size might lower latency at the cost of the consistency and believability properties that make LLM-based NPCs compelling* in the first place. Researchers and developers could benefit from studies that treat this as a controlled variable, measuring, for a fixed NPC task, how response quality, hallucination rate, and character consistency degrade as model size decreases, so that the latency-quality frontier can be mapped rather than assumed.

*[Barrier#2] Consistency and quality assurance.:* Commercial games need deterministic content review, but LLM outputs are non-deterministic. Even at temperature zero, minor context changes can alter outputs. Toward Dynamic NPC [33] reports that short personality prompts "can't cover all scenarios" and that GPT fills gaps with unintended violence.

*[Barrier#3] Cost at player scale.:* A title with one million daily active players generating 50 LLM-based NPC exchanges each produces 50 million API calls per day. This suggests a potentially significant financial burden for studios operating at consumer scale, though actual costs would depend heavily on caching strategies, model selection, and dialogue frequency. No reviewed paper quantifies cost at player scale; the previous estimate is the authors' own calculation based on publicly available API pricing, intended to illustrate the scope of the problem. Running the model locally avoids API costs but means either bundling large model files with the game or maintaining your own server.

*[Barrier#4] Multiplayer synchronisation.:* All reviewed architectures assume a single player interacting with an LLM-based NPC; Section IV-D discusses the synchronisation, consistency, and security challenges that arise in multiplayer settings.

*[Barrier#5] Content rating and regulatory compliance.:* LLMs can say almost anything, which makes it impossible for studios to guarantee a content rating. Players in the Vaudeville study [6] actively jailbroke NPCs and shared working exploits with each other; GROMIT acknowledges the same problem [23] but leaves the fix to future work.

*[Barrier#6] Localisation.:* Frontier LLMs work noticeably worse in non-English languages, and the gap widens for languages underrepresented in training data. Fine-tuning smaller models for specific languages helps. The Japanese escape game [42] study, for example, identifies ELYZA and rinna as candidates. The issue here, though, is that each language needs its own pipeline, which multiplies cost and the chance of culturally tone-deaf outputs.

### C. Copyright, Security, and the Impact on Creative Labour

The corpus does not address whether LLM-based NPC dialogue reproduces copyrighted material, leaving ownership of generated story beats an open question. Fine-tuning on domain-specific data [12, 28] partially mitigates this but cannot eliminate the risk when base model training data remains undisclosed. Two papers document active jailbreaking in

player-facing deployments [6, 23], and prompt injection via in-game text represents an additional unresolved attack surface, yet no reviewed paper brings a security-oriented workaround beyond prompt-level defences.

Beyond security, the corpus neither addresses the privacy obligations (GDPR, COPPA) arising from routing player dialogue through commercial APIs, nor measures whether the labour shift from writing dialogue to engineering prompts [4, 28, 43] actually reduces authoring cost, and several systems pair LLM output with text-to-speech [29, 37, 38], leaving the legal and ethical stakes of synthesised performance unexamined.

### D. Technical Roadblocks on the Software Side

*[S.E. Challenge #1] Game engine integration.:* Unity and Unreal were built for deterministic game logic, not asynchronous API calls with multi-second response times, so integrating an LLM requires either stalling the main thread or building custom async systems to keep animations and NPC logic running while the model processes. As of the reviewed period, no standardised LLM integration layer had emerged in mainstream engines, and third-party packages like LLMUnity [28] only partially fill the gap.

*[S.E. Challenge #2] Determinism and save-state reproducibility.:* A game cannot trivially save and restore runtime-generated LLM-based NPC dialogue: memory, conversation history, and model random seed would all need serialisation; critical for branching narratives where LLM-based NPC dialogue has downstream consequences.

*[S.E. Challenge #3] Testing and quality assurance methodology.:* Standard quality assurance does not work on outputs that are different every time, and none addresses what happens when the underlying model gets updated or retired; a real concern given that DEPS [5] and ADAM [18] rely on model versions no longer available.

*[S.E. Challenge #4] Context window limits as a game-state bottleneck.:* An open-world NPC realistically needs faction standings, active quests, and hundreds of world-state flags; far more than fits in a context window, and RAG only helps if the retrieval step already knows what the player is trying to do, which is precisely what the LLM is being asked to figure out.

*[S.E. Challenge #5] Multiplayer and shared-world state synchronisation.:* Multiplayer LLM-based NPCs introduce a synchronisation problem (IV-B) and an additional attack surface, where adversarial players can influence an NPC's state in ways that affect all subsequent players in that session.

None of these five challenges has a reviewed solution; closing them is a prerequisite for commercial deployment.

> Longer-term adoption of LLM-based NPCs hinges on three developments: (1) cost and speed improvements for sub-2 B parameter local models; (2) legal clarity on generated-content ownership and training-data licensing; and (3) a shared evaluation benchmark. None requires a research breakthrough, but none is imminent.

### E. Threats to Validity

Extraction subjectivity is the primary internal threat. Inconsistent terminology across papers required interpretive judgment. For example, distinguishing a planner+LLM architecture from an augmented single-agent system is not always clear-cut. We mitigated this with a fixed coding schema applied consistently across all 42 papers.

Comparing player-facing NPC studies (20/42) against broader simulation/agent studies (22/42) for architecture, latency, safety, and reproducibility, we get two groups. For clarity, this was introduced in the Results for RQ4. Player-facing papers prefer single-agent prompt pipelines (17/20, 85%) over the multi-agent and planner+LLM designs, which are more common in broader-agent work (11/22, 50% and 13/22, 59%). Broader-agent papers release code and artifacts more often (code: 10/22, 45% vs. 3/20, 15%; any artifact: 15/22, 68% vs. 9/20, 45%). Real-time constraints and safety coverage are comparable across subsets (85% vs 68%; 55% vs 50%), showing these are overall corpus patterns. These differences don't weaken the review's findings. The reproducibility gap (RQ8), the lack of safety guardrails (RQ9), and the latency-quality compromise (IV-B) hold within both subsets. We still have the broader corpus from this, with "LLM-based NPC and adjacent virtual-agent systems" where a claim involves both groups.

The corpus covers only three databases (ACM Digital Library, IEEE Xplore, Scopus) and papers published between 2023 and early 2026. Industry technical reports, workshop papers without proceedings, and unindexed preprints are systematically absent.

Rapid technological change threatens this review's durability. Findings tied to specific model generations (GPT-4, GPT-4o, 3–7 B local models) may date quickly; readers should treat quantitative results as indicative of the 2023–2026 period.

## V. Conclusion

NPCs have always been bound by what designers explicitly script. LLMs introduced the possibility of characters that converse openly, adapt to unforeseen situations, and exhibit behaviours not hand-authored. This review systematically examines whether the field has realised that possibility, for whom, and under what conditions. Across 42 papers published between 2023 and 2026, the findings reveal substantial architectural diversity alongside a critical failure to build shared infrastructure, with reproducibility, evaluation validity, and runtime safety all remaining structurally untouched.

Future work should prioritise the following: First, the field would benefit from a shared benchmark so researchers can compare believability and coherence claims directly. Second, long-term cross-session memory remains the most critical unsolved problem. Third, hybrid designs offer the clearest path to commercial viability, yet only two papers in this corpus test them. Addressing these gaps, alongside the falling cost and latency of smaller local models, could close the distance between research prototypes and shipped games.

## REFERENCES


[1] I. Millington and J. Funge, *Artificial Intelligence for Games*, 2nd ed. Morgan Kaufmann, 2009.

[2] R. Trukhin, J. F. Isaksson, and M. Palosaari Eladhari, "Llm-powered npcs: Evaluating the impact of large language models on npc design," in *Interactive Storytelling: 18th International Conference on Interactive Digital Storytelling, ICIDS 2025, Saint Julian, Malta, December 1–5, 2025, Proceedings, Part II*. Springer-Verlag, 2025.

[3] P. Sweetser, "Large language models and video games: A preliminary scoping review," in *CHI 2025: CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2025.

[4] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, 2023.

[5] Z. Wang, S. Cai, G. Chen, A. Liu, X. Ma, Y. Liang, and T. CraftJarvis, "Describe, explain, plan and select: interactive planning with large language models enables open-world multi-task agents," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2023.

[6] S. R. Cox and W. T. Ooi, "Conversational interactions with npcs in llm-driven gaming: Guidelines from a content analysis of player feedback," in *2024 IEEE Conference on Games (CoG)*, 2024.

[7] M. Yin, E. Wang, C. Ng, and R. Xiao, "Lies, deceit, and hallucinations: Player perception and expectations regarding trust and deception in games," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024.

[8] G. Radež and C. Bohak, "Integrating environmental awareness into npcs: Contextual conversational interaction in games," in *HCI SI 2024 (Human-Computer Interaction Slovenia 2024)*, 2024.

[9] R. Ploug, E. Rimer, A. K. Skov Petersen, and M. Scirea, "Open-ended npc dialogue favors casual players: A pilot comparison of three llm-driven dialogue systems," in *2025 IEEE Conference on Games (CoG)*, 2025.

[10] D. Singh, J. Banerjee, and J. Pandey, "Ai-driven npc dialogues for immersive gameplay: Integrating openai's nlp technology in unity-based games," in *2025 International Conference on Intelligent Computing and Virtual & Augmented Reality Simulations (ICVARS)*, 2025.

[11] A. Shaikh, H. Chinchakar, V. Jagtap, A. Kulshreshtha, and P. Jain, "Orchestrating adaptive ai in video games using dynamic sentiment modulation and dual-memory architectures," in *ICT Analysis and Applications*. Springer Nature Switzerland, 2026.

[12] C. Valliyammai, K. Jennie, K. Rameshbabu, and D. P. Prshanth, "Multi-agent interactive game simulation for enhancing child safety," in *IntelliSys 2025, published in Springer LNNS 1553*, 2025.

[13] B. Sarkar, W. Xia, C. K. Liu, and D. Sadigh, "Training language models for social deduction with multi-agent reinforcement learning," in *AAMAS '25: Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2025.

[14] Y. Li, L. Sun, and Y. Zhang, "Metaagents: Large language model based agents for decision-making on teaming," *Proc. ACM Hum.-Comput. Interact.*, vol. 9, no. 2, 2025.

[15] H. Zhang, J. Yin, M. Jiang, and C. Su, "Can agents spontaneously form a society? introducing a novel architecture for generative multi-agents to elicit social emergence," in *Adjunct Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST Adjunct '25. Association for Computing Machinery, 2025.

[16] C. Wu and G. Li, "Enhancing generative agents in social simulations: Bridging memory, emotion, and governance for realistic social simulations," in *Intelligent Systems and Applications*, K. Arai, Ed. Springer Nature Switzerland, 2025.

[17] Z. Zhao, W. Chai, X. Wang, B. Li, S. Hao, S. Cao, T. Ye, and G. Wang, "See and think: Embodied agent in virtual environment," in *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VIII*, 2024.

[18] S. Yu and C. Lu, "Adam: An embodied causal agent in open-world environments," 2024. [Online]. Available: https://arxiv.org/abs/2410.22194

[19] S. Xu, H.-N. Wen, H. Pan, D. Dominguez, D. Hu, and X. Zhang, "Classroom simulacra: Building contextual student generative agents in online education for learning behavioral simulation," in *HCII 2025 (HCI International 2025), published in LNCS 15816 by Springer Nature Switzerland*. Association for Computing Machinery, 2025.

[20] X. Sun, L. Wang, Y. Li, J. Li, M. Poesio, J. Frommel, K. Hindriks, and J. Pei, *Talking-to-Build: How LLM-Assisted Interface Shapes Player Performance and Experience in Minecraft*. Association for Computing Machinery, 2025.

[21] S. Liu, Y. Li, K. Zhang, Z. Cui, W. Fang, Y. Zheng, T. Zheng, and M. Song, "Odyssey: empowering minecraft agents with open-world skills," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, ser. IJCAI '25, 2025.

[22] P. Cai, C. Zhang, Y. Fu, H. Yuan, and Z. Lu, "Creative agents: Empowering agents with imagination for creative tasks," in *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*. PMLR, 2025.

[23] N. Jennings, H. Wang, I. Li, J. Smith, and B. Hartmann, "What's the game, then? opportunities and challenges for runtime behavior generation," in *2025 IEEE Conference on Games (CoG)*. Association for Computing Machinery, 2025.

[24] X. Peng, J. A. Quaye, S. Rao, W. Xu, P. Botchway, C. Brockett, N. Jojić, G. DesGarennes, K. Lobb, M. Xu, J. Leandro, C. Jin, and B. Dolan, "Player-driven emer-

gence in llm-driven game narrative," in *2025 IEEE Conference on Games (CoG)*, 2025.

[25] D. Li, Z. Yang, G. Qi, S. Pang, G. Shang, Q. Ma, and Z. Yang, "Openmap: Instruction grounding via open-vocabulary visual-language mapping," in *Proceedings of the 33rd ACM International Conference on Multimedia*, ser. MM '25. Association for Computing Machinery, 2025.

[26] E. Zhou, Y. Qin, Z. Yin, Z. Shi, Y. Huang, R. Zhang, L. Sheng, and J. Shao, "Chain-of-imagination for reliable instruction following in decision making," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

[27] C. I. Păduraru, M. G. Paduraru, and A. Ştefănescu, "Enhancing game ai behaviors with large language models and agentic ai," in *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering*, 2024.

[28] C. B. Oudrup, F. B. Krog, J. Winkel Missel, L. L. Abildgaard, M. K. Hansen, N. B. Jensen, and H. S. Fog, "Killer on board: Addressing the narrative paradox by utilizing llm-driven npcs," in *Interactive Storytelling: 18th International Conference on Interactive Digital Storytelling, ICIDS 2025, Saint Julian, Malta, December 1–5, 2025, Proceedings, Part I*, 2025.

[29] A. Bellucci, G. Jacucci, K. D. Trung, P. K. Das, S. V. Smirnov, I. Ahmed, and J.-L. Lugrin, "Immersive tailoring of embodied agents using large language models," in *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2025.

[30] Z. Li, X. Xu, Z. Xu, S.-N. Lim, and H. Zhao, "LARM: Large auto-regressive model for long-horizon embodied intelligence," in *Proceedings of the 42nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, Eds., vol. 267. PMLR, 2025.

[31] M. M. Jahangiri and P. Rahmani, "Balancing game satisfaction and resource efficiency: Llm and pursuit learning automata for npc dialogues," in *2025 IEEE Conference on Games (CoG)*, 2025.

[32] X. Ji, Z. Pan, X. Gao, and J. Pan, "Text-guided synthesis of crowd animation," *ACM Transactions on Graphics*, vol. 43, no. 4, 2024.

[33] J. Stojanov, L. R. Kouzelis, and O. Spantidi, "Toward dynamic npc interactions: Integrating gpt-driven agents in 3d virtual environments," in *Advances in Visual Computing*, 2025.

[34] J. Li, H. Zhou, W. Jian, and Z. Liu, "Bartender's advice: A cbt-based serious game for anxiety relief utilizing generative agents," in *HCI in Games: 7th International Conference, HCI-Games 2025, Held as Part of the 27th HCI International Conference, HCII 2025, Gothenburg, Sweden, June 22–27, 2025, Proceedings*, 2025.

[35] L. Kruse, S. Rings, and F. Steinicke, "My focus was on the game: Investigating the use of ai assistants in a virtual reality exergame for older adults," in *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2025.

[36] Y. Zhao, J. Pan, J. Cao, J. Zhang, Y. Dong, Y. Wang, P. Hansen, and G. Wang, "Unlocking the power of speech: Game-based accent and oral communication training for immigrant english language learners via large language models," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2024.

[37] M. Schmidbauer, J. Pirker, E. Mayer, and T. Odaker, "Social presence in virtual reality: A comparative study of ai npcs and human instructors," in *Proceedings of the 2025 18th ACM SIGGRAPH Conference on Motion, Interaction, and Games*. Association for Computing Machinery, 2025.

[38] N. Zargham, A. Merkord, A. Safari, L. T. Dratzidis, and R. Malaka, ""can you feel me now?": Exploring player empathy in ai-based npc conversations," in *Entertainment Computing – ICEC 2025: 24th IFIP TC 14 International Conference, Tokyo, Japan, August 27–30, 2025, Proceedings*, 2025.

[39] Q. Zhang, R. Wen, L. B. Hendra, Z. Ding, and R. LC, "Can ai prompt humans? multimodal agents prompt players? game actions and show consequences to raise sustainability awareness," in *2024 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. Association for Computing Machinery, 2024.

[40] Q. Sun, H. Qiao, J. Zhu, Y. Guo, Y. Jin, T. Yuan, S. Wang, and H. Mi, "Myself, my child or my friend: Understanding player relationships with ai-generated game characters," in *CHI EA '25: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2025.

[41] S. Chetty, H. Le Vu, H. Wang, and R. Smyth, "An llm framework for inferring household energy consumption through behaviour simulation," in *IntelliSys 2025, published in Springer LNNS 1553*. Association for Computing Machinery, 2025.

[42] M. Sakura, K. Tasaka, M. Numao, and S. Nishide, "Npc dialogue generation in a custom-built game using generative ai," in *2025 IEEE 14th Global Conference on Consumer Electronics (GCCE)*. IEEE, 2025.

[43] S. Buongiorno, L. J. Klinkert, Z. Zhaung, T. Chawla, and C. Clark, "Pangea: Procedural artificial narrative using generative ai for turn-based, role-playing video games," in *Proceedings of the Twentieth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2024.

[44] A. Marincioni, M. Miltiadous, K. Zacharia, R. Heemskerk, G. Doukeris, M. Preuss, and G. Barbero, "The effect of llm-based npc emotional states on player emotions: An analysis of interactive game play," in *CHI 2025: CHI Conference on Human Factors in Computing Systems*, 2025.