
Proceedings Paper:
Ternar, Dan-Alexandru, Denisova, ALENA, Cunha, João Miguel et al. (2026) Generative AI in Game Development:A Qualitative Research Synthesis. In: 2026 CHI Conference on Human Factors in Computing Systems:Proceedings. 2026 CHI Conference on Human Factors in Computing Systems, 13-17 Apr 2026 Human factors in computing systems. ACM, ESP. (In Press)

https://doi.org/10.1145/3772318.3791206

Generative AI in Game Development: A Qualitative Research Synthesis

Dan-Alexandru Ternar
Department of Computer Science
Aalto University
Espoo, Finland
alex.1.ternar@aalto.fi

Alena Denisova
Department of Computer Science
University of York
York, United Kingdom
alena.denisova@york.ac.uk

João Miguel Cunha
CISUC, Department of Informatics Engineering
University of Coimbra
Coimbra, Portugal
jmacunha@dei.uc.pt

Annakaisa Kultima
Aalto University
Helsinki, Finland
annakaisa.kultima@aalto.fi

Christian Guckelsberger
Department of Computer Science
Aalto University
Espoo, Finland
christian.guckelsberger@aalto.fi

Abstract

Generative Artificial Intelligence (GenAI) is currently reshaping game development practices, production pipelines, and value networks in an unprecedentedly pervasive manner with cascading consequences remaining unclear. In the last five years since GenAI's inception, a growing body of qualitative research has explored these early transformations from different settings and demographic angles. However, these studies often contextualise and consolidate their findings weakly with related work; for research to keep up with and support stakeholders in this development, the current moment calls for a synthesis of the findings emerged thus far. Here, we address this need through a qualitative research synthesis via meta-ethnography. We followed PRISMA-S to systematically search the relevant literature from 2020-2025, including major HCI and games research databases. We then synthesised the ten eligible studies, conducting reciprocal translation and line-of-argument synthesis guided by eMERGe, informed by CASP quality appraisal. We identified nine overarching themes, provide recommendations, and contextualise our insights in wider game production trajectories. With this work, we seek to provide practitioners, researchers and policy-makers with grounded insights to guide practice, research and governance.

CCS Concepts

• **Human-centered computing** → *Empirical studies in HCI*; *HCI theory, concepts and models*; *Computer supported cooperative work*; • **Applied computing** → **Computer games**; • **Computing methodologies** → *Generative and developmental approaches*; • **General and reference** → Surveys and overviews.

Keywords

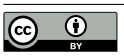
generative AI, genAI, game development, systematic review, qualitative research synthesis, qualitative research appraisal, meta-ethnography, large-language models, LLMs, text-to-image generation

ACM Reference Format:
Dan-Alexandru Ternar, Alena Denisova, João Miguel Cunha, Annakaisa Kultima, and Christian Guckelsberger. 2026. Generative AI in Game Development: A Qualitative Research Synthesis. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3772318.3791206

1 Introduction

Generative Artificial Intelligence (GenAI) has disrupted and is transforming creative workflows across most industries and forms of creative practice. Game development, traditionally an early adopter of new technologies, is no exception. Transformation through GenAI has been identified as a major and lasting trend in this domain [31]. The rapid uptake of GenAI in commercial game production has also been confirmed through publishing data: recent reporting on *Steam* submissions indicates that at least[1] 7% of all games released on the platform now disclose some form of AI usage for content creation, up from only 1% the year before [32]. Around 60% of disclosed implementations involve visual asset generation, demonstrating how art pipelines are a central entry point for GenAI in game production.

The adoption of GenAI in game production calls for researchers not only to improve systems and interaction modalities, but also to shape a societally and economically sustainable future of human-AI co-creation [25] and support the advancement of games as diverse cultural artefacts. Due to its complex and rapidly evolving nature,



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791206

[1]This likely under-represents the actual scale of adoption since disclosure is inconsistent. Reflecting this trend, *Valve* (a video games company) has updated its *Steam* (online platform for buying, downloading, and playing video games) content survey to require disclosure of both pre-generated AI content (produced during development) and live-generated AI content (produced during gameplay).

we deem *qualitative research* the prime vehicle to explore this phenomenon. Existing qualitative studies have, amongst others, investigated how GenAI reshapes workflows, raises ethical questions, and transforms relations between developers, artists, and players. Crucially though, while scholars have dedicated studies to the adoption of GenAI in game production at a rarely seen and still increasing level of intensity, these investigations are often *limited* to specific types of e.g. target demographic and production setting, or only a narrow aspect of the overall phenomenon. In addition, the research landscape remains *fragmented*, reflecting the "one-off problem of failing to build upon prior work" [16] – a broader pattern in qualitative research. While some qualitative studies (Sec. 3.2) notably relate to previous work more extensively, this is (naturally) done in the service of distinguishing and motivating their own focus or contributions, and not from a neutral perspective. This makes it difficult for researchers, practitioners, and policymakers to grasp the big picture and assess how strongly smaller observed phenomena are empirically supported. A *synthesis* is needed to translate insights across studies, identify consistencies and contradictions, and highlight areas requiring further exploration – both in terms of the studies phenomena and the methodology used.

Against this backdrop, our aim was to *systematically review and synthesise qualitative research on the adoption and impact of GenAI in game production*. Specifically, we seek to answer the following research questions (RQs):

- **RQ1** How do existing qualitative studies differ, e.g. in research questions, methodology, demographics and setting?
- **RQ2** How do the findings from these qualitative studies characterise the ways in which GenAI is adopted, used and perceived within game development workflows?
- **RQ3** In what ways do findings across studies converge, complement, or contradict one another?
- **RQ4** Which conceptual, empirical, or methodological gaps remain? How might these shape future research priorities?

We address these questions through a *meta-ethnography* as the arguably most established, interpretive Qualitative Research Synthesis (QRS) method (see Sec. 4 for further justification). QRS has been highlighted as a systematic and reasonably robust method [11, p. 69] capable of identifying saturated as well as underexplored areas of research [11, p. 120]. To ensure methodological rigour, we follow the *eMERGe* reporting guidelines [19] for meta-ethnography, documenting our search and screening via *PRISMA-S* [62], and appraising primary study quality with the *Critical Appraisal Skills Programme (CASP)* [35]. We focus on studies from 2020-2025 investigating the use of publicly available GenAI in "offline" game production (see Sec. 5.1 for further scoping). Drawing together insights from primary research, this synthesis not only consolidates a scattered evidence base, but also probes the generalisability of observed phenomena.

The result is the first integrated qualitative evidence base on GenAI's impact on game production, offering *industry stakeholders* a clearer picture of emerging practices to adopt or avoid while equipping *researchers*, *funders* and *policymakers* with grounded insights to guide future investigations, support decisions and governance. More specifically, we make at least three *contributions*. Firstly, in answering **RQ1**, we map the existing qualitative research landscape on GenAI in game development in terms of what studies sought out to do, in which context, and by which means, thus providing a much needed overview and signpost to studies with specific foci. Secondly, through interpretative work on **RQ2/3**, we abstract the core phenomena found by qualitative researchers and make transparent how they are similarly supported, enriched or contradicted by different studies, thus providing a nuanced overview of a rapidly evolving research complex to guide practice, research and governance. This includes, for instance, how practitioners negotiate the integration of GenAI in their creative and technical workflows, and how GenAI reshapes developers' perceptions of creative agency, authorship, and craft accountability within game development. Based on these insights, we derive implications and recommendations for all stakeholders. Thirdly, via **RQ4** and supported by **RQ1/3**, we provide the necessary insights to inform future research priorities or caution decision-making based on incomplete insights.

2 Background

While generative systems date back to early computer art, e.g., [12, 68], today's Generative Artificial Intelligence (GenAI) refers mainly to machine learning systems, particularly Transformers [72] and Diffusion Models [22]. GenAI is reshaping game design and development, creating both opportunities and challenges across creative and technical processes [61]. We briefly outline the advancements of GenAI relevant to this work, focusing on text, image, 3D and audio content.

The development of Large Language Models (LLMs) has transformed natural language processing, enabling models to generate highly realistic text from minimal examples [7]. GPT-1 (2018) [58] showed that strong performance across diverse language tasks was possible, paving the way for today's state-of-the-art systems. LLMs continue to evolve rapidly, shaping both game writing and technical workflows. GPT-3 and 3.5 [7] popularised conversational interaction through *ChatGPT*, while GitHub *Copilot* demonstrated how code generation could accelerate engine scripting and gameplay logic. Google's latest *Gemini 2.5* model adds advanced multimodal reasoning, allowing integration of text, image, and audio in design workflows [24, 50]. Unity *Muse* [71] (introduced in Unity 6) now embeds LLM assistance directly into engine editor workflows, enabling code completion, asset search, and interaction prototyping out-of-the-box.

With advances in LLMs, Text-to-Image Generation (TTIG) systems emerged [7]. Caption-conditioned models had already shown that text could guide image generation [41], and CLIP [57] strengthened the field by learning visual concepts from natural language. OpenAI's *DALL·E* (2021) [48, 60] combined a Transformer [72] with CLIP [59], while *DALL·E 2* [49] replaced the Transformer with Diffusion [22], producing images that were more realistic and faithful to prompts. Systems such as *DALL·E 3*, *Midjourney* [45], *Stable Diffusion* [1, 64] are now widely used for creative work.

Naturally, AI's capabilities have been extended from two- to three-dimensional content generation. Text2Mesh [44] introduced text-driven colour and geometry editing of meshes, while CLIP-Forge [65] proposed zero-shot text-to-3D generation using CLIP embeddings without paired data. Recent systems can now transform simple text descriptions into 3D models [14]. Tools such as

Alpha 3D [3], *Luma AI* [36] and *Meshy.ai* [42] create textured, game-ready assets from text or image prompts. *Promethean AI* [56] assists designers by automating environment assembly [4], while *Genie 3* generates interactive 3D worlds from a single prompt with evolving dynamics [53]. Similarly, Tencent's *Hunyuan-Game model* [34] extends generation to entire asset sets, including characters, effects, and video snippets.

As our final group of GenAI to introduce, generative audio systems are beginning to automatise sound and voice production. Tools such as *Soundraw* [69] provide AI-assisted adaptive music composition, while *ElevenLabs* [17] enables lifelike voice acting with controllable accents and emotions. Diffusion-based systems like *Producer.ai* [55] and research frameworks such as *Meta's Audiocraft* [43] expand possibilities for procedural soundscapes. These developments hint at future workflows where background scores, diegetic sound effects, and even NPC voices can be prototyped without traditional recording pipelines, providing opportunities but also raising concerns about replacing skilled people [18, 74].

3 Related Work

To our knowledge, no prior work has systematically synthesised qualitative research on the impact of GenAI on game development or developer experiences. We position our synthesis relative to adjacent reviews within games, clarifying differences in scope and methods. We also relate to the few primary studies captured by our synthesis that connect their findings with previous work, complimenting their scholarship and further motivating our approach.

3.1 Reviews on Generative AI in Games

Sweetser [70] conducted a scoping review and reflexive thematic analysis of early research (2022-Spring 2024) on LLMs and video games across five application themes. While most reviewed papers are technical, some qualitative observations from developers' interaction with GenAI are included, making it closest to our work. Key differences lie in technical focus, methodology, and timeframe. We focus exclusively on qualitative studies of developer experiences, over a longer timeframe (2020-Summer 2025), and, through meta-ethnography, provide a more systematic and conceptually richer synthesis of existing work, which preserves the interpretations of the primary literature.

Similarly, Moon et al. [46] combine qualitative, empirical and quantitative technical work to assess the relevance and challenges of GenAI in educational game design across mutiple stakeholder groups. The review and analysis methodology as well as exact timeframe of covered literature is left opaque, and the report does not connect claims directly to specific qualitative insights from the primary literature. We, in contrast, provide a systematic, grounded synthesis of exclusively qualitative studies on GenAI's impact on game production, focusing on developers' insights while covering diverse production contexts and game types, including educational games.

Other reviews draw on technical literature even more substantially to identify opportunities and challenges in using specific GenAI technologies – often research prototypes – for game development. Focusing on General Pre-Trained Transformer (GPT) models and technical venues, Yang et al. [76, 77] provide a scoping review of GPT applications to games from 2020 to 2024 across Procedural Content Generation (PCG), mixed-initiative game design, mixed-initiative gameplay, playing games, and game user research. The reviews do not report a specific literature search and analysis standard. Other technical reviews embrace a broader variety of AI but narrow their focus on specific applications in the game development pipeline. Using PRISMA, Wu et al. [75] systematically review papers demonstrating and advocating the use of GenAI for game character creation across concept, modelling, animation, and behaviour design (2019-Autumn 2024). While their emphasis is on design-time support, Maleki and Zhao [39] review methods to implement PCG primarily at runtime, with a focus on LLMs. Their review and GPT-3.5d analysis covers literature from 2019 to 2023 but does not mention a specific standard. Ribeiro et al. [63] employ PRISMA to systematically review applications of image generation models for image-based game asset production, with a focus on image quality metrics. With papers from 2016 to 2023, the review also covers techniques that are not associated with GenAI in the popular sense (cf. Sec. 2).

In contrast to these reviews, we focus exclusively on *synthesising* rich, *qualitative* insights from studies dedicated to documenting *game developers' experience* with GenAI. Similar to the technical reviews mentioned above, this data also conveys opportunities and shortcomings of GenAI applications but is grounded in *concrete, in-situ insights from actual game development* rather than researcher proposals on prototype potentiality. In common with these reviews, we also seeks to identify gaps and opportunities in research. However, these concern *gaps in evidence and methodology*, not in technology or specific applications. To support the transfer of insights, we focus on serviced, publicly available systems and exclude work on non-public research prototypes. Moreover, we focus on GenAI technologies more generally, rather than specific flavours, and on its impact on game development across contexts, rather than a specific aspect of the pipeline. Arguably most importantly, we conduct a methodologically rigorous, purely qualitative research synthesis via meta-ethnography [10, 47], following established standards across the steps of literature search [PRISMA-S, 62], quality appraisal [CASP, 35] as well as analysis and synthesis [eMERGe, 19], facilitating the emergence of new conceptual insights. Epistemologically speaking, while the existing reviews primarily aggregate first-order interpretations from researchers, we interpret researchers' second-order constructs of their participants' experiences into third-order constructs as part of our synthesis, using participants' first-order interpretations to ground the analysis and synthesis (see Sec. 4.1 for a distinction between aggregative/interpretative synthesis and$1^{st}/2^{nd}/3^{rd}$-order interpretations).

3.2 Consolidation in Primary, Empirical Studies of GenAI in Games

Across our corpus of primary studies, only six out of 10 relate their findings to existing empirical work within game development. Here, the depth and scope of integration vary considerably. The first studies on the topic [e.g. 21, 73] naturally integrate their findings mostly with work outside games. However, even among later studies, substantive cross-reference to other empirical studies

in game production remains rare. The most notable exceptions are Boucher et al. [5], who situate their findings within debates on authorship and labour in- and outside games, and Alharthi [2], who primarily relates to empirical studies to link production practices with broader questions of creativity and industry adoption.

We provide a more comprehensive, neutral and systematic re-integration of these individual findings with the larger and up-to-date body of qualitative research. Given the interpretative nature of meta-ethnography, the partial integrations in the primary literature above can be used as contrast and complement our synthesis.

4 Methods

This paper aims to draw a bigger picture from existing qualitative research on how GenAI has been adopted, received, and integrated within game development. To this end, we conducted a *Qualitative Research Synthesis (QRS)* and a quality appraisal of the underlying studies. While common in e.g. medical research [10], our methods are only gaining traction in HCI and games research. Here, we introduce and motivate the core methods of this study. We sketch our collaboration across all methods in Sec. 14, and provide a detailed account, reflexive procedures and positionality in Appx. B.

4.1 Qualitative Research Synthesis

Qualitative Research Synthesis (QRS) encompasses methods for systematically combining the data or interpretive findings from multiple qualitative studies to generate new knowledge and theory [16] about a phenomenon. *aggregative QRS methods* typically take a realist/pragmatist epistemology to pool and describe the original data (e.g., interviews) from large amounts of primary studies. In contrast, *interpretive methods* take an intepretivist/constructivist stance to synthesise the interpretive findings (e.g., codes developed on interviews) from often smaller sets of primary studies [16].

We conducted an *interpretive QRS* because (i) it can extend prior conceptualisation and theory; (ii) comprehensive raw data are not available in most relevant primary studies (Sec. 8); (iii) it enables critical appraisal of the strengths and weaknesses of the original research; and (iv) it makes contextual and human diversity features more apparent across studies [16, quoting Paterson, 2012] – Campbell et al. [11, p. 123] note its empowering function in involving "methods for combining multiple voices to seek new interpretations, rather than dismissing single case studies as locally bound". While most relevant to our second RQ (Sec. 1), (i)-(iii) are pivotal to addressing our third and fourth RQs .

Following common QRS terminology [e.g. 6, 47], we refer to raw data in the primary literature as "1st-order interpretations", emphasising that e.g., interview statements also constitute an individual's interpretation of their experience. "2nd-order interpretations" then denote authors' interpretations of this data through qualitative analysis, shaped by theories and individual perspectives. Finally, "3rd-order interpretations" emerge from the synthesis approach by critically comparing these second-order interpretations. We later extend this taxonomy slightly in Sec. 7.

4.2 Meta-Ethnography

We conducted a meta-ethnography, the pioneering and most widely used interpretive QRS method [16]. Developed by Noblit and Hare [47] and later extended [19], it introduces seven steps widely adopted in interpretive QRS:

(1) Selecting meta-ethnography and getting started
(2) Deciding what is relevant
(3) Reading included studies
(4) Determining how studies are related
(5) Translating studies into one another
(6) Synthesising translations
(7) Expressing the synthesis
(8) Selecting meta-ethnography and getting started
(9) Deciding what is relevant
(10) Reading included studies
(11) Determining how studies are related
(12) Translating studies into one another
(13) Synthesising translations
(14) Expressing the synthesis

Meta-ethnography enables conceptual translation across studies while preserving their contextual and epistemic integrity [16]. In step (5), we specifically employ *reciprocal translation*, relating 2nd-order interpretations across studies while noting similarities and differences [16] and supporting with 1st-order interpretations when needed.

The *synthesis* (step 6) is an inductive, iterative process in which the researcher produces 3rd-order interpretations as conceptual abstractions by (potentially) re-analysing, relating and extending the original study authors' 2nd-order interpretations, while carefully preserving their grounding in the 1st-order interpretations. Although described as separate phases, synthesis can start during translation – as was also the case here. The synthesis culminates in a *line-of-argument* narrative in which 2nd-order interpretations from different primary studies are re-narrated, structured by our new 3rd-order interpretations. The result "says something about the whole [phenomenon] based on studies of the parts" [11, p. 64] and forms, also due to its interpretive nature, a new complement to the primary studies [11, p. 121].

For precision and transparency, we report our meta-ethnography in accordance with the 19 criteria outlined in the dedicated eMERGe framework [19].

4.3 Qualitative Research Appraisal

Acknowledging the ongoing debate on the value of quality appraisal in qualitative synthesis [10, 15, 35], we incorporate formal study quality assessment. This can also contribute as interpretive aid and encourage the closer and repeated reading of primary studies [11, p. 122]. Following Majid and Vanstone's [2018] decision tool, we selected the Critical Appraisal Skills Programme (CASP) Qualitative Checklist as (1) the most common appraisal tool in QRS that is (2) domain-agnostic, (3) short, and (4) relatively easy to employ. We use Long et al.'s [35] updated and optimised version.

In line with guidance that appraisal should inform, but not dictate, interpretive decisions in meta-ethnography [10, 11], we use quality appraisal not to determine study inclusion, but to (1) support interpretive confidence of each study's findings during extraction

and translation [35] and to (2) critically reflect on methodology as a lens to pinpoint gaps in research (our fourth RQ, Sec. 1). We report this weighting alongside the references to the primary sources in the translation maps and the code book to make clear where interpretive grounding is stronger and where further work is needed.

To reconcile a study's conceptual contribution with its methodological rigour, we decided not to adopt the popular taxonomy of "Key Paper", "Satisfactory", and "Fatally Flawed" [15]. For one, the latter category might appear antagonising and hinder our goal to constructively encourage methodological rigour. Moreover, Campbell et al. [11] found that a study's methodological quality (e.g. "fatally flawed") was not always reflective of its conceptual richness (e.g. a "key paper"); similarly, Dixon-Woods et al. [15] document reviewers' dilemmas when assessing studies they found to be highly relevant despite flaws in their research conduct. Therefore, our approach separates these two judgements. Following the updated CASP framework [35], each paper is assigned a quality rating of Low, Medium, or High. Complementing this, we flag studies that function as "Key Papers" conceptually, i.e. that are "conceptually rich and could potentially make an important contribution to the synthesis" [40]. Our two-pronged appraisal system ensures that we are not overlooking conceptual innovation whilst still maintaining a clear view of the methodological rigour underpinning our synthesis.

5 Selection of Primary Studies

Searching and screening literature for a qualitative synthesis is acknowledged to be a challenging and multi-faceted process [10]. To foster transparency and reproducibility, we implement the Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension (PRISMA-S) guidelines for reporting literature searches in systematic reviews [62], thus adding detail to eMERGe steps 5–7. The PRISMA process is summarised in Figure 1.

5.1 Eligibility Criteria

Inclusion. Our eligibility criteria (Table 1) were developed iteratively and collaboratively (details in Appx. B) at the outset of the review. We included studies that report *original empirical research* using *qualitative data collection* and *qualitative analysis* methods to examine the use of *GenAI* systems within *game development* contexts. We deemed mixed-methods studies eligible if the qualitative component was substantial and clearly reported. With this, we express our focus on studies that offer insight into GenAI's practical impact over speculative or theoretical perspectives. To ensure relevance to current practices and a match of what is dominantly discussed under the umbrella of "Generative Artificial Intelligence", only studies published in the *last five years* (January 2020 - June 2025) were included, i.e. since publicly accessible GenAI tools became widely available to our demographic. Studies were considered if they reported on participants' *direct experience of game development*, either in professional roles or through situated creative production such as game jams or study assignments. We deliberately chose this wider focus as it includes people outside game industry who may eventually join the industry or contribute to game production from outside.

Exclusion. We excluded work that relied on custom-built or non-public GenAI systems, as the corresponding findings may not easily generalise. We moreover excluded work that focused on technical implementation and evaluation, as well as studies that gathered data from players, students, or other stakeholders outside of a development capacity.

Limits and restrictions. We limit our synthesis to qualitative studies in acknowledgement that GenAI is a relatively new and rapidly evolving phenomenon within the game development industry, and thus benefits from the rich, nuanced, and exploratory insights offered by qualitative inquiry.We acknowledge the value of quantitative inquiry, which can be further informed through the qualitative insights and research gaps identified here.

5.2 Search Strategy

We conducted a systematic literature search across academic databases relevant to HCI and games research, complemented with manual and open searches. We initially searched the ACM Digital Library and Scopus via Elsevier, as they index the major publication venues in HCI and technical game research. We further conducted manual searches of the DiGRA and the Foundations of Digital Games (FDG) proceedings to account for relevant work that may not be well indexed elsewhere. Finally, we searched openly on Google Scholar to identify preprints and peer-reviewed work published in other venues. Two additional measures taken to expand the identified corpus are described in Sec. 5.3.

The search terms were developed iteratively, informed by prior literature and exploratory queries. Terms were grouped into four primary blocks: (1) game development contexts, (2) GenAI systems and architectures, (3) qualitative research design, and (4) conceptual framings and interaction modalities. Searches were piloted in the ACM Digital Library and Scopus and then extended to the conference proceedings and Google Scholar. Boolean operators and wildcards were used to manage linguistic variation, and search strings were adapted to suit the syntax requirements of each platform. Full search strategies, including all query variants, are included as Supplementary Materials.

5.3 Selection Process

Individual database yields are reported in Fig. 1. For Google Scholar, the first 100 results were sorted by relevance and selected for review. After manually removing duplicates, 320 records remained. Screening and selection followed a consensus-based approach (details in Appx. B), ensuring agreement on inclusion decisions at both abstract/title screening and full-text review stages. Based on the review of titles and abstracts, we identified seven clearly eligible and 17 borderline studies that were subjected to detailed evaluation via full-text close reading with a focus on methodology and findings. This closer inspection confirmed three more papers, establishing a preliminary corpus of 10 studies. While modest in scale, the final corpus reflects the specificity of our eligibility criteria as encoded in the search query.

To ensure our primary studies pool was as comprehensive and up-to-date as possible, we took two further steps. First, we conducted backward citation searching, manually examining the bibliographies of the ten included articles for any relevant studies.

Table 1: Eligibility Criteria

Criterion	Description
Publication Period	January 2020 – June 2025.
Study Design	Reports original empirical research using qualitative or substantial mixed-methods designs.
Study Focus	Examines the application and experience of GenAI in game development contexts.
Participants	Must act in a game development capacity (e.g., professionals, game jam participants).
Scope of Technology	Publicly available GenAI systems.

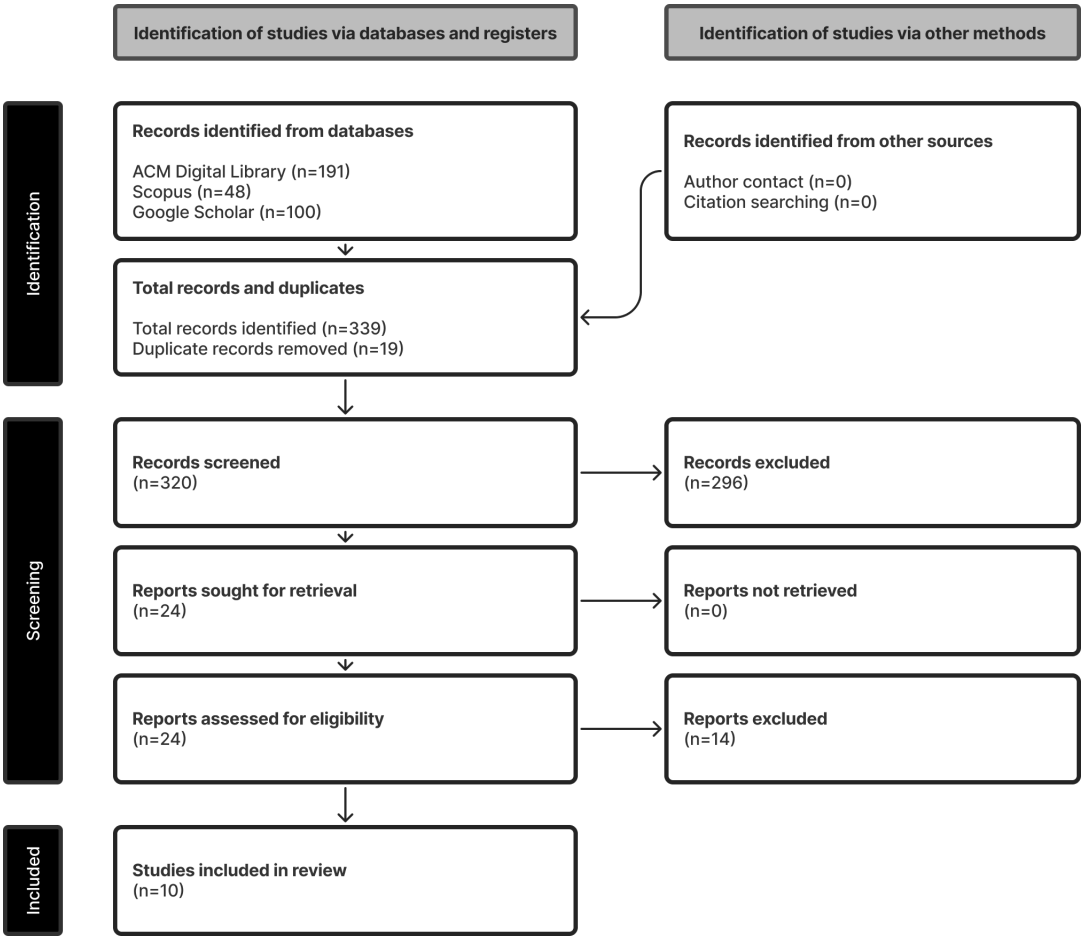


Figure 1: PRISMA Flow Diagram

Second, we contacted the first and last authors of these publications to enquire about any in-press or forthcoming work meeting our eligibility criteria. Neither step yielded further studies; while we had constructive dialogue with authors, their current research was either out of scope or still in progress when our analysis began.

6 Quality Appraisal

The qualitative research appraisal (Sec. 4.3) of each paper was done collaboratively (details in Appx. B) with the modified 11-question CASP checklist [35], and detailed justifications are provided as Supplementary Materials. Following procedure by Long et al. [35], we selected tipping-point criteria to decide on borderline weightings, namely rigour of the analysis and the clarity with which findings were grounded in the data. Tbl. 2 provides a per-criterion overview, highlighting that all studies clearly stated the

aims of research (criterion 1) and proposed an appropriate research design (criterion 3). We found the highest variation for criteria 7–9, with several studies not or only somewhat considering the researcher–participant relationship, addressing ethical issues, or performing a rigorous data analysis. The resulting appraisal weightings ("Low","Medium","High") for individual papers are presented in Tbl. 4.

7 Data Extraction

Each included study was initially subjected to a close reading by one team member to extract the synthesis materials (see Appx. B for collaboration details). Because "all [2^{nd}-order] interpretations must be grounded in the [primary] texts to be synthesised" [47], we extracted 2^{nd}-order interpretations if their grounding in 1^{st}-order interpretations was made explicit, and were inclusive where the link was implicit but clear in context. CASP ratings (Sec. 4.3) informed this judgement pragmatically: in studies appraised as higher quality, implicit grounding was more readily accepted as sufficiently evidenced, whereas in lower-rated studies we sought explicit textual support before extracting an interpretation. Extraction and subsequent translation was done in Atlas.ti 25 (Desktop), treating each primary source as separate document.

We encountered two additional challenges. Firstly, primary studies feature interpretations of similar phenomena but at varying depth (e.g. one code or a theme with subcodes). Secondly, 2^{nd}-order interpretations are given at various levels of conciseness (e.g. a code vs. an interpretive paragraph) and/or scarcity (e.g., some authors letting participants' 1^{st}-order interpretations speak largely for themselves). To address the risk of rich, readily extractable interpretations biasing the extraction of less concise, scarcer or absent 2^{nd}-order interpretations, we first engaged with studies featuring the latter. To systematically handle conceptual depth differences, we adopted a two-step approach to the extraction:

(1) An initial extraction of top-level 2^{nd}-order interpretations was conducted across all studies.
(2) We then returned to studies where an already identified 2^{nd}-order interpretation was elaborated (e.g., Boucher et al. [5] on ethical resistance) to extract their more granular sub-themes in explicit relation to the parent concept.

Our preservation of context and meaning of interpretations within and across studies followed guidance by France et al. [19]: we retained the original wording and framing of 2^{nd}-order interpretations, documented their provenance, and preserved the hierarchical links between interpretations provided by the primary studies. Where papers offered extensive 1^{st}–order but minimal researcher interpretation[e.g. 21], we conducted our own thematic analysis to generate what we term and later report as "2^{nd}-order interpretations augmented". We clearly highlight these new interpretations in the visual documentation of our reciprocal translation, also showing all extracted interpretations (Sec. 9). Extracted interpretations were reviewed collaboratively, leading to refinement or merging to enhance clarity and coherence prior to synthesis. We engaged in memo-taking to preserve these insights for the analysis to follow.

8 Description of Included Studies

We summarise the included studies across three tables: Tbl. 3 outlines each study's research questions/objectives; Tbl. 4 reports participants, data collection/analysis, duration, and CASP appraisal; and Tbl. 5 situates them by type of game considered, GenAI used, purpose, setting, scale of practice and geographic scope. To contrast adoption contexts, we split the literature according to their primary production purpose (*Purpose*, Tbl. 5). Group A ("Learning") comprises studies in learning and training environments, where GenAI is used for skill development. Group B ("Production") covers professional contexts, where GenAI serves the production of public-facing assets or complete games. Moreover, we illustrate the study collection intervals together with key GenAI milestones in Fig. 2.

8.1 Group A: Learning

These four studies show how educational settings offer diverse perspectives on GenAI for game production. Boucher et al. [5] studied a US university summer programme simulating commercial game development with small, diverse intern teams. Its industry-readiness focus fostered debates on authorship and ethics alongside tool use, framed by the authors as resistance where scepticism toward GenAI intertwined with questions of identity and ethics in emerging professional cultures. In a UK classroom, French et al. [20] examined early integration of LLM and TTIG tools into individual game projects. Unlike the collaborative, industry-simulated setting of Boucher et al., this study highlighted how students negotiated technical limitations, with expectations defined more by tool deficiencies than transformative potential. Majgaard [37] provide a comparable classroom approach, studying Danish engineering students building serious games in small teams. Here, GenAI was treated as an additional teammate whose output required constant verification. Shifting toward a production-oriented model, Shields et al. [67] documented a US serious games project where students and academics develop an educational visual novel. Like Boucher et al., the work simulated professional production but within academic constraints, with managing GenAI assets proving as time-consuming as traditional art.

8.2 Group B: Production

These studies span a range of professionals settings, from cross-regional industry surveys to focused case studies, time-limited events, and large-scale analyses of community discourse. Alharthi conducted a mixed-methods study with professionals from the MENA region, Europe, North America, and Asia. A survey identified ideation, asset creation, and programming as primary use-cases. Follow-up interviews revealed both enthusiasm and concern in terms of potential efficiency gains and erosion of creative authenticity when integrating GenAI into established workflows. Where Alharthi [2] mapped practices across multiple studios, Begemann and Hutson [4] turned inward to their own ten-day project. Acting as developers, their diary study focuses on the integration of TTIG and 3D generation tools, highlighting a persistent gap in the production of high-quality 2D concept art and usable 3D assets. Studying a ChatGPT-based game jam, Grow and Khosmood [21] offer an even more time-constrained perspective. In contrast to

Table 2: Summary of CASP appraisal of included studies (N = 10). Note: CASP scores reflect independent criteria, i.e. a study may receive a "Yes" for design appropriateness even if its later execution (e.g., data collection or analysis) was flawed.

CASP Criterion	Yes	Somewhat	No	Can't Tell
1. Was there a clear statement of the aims of the research?	10	0	0	0
2. Is a qualitative methodology appropriate?	7	3	0	0
3. Was the research design appropriate to address the aims?	10	0	0	0
4. Are the study's theoretical underpinnings clear and coherent?	9	1	0	0
5. Was the recruitment strategy appropriate to the aims?	4	1	0	5
6. Was the data collected in a way that addressed the research issue?	8	2	0	0
7. Was the researcher–participant relationship adequately considered?	2	1	4	3
8. Have ethical issues been taken into consideration?	5	2	1	2
9. Was the data analysis sufficiently rigorous?	4	3	3	0
10. Is there a clear statement of findings?	9	1	0	0
11. How valuable is the research?	10	0	0	0

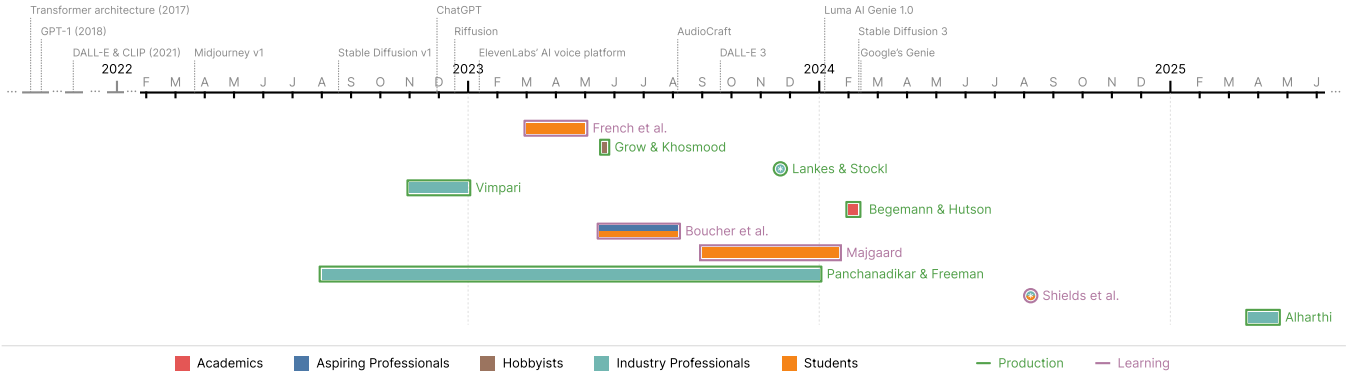


Figure 2: Timeline with key GenAI release dates and data collection intervals for the ten included studies. Studies are colour-coded for demographic focus, and outlines distinguish the study context (green for production; purple for learning). Detailed data for [2] and Begemann and Hutson [4] were obtained via direct correspondence. Data for Boucher et al. [5] and Majgaard [37] were supplemented with information from public online sources. For Panchanadikar and Freeman [52], the interval reflects the date range of collected social media posts. It was not possible to determine the data collection period for Lankes and Stockl [33] and Shields et al. [67] and these are marked in a circular shape.

previous pipelines, this context prioritised rapid prototyping and improvisation, allowing participants to explore unconventional uses of GenAI. Post-event reflections revealed benefits of faster idea generation, but also a reduced sense of authorship. Lankes and Stockl [33] explored similar uses of ChatGPT in longer-term indie studio practice. Interviews with three Austrian experts on design exercises reveal that, while GenAI supported initial ideation, the outputs were not readily transferable to production without extensive refinement. While Lankes and Stockl focused on a small number of experts, Panchanadikar and Freeman [52] scaled outward to examine more than 3000 posts from international online developer communities. Discussions often portrayed GenAI as a co-worker that could accelerate solo workflows, particularly for non-specialist tasks, while also producing errors that required close supervision. Concerns were also raised about asset quality, copyright issues, as well as over-reliance on the technology. The study further captured how community norms shaped this discourse, with enthusiasm for innovation tempered by peer-to-peer caution.

The first study of GenAI in a specific industry, Vimpari et al. [73] offered yet another perspective through interviewing professionals from Finnish game studios of different size on their use of TTIG. Their study reveals both optimism about the technology's creative potential as well as unease over ethical challenges and commercial pressures to adopt new tools. The authors note the pragmatic framing of tensions, attempting to balance artistic ambitions with the realities of market competitiveness.

9 Translation

Procedure. Similarly to Campbell et al. [11, pp. 55-57], we (a) applied reciprocal translation in two stages (first within, and then across the Learning and Production groups) and (b) adopted a chronological procedure with an "index" paper: using the oldest primary study as the index against which other were compared [11, p. 47, 65], we proceeded in publication order, adding one paper at a time and iteratively updating the synthesis. In the rest of the paper,

Table 3: Overview of the included studies (in chronological order) as well as their research questions or objectives. Quotation marks denote author-stated content; unquoted items are inferred from the paper. Further details are provided in subsequent tables.

Study	Research Questions / Objectives
[20] French et al., 2023 *Creative use of OpenAI in Education*	• "Describe and evaluate our experiences using AI tools with BSc Games Programming undergraduates as part of their coursework" • "Provide [students] with opportunities for focused engagement with OpenAI that would enhance their technical and problem-solving skills, as well as refine their abilities to analyse the current capability and potential of the technology, based on their own experiences" • "Enhance communication skills in relation to AI technologies, moving away from social-media-framed hype and towards a more rigorous, well-informed perspective" • "Reduce students' fears around AI replacing them in the future workplace"
[21] Grow & Khosmood, 2023 *ChatGPT Game Jam*	• Explore the potential of LLMs in game development by designing, convening, and evaluating a small ChatGPT Game Jam. • Provide insights about the current capabilities and shortcomings of LLM-based game production.
[33] Lankes & Stockl, 2023 *AI-Powered Game Design*	• Investigate whether current AI chatbots can effectively support designers in specific design tasks, focusing on the professional perspective of game designer • Shed more light on the potential of AI chatbots like ChatGPT to support game design processes in general
[73] Vimpari et al., 2023 *Adapt or Die: TTIG in Game Dev*	• "What are professionals' perceptions and attitudes towards TTIG systems and their future?" • "How are TTIG systems adopted and used in the creative practice now?" • "How will TTIG systems change and be used in future creative practice?"
[4] Begemann & Hutson, 2024 *Empirical Insights into AI-Assisted Game Dev*	• Explore the potential of generative AI tools in revolutionising game development pipelines by alleviating current limitations through a synergistic approach. • Illuminate the pathways through which AI can streamline and enrich the game development process, contributing to a broader discourse on the synergies between AI technologies and creative industries.
[5] Boucher et al., 2024 *Early Career Game Devs & Gen AI*	• "How, if at all, are (early game devs) using GAI in their development process?" • "What harms or benefits do they identify in using this technology?" • "How does their position as emerging professionals relate to their perceptions and usage of this technology?"
[37] Majgaard, 2024 *A Pilot Study: Engineering Students use GenAI*	• "How can GenAI be used in the development of game-like applications for educational use?" • "How can we promote students' reflections on GenAI?"
[52] Panchanadikar & Freeman, 2024 *I'm a Solo Developer Indie Devs Online*	• "What are the perceived novel opportunities and urgent risks of generative AI for indie game developers' efforts to innovate game development and production?" • "How can we design future generative AI technologies to enhance such opportunities and mitigate risks to better support these developers' efforts?"
[67] Shields et al., 2024 *Generating Together: Educational Visual Novel*	• Understand how new generative AI technologies might integrate within video game development to support narrative and artistic productivity • Describe lessons learned from attempting to integrate LLMs and text-to-image models into the development of an educational visual novel
[2] Alharthi, 2025 *Generative AI in Game Design*	• "How do game designers and developers perceive the value and potential of generative AI tools in their workflows?" • "How do generative AI tools influence creativity, productivity, and efficiency in game design and development?" • "What are the primary challenges and concerns game designers and developers face when using generative AI?"

we use "scope" to denote the level or focus of a 2nd-order interpretation (e.g. whether a finding refers to asset inclusion, pipeline efficiency, or adoption attitudes). Interpretations were related using a custom vocabulary of connectors in Atlas.ti:

Translates Conceptually equivalent at the same scope.
Supports Adds further first-order interpretations of the same type.
Enriches Introduces a new dimension to an existing concept.
Contradicts Opposes another interpretation at the same scope.
Originally part of Retains an author-defined interpretive hierarchy from the primary source.

Fig. 3 illustrates their use. The first four relationships connect nodes from different primary studies, and the last enables us to distinguish original hierarchies from those developed through the synthesis process. Relations were assigned collaboratively, with details provided in Appx. B. Once each group's network was internally connected, they were brought together for cross-group translation using the first four types of annotators. Where the scope was ambiguous, we returned to the primary texts to verify the interpretation, giving greater scrutiny to lower-rated studies (Tbl. 2). The final integrated network retained all relation types, with group-specific and cross-group structures remaining visible for subsequent synthesis (Supplementary Materials).

Table 4: Overview of participant numbers (N) and demographics, methods and duration of data collection in months, days, weeks or not stated (—), methods of data analysis, and CASP appraisal weighting – Low, Medium, orHigh.

			Collection			
Study	Demographic	N	Methods	Duration	Analysis Methods	CASP
[20]	Students	5	Case studies	9 Months	Case study discussion	Low
[21]	Hobbyists	9, 2*	Online survey	4 Days	Case study discussion	Low
[33]	Industry Professionals	3	Task-based assessment using ChatGPT; semi-structured interviews	—	Qualitative content analysis, inductive	High
[73]	Industry Professionals	14	Semi-structured interviews; online pre-interview survey	2 Months	Template analysis, inductive	High
[4]	Academics	2	Diary/Logbook format	10 Days	Statistical analysis	Medium
[5]	Students, Aspiring Professionals	26	Semi-structured interviews; embedded researcher observation	11 Weeks	Thematic analysis, inductive & deductive	High
[37]	Students	36	Student reports, classroom log notes	20 Weeks	"Grounded Theory" (?)	Medium
[52]	Industry Professionals	3091*	Online post mining	17 Months	Thematic analysis, inductive	High
[67]	Industry Professionals	9	Postmortem case study	—	Reflective analysis as case studies	Low
[2]	Industry Professionals	42 / 9 *	Online survey; semi-structured interviews	—	Thematic analysis, inductive	Medium

* [21]: 9 pre-survey and 2 post-survey ($\subseteq$ 9); [52]: 3091 online posts; [2]: 42 online survey and 9 interview ($\subseteq$ 42).

Table 5: Overview of study foci, identifying (1) type of game; (2) type of GenAI used: Large Language Models (LLM), Text-to-Image Generation (TTIG), Image-to-3D (I23D) and Text-to-3D (T23D), Audio and Speech Generation (Aud. / Spe.); (3) primary production purpose; (4) setting; (5) scale of practice: solo, 2 to 11 people (S, small), 12 - 50 (M, mid-size); more than 50 (L, large); and (6) geographic scope. Studies focused on a particular model are marked accordingly: C (LLM column) for ChatGPT and D (TTIG column) for DALL·E.

		Type of GAI							
Study	Game Type	LLM	TTIG	T23D / I23D	Aud. / Spe.	Purpose	Settting	Scale	Geographic Scope
[20]	Entertainment	C	D			Learning	Classroom	Solo	UK
[21]	Entertainment	C				Production	Event Based	Solo	USA
[33]	Mixed	C				Production	Industry	S	Austria
[73]	Entertainment		×			Production	Industry	S, M, L	Finland
[4]	Entertainment		×	×		Production	Research	S	USA
[5]	Entertainment	×	×			Learning	Industry	S	USA
[37]	Serious	×	×		×	Learning	Classroom	S	Denmark
[52]	Mixed	×	×		×	Production	Industry	Solo, S	Global
[67]	Serious	×	×			Learning	Research	S	USA
[2]	Mixed	×	×		×	Production	Industry	S, M	MENA, Europe, NA, Asia

Outcome. Relating the studies yielded a dense *Production* network (79 unique interpretations, 105 links), a more compact *Learning* network (37 unique interpretations, 35 links), and a cross-group layer (24 links). From here onwards (incl. figures), colour-code each of these primary groups as exemplified above. Across all three spaces, most relations either *Enrich* or *Support* an existing interpretation rather than asserting strict equivalence (*Translates*). In *Production* over half of the links (54%) *enrich* the linked interpretation with new conditions or mechanisms, while 18% *support* the same claim with further 1st-order interpretations. Only 12% *translate* directly, and *contradictions* are very rare (1.9%). In *Learning*, the pattern shifts slightly towards corroboration, with *Supports* (49%) roughly matching *Enriches* (43%) and a similar scarcity of *Translates* and *Contradicts*. This scarcity is not surprising, given the overall

modest number of primary studies and the HCI research tendency not to replicate but complement existing work.

Similarly, cross-environment links are dominated by *Enriches* (83%), with very few *Translates* and *Contradicts*. This pattern reflects that when *Learning* and *Production* connect, they add contextual detail but, as of now, do not identify fully equivalent phenomena across domains. Where contradictions occur, they often mark boundaries in framing, e.g. *"LS23 - GAI as colleague"* vs. *"BC23 - AI as Tool"* reflecting differences in role conceptualisation between environments, which is expanded on in our Discussion (12).

In the resulting relational map, Each node retained the verbatim 2nd-order interpretation from its source, linked to the underlying 1st-order data in Atlas.ti for direct traceability. Fig. 3 shows a close-up of the "ideation support" translation cluster, illustrating how the relational patterns identified above play out in practice. Links to *"AH25 - Prototyping and Efficiency"* and *"VG23 - Systems strengths & weaknesses - inspiration"* enrich the core idea by adding mechanisms and context-specific detailing, whilst links such as *"MJ24 - Brainstorm"* support the claim by providing concrete examples from the classroom. Contributions to the capsule come from 8 primary sources (LS23, GK23, etc.).

10 Synthesis Procedure

The cross-group relational map from reciprocal translation (Sec. 9) forms the basis for our synthesis. Clusters were used as the primary organisational unit, and *Translates* and *Contradicts* relations formed our comparative baseline, establishing where studies aligned or diverged before considering how interpretations were extended or substantiated. For each connection, we re-read the primary study extracts and recorded memos on scope, framing, and (dis-)agreement. We then examined *Enriches* connections to understand how core interpretations were extended, re-contextualised, or conditionally bounded, and *Supports* to locate where arguments gained density through 1st-order evidence.

We then turned these observations into provisional narratives for each cluster, noting where interpretations bridged clusters or sat at their margins. Where an interpretation connected to multiple clusters (different colours in synthesis map), allocation was based on contextual fit, with decisions reviewed collaboratively (details in Appx. B). This process yielded nine *synthesis capsules*, representing our nascent 3rd-order interpretations, each corresponding to one or a merger of several clusters. Each capsule brings together the interpretations, conditions, corroborations, and scope limits of one phenomenon. Short identifiers (e.g., *C1.2*, *C8.3*) for each underlying interpretation will be used consistently with the colour-coding in the remaining text to indicate where individual 2nd-order interpretations sit within the synthesis, and to maintain linkage with the 1st-order evidence. The resulting "synthesis map", an update of the original "translation map", is available in the Supplementary Materials. Capsule labels and narratives express our 3rd-order interpretations, which were then refined into the following line-of-argument synthesis, linking the nine clusters into a coherent account of how GenAI's impact has been studied across learning and production settings.

11 Outcome of Synthesis

We complement a *line-of-argument synthesis* (Sec. 4) with brief paragraphs of *implications and recommendation*. The structure is provided by our 3rd-order interpretations, corresponding to the nine identified capsules (**C1**-**C9**) that each capture a distinct but connected aspect of GenAI's impact. These span production-level practices, socio-technical positioning, adoption dynamics, governance, and authorship consequences, presented in an order that surfaces both their internal logic and their cross-linkages. Parenthetical code references (e.g., *C1.2*, *C8.3*) point to the respective capsule (e.g. **C1**, **C8**) and individual interpretations extracted previously; colours follow the evidence-group scheme introduced in the translation stage (Sec. 9). Via the full codebook (Supplementary Materials), each interpretation can be traced back to its primary study and CASP score. Interpretive decisions during the synthesis were cross-checked collaboratively (details in Appx. B), with alternative framings compared against and grounded in the primary studies.

11.1 C1: Human-in-the-loop refinement as the production norm

Synthesis. Generative outputs operate as provisional artefacts which require expert intervention before they can be accepted for production. The studies reviewed are consistent in this observation, and that correction is integral to the work rather than a marginal contingency (*C1.1*, *C1.2*, C1.3). In professional practice, both indie developers and those from bigger studios describe iterative prompting as a central mechanism for recovering creative intent and reconciling the model's tendencies with the stylistic and functional requirements of the project (*C1.4*, *C1.5*). Studies in educational settings convey a similar stance, albeit with different objectives: LLMs are positioned as programming tutors and scaffolds rather than autonomous generators, and in-engine runtime code generation during development is constrained to relatively simple tasks (*C1.6*, *C1.14*, *C1.15*). Students' accounts of routinely editing outputs before they can be applied further highlight the normality of such intervention (*C1.16*). Early stage ideation and its benefits are further discussed in **C2**. In code-based tasks, practitioners engage in iterative exchanges with the system in which error messages are used diagnostically to structure subsequent prompts, allowing gradual convergence on a working solution (*C1.7*). In image generation, stylistic stability is sensitive to asset class, with backgrounds proving more variable than character designs, and acceptable results are more often the outcome of selective curation across multiple outputs than of single, high-fidelity generations (*C1.8*, *C1.9*). Even when outputs are close to acceptable, preparatory work such as training and asset conditioning remains time-intensive and requires expertise (*C1.10*). The need for refinement is even stronger in 3D production, where unedited outputs are rarely suitable for direct inclusion due to technical and stylistic tensions with project constraints (*C1.11*). How these constraints appear at engine boundaries is detailed in **C4**, and the organisational controls that stabilise inclusion are set out in **C8**. At this juncture, human editorial decision-making determines which variants anchor subsequent work, which deviations from intent are acceptable, and which corrective strategies preserve alignment with the project's overall requirements (*C1.12*, *C1.13*). The data thus repositions prompting from the simple

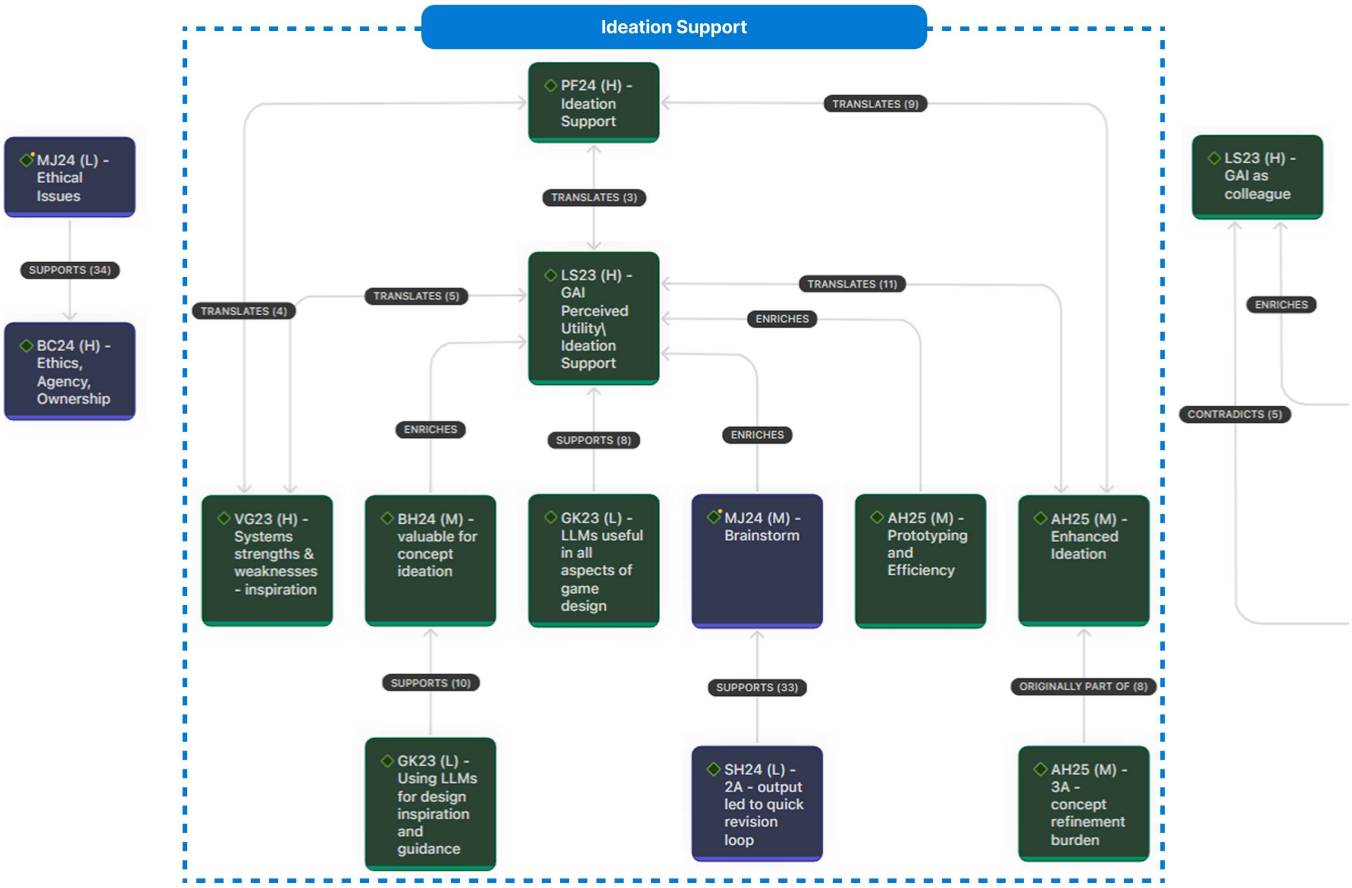


Figure 3: Exemplary translation cluster from the full synthesis map (Supplementary Materials) with the "ideation support" interpretation at the top. The hues signal that most interpretations stem from the *Production* group with only two contributing from the *Learning*.

act of query formulation to a form of progressive specification work in which the developer controls the range of acceptable solutions while maintaining coherence across the project. No primary evidence suggests that GenAI can produce inclusion-ready artefacts without human revision. Differences between studies concern the efficiency of refinement loops (**C3**) and the extent of adoption (**C6**), but these operate at distinct analytical scopes and do not contradict the claim that human intervention remains necessary at the point of inclusion.

Implications and Recommendations. To capture actual working practices, GenAI evaluation should extend from static measures of output quality to capturing characteristics of the refinement process, including time-to-acceptable-edit, the latency with which errors become visible, and the degree of control available in resolving them (*C1.7*–*C1.11*). Reporting should be disaggregated by asset class and pipeline stage to make evident where acceptance rates are higher or lower, and to inform targeted technical and organisational interventions (*C1.8*–*C1.11*). Tool design should support accountable refinement by preserving edit histories, enabling side-by-side comparison of variants, and recording the rationale for inclusion decisions (*C1.12*, *C1.13*). In education, assessments might place greater value on refinement competence (i.e. the capacity to diagnose, steer, and integrate outputs) rather than overemphasising the quality of first-attempt generation (*C1.6*, *C1.14*–*C1.16*). While the evidence base is strongest for code and 2D imagery, findings for 3D outputs, although convergent, are sparse and corresponding claims should be interpreted as indicative (*C1.11*).

11.2 C2: Early-stage ideation scaffold, not autonomous authorship

Synthesis. Across all settings, the primary value of GenAI in game development lies in their contribution to ideation rather than in any capacity for autonomous authorship, described as routine ideation support (*C2.1*, *C2.2*) or enhanced ideation (*C2.3*). Studio reflections further reinforce this premise, emphasising the utility of generators for broadening the range of ideas prior to human selection (*C2.4*). In game jams, LLMs are used explicitly for design inspiration and problem-solving guidance (*C2.5*). Ethnographic analyses offer greater specificity about where generative inspiration is reliable and how it is enacted. Users articulate system strengths

and weaknesses in relation to inspiration (C2.9), and engage in exploratory, playful routines within an established creative process (C2.7, C2.8). Educational settings provide convergent evidence, showing that GenAI can produce narrative variety that expands the range of options available for curation (C2.6). These accounts suggest that ideation is most effective under prototyping regimes where speed, disposability, and low-fidelity exploration are prioritised (C2.10). The productivity of such early-stage work is enhanced by explicit role design, with systems positioned as creative assistants within collaborative workflows that maintain clarity about human and machine responsibilities (C2.11, C2.12). In classroom settings, this is complemented by reports of brainstorming and knowledge gains that accelerate team progress (C2.13, C2.14), as well as explicit role allocations that preserve human judgement in key creative decisions (C2.15). A single counterpoint in the corpus tempers any tendency to overgeneralise from ideation capacity to autonomous authorship. Arguably little evidence from jam practice shows that producing a complete game from a single prompt remains difficult (C2.16). This reinforces that GenAI's ideation function is bounded: human framing, evaluation, and integration are constitutive of the process.

Implications and Recommendations. The implication for both research and practice is that evaluation criteria should be tailored to the ideation role. Systems should be assessed on the quality and diversity of options surfaced, their ability to support constraint-aware suggestion, and the degree to which they enable clarity in partner and workflow roles. Analyses should avoid conflating early-stage variety with later production efficiency; where efficiency does occur, it is likely to be context-specific and is examined in **C3**. These ideation outputs are provisional and require subsequent refinement for inclusion; the conditions and costs of that refinement are addressed in **C1**.

11.3 C3: Efficiency claims contested

Synthesis. Across the corpus, efficiency is frequently asserted, yet the evidence shows it to be contingent upon development phase, asset class, and practitioner expertise. Expert accounts describe time saving in everyday practice (C3.1), and review-level analyses identifies resource efficiency and faster completion for bounded activities (C3.2, C3.3). Classroom reports note faster progress but only in specific exercises (C3.5), while solo developers report advantages of focused, cost-effective development (C3.11). Many of these reports arise in early stages, as noted in **C2**. Studio reflections place the value of acceleration in rapid prototyping and concept ideation, but only under conditions of sufficient AI expertise to make use of the system productively (C3.9, C3.10). Review evidence clarifies that such advantages apply to specific tasks rather than to any wholesale acceleration of the development pipeline (C3.2, C3.3). In this reading, efficiency emerges as a conditional and situated property that depends on the alignment between the scope of the task, the model's affordances, and the practitioner's capability to configure, prompt, and integrate outputs effectively. Realised gains also depend on integration fit and organisation controls, which are detailed in **C4** and **C8**. The corpus also documents clear limits to efficiency claims: art teams report that image generation did not accelerate the production of background assets at the point of inclusion (C3.6), and model preparation remains time-consuming even where generation is rapid (C3.7). In controlled classroom exercises, manual editing proved faster than generative assistance when the specification was tight (C3.8). At the scale of individual labour, configuration, prompting, and verification overheads can erode net benefits, reframing some time-saving claims as speculative investments that do not consistently pay off (C3.12). These findings support a qualified conclusion in which efficiency is not a general property of GenAI in game development but a phase and context dependent outcome. Gains are most likely in early-stage work where disposability and rapid iteration are valued, and least likely in downstream asset inclusion and preparation where refinement and integration work, discussed further in **C1**, reabsorb much of the time apparently saved upstream.

Implications and Recommendations. Evaluation should be conducted at the scale of discrete activities rather than entire pipelines, with reporting that pairs elapsed time with rework avoided and verification burden. Expertise prerequisites should be made explicit, and efficiencies in prototyping should not be generalised to production contexts that are contradicted by evidence from art and preparation stages. Evidence is strongest for prototyping and educational micro-tasks, while findings from downstream contexts are more consistently neutral or negative. Measured effects are further shaped by pipeline constraints and governance arrangements, treated in **C4** and **C8**.

11.4 C4: Pipeline, integration, and artefact constraints

Synthesis. Across settings, present tools are reported to misalign with game-production pipelines, which require packaged, versioned artefacts, traceable provenance, and engine-conformant formats. Weak evidence from a game jam setting suggests that model outputs are optimised for text interaction rather than asset integration, producing a modality mismatch that necessitates adapters, post-processing, and human evaluation before inclusion (C4.1). Classroom settings support such affordance gaps more strongly, e.g. highlighting that debugging via prompts is difficult when errors need targeted inspection within an engine context (C4.2). Similarly, the absence of standardised transparency support creates barriers for straightforward importing of outputs into art pipelines (C4.3). Unity coding with ChatGPT still presupposes prior programming competence, which sustains reliance on existing expertise rather than enabling turnkey integration (C4.10). Unity runtime code generation has been found useful only for a narrow set of 3D primitives (C4.11). Sensitivities to labelling and training data, and to inference parameters, surface as practical hazards. For example, higher values for the LLM temperature parameter produce more diversity but also nonsensical output that requires detection and curation (C4.12–C4.13). Jam participants add that generated code comments and rationales are often not optimal for explanation or hand-off, which would limit integration especially in bigger projects (C4.14). Students also describe hallucination as a routine risk to be managed (C4.15). In art pipelines, stability is contingent on asset class, with background images exhibiting greater instability than character assets and therefore demanding more corrective work prior to import (C4.4). Studio reflections consolidate these observations

into process-level requirements. Teams report that current technology requires refinement to meet industry standards, that tool selection materially affects output quality, and that the generation of 3D models remains challenging for production purposes (*C4.5*, *C4.8*–*C4.9*). The studios respond by formalising integration as a designed and testable process: dedicated evaluation methods are required at integration gates to institutionalise model choice, acceptance thresholds, and regression checks over time (*C4.6*). Usability and integration dependencies are identified as system conditions that mediate realised value (*C4.7*). The surrounding infrastructure provides the organisational means to manage misfits and variation, as explored in **C8**. No data supports frictionless, end-to-end integration. Apparent successes in practice remain local and depend precisely on the adapters, expertise, and evaluation infrastructures documented above, which means they do not refute the underlying constraint. Thus, without structured post-processing, explicit acceptance criteria, and accountable checks, generated materials fail to satisfy pipeline requirements across both code and art contexts (C4.1–*C4.10*, C4.13). Claims about time saved should be read with the task-level analysis in **C3** in mind.

Implications and Recommendations. We recommend to treat integration as a first-class design problem. Specifications should state asset-class acceptance criteria and error budgets; adapters for compositing and import, including transparency handling and structured outputs, should be engineered as part of the workflow rather than improvised ad hoc. Debugging affordances need to extend beyond dialogue, with facilities for inspection, tracing, and deterministic reproduction within the engine. Integration gates should be supported by evaluation harnesses and provenance capture at handoff, so that model updates and parameter changes are auditable. Progress ought to be reported as reductions in post-processing effort, hallucination incidence, and re-integration time, rather than as claims of seamless substitution. In short, value accrues when integration work is made visible and improvable within the pipeline, not when generation alone is optimised without regard to the artefact constraints that ultimately govern inclusion (see **C8** for further discussion on governance).

11.5 C5: Socio-technical positioning: assistant, colleague, partner

Synthesis. Practitioners often describe GenAI in game development as a collaborator-like aide that augments rather than replaces human creativity. Expert developers report engaging with these systems much as they would with other colleagues, coordinating critique, exploration, and turn-taking (*C5.1*). For independent developers, the same stance appears in how they characterise the system as a "co-worker" and through collaboration routines that structure feedback cycles and sustain momentum in small teams (*C5.2*, *C5.3*). Across the entire industry, we can generalises this pattern as creative augmentation, making explicit that the function of the technology is to enhance the creative process rather than to automate it (*C5.5*). In learning contexts, students position the system as a teammate, integrating it into small-group coordination and using it to support collective progress (*C5.6*). Whether this stance becomes routine depends on adoption patterns by role and setting, examined in **C6**. The durability of this collaborator stance depends on deliberate socio-technical configuration. Treating GenAI as social entities allows to prescribe persona design and conversational conventions that align model behaviour with human workflow needs (*C5.4*). These mechanisms underpin the assistantship role, enabling it to generalise across contexts while maintaining accountability to human judgement. They also help explain variation in metaphor use: where the interaction is scaffolded with social design features, colleague and partner framings become more acceptable and robust. Professional identity and craft accountability, however, set clear boundaries on personification. Industry ethnography shows that some practitioners reject colleague metaphors for GenAI, positioning the system as a tool when authorship, skill recognition, and responsibility are at stake (*C5.7*). This refutation does not challenge the augmentation premise itself; rather, it specifies the contexts in which certain social metaphors are used. GenAI as assistants forms a baseline across both industry and learning settings, with colleague and partner roles remaining contingent on the absence of other conflicts. Further concerns about legitimacy and labour protections that shape these limits are discussed in **C7**.

Implications and Recommendations. For both research and practice, the implication is to design for assistantship as the durable baseline. Systems should provide controls for persona and tone, make turn-taking and critique affordances visible, and link generative contributions to artefact-level justifications. Evaluation of collaboration features should consider how well they support human-led decision making and preserve craft accountability. Where colleague metaphors are desired, their social design needs to be explicit, as in the persona-based approaches of **C5.4**, and paired with attribution and provenance measures discussed in **C9**, as well as with protections for ownership and credit addressed in **C7**. In this way, interactional convenience is supported without blurring authorship or diminishing professional responsibility.

11.6 C6: Access, democratisation, and adoption

Synthesis. GenAI is reported to lower entry barriers to game development, widening participation while revealing uneven patterns of adoption. Independent participants describe a democratising effect, with tools motivating and enabling a broader range of people to engage in game making (*C6.1*, *C6.2*). Expert developers express optimism about the potential for adoption (*C6.7*), and classroom settings reveal detailed accounts of practical on-ramps, enumerating uses such as code generation, dialogue agents, Unity runtime code and procedural content generation, and real-time creation (*C6.18*–*C6.23*). Industry settings align these gains with perceived efficiency in prototyping regimes (*C6.9*, *C6.12*). The persistence of participation gains depends on fit and skill. Studio reflections locate sustained use where AI expertise is present and benefits accrue in concept generation (*C6.15*–*C6.16*), while asset-class boundaries (e.g. 3D models) temper early optimism (*C6.14*). Ethnographic accounts document adoption as a situated process embedded within specific communities, routed through assessments of system strengths and weaknesses, ease of use, and evolving perceptions of system capabilities and development trajectories (*C6.5*, *C6.11*, *C6.38*–*C6.39*, *C6.41*–*C6.44*). Perception shifts are shaped both by direct interaction with the tools and by the socio-economic and organisational contexts in which they are deployed. Content qualities also influence

immediate utility, with overly verbose output cited as degrading day-one usefulness (*C6.46*). Constraints on diffusion are visible in both review and classroom evidence, which describe adoption as limited or experimental rather than widespread (*C6.4*, *C6.17*). Industry ethnography records predominant scepticism in craft-intensive roles where human skill, authorship, and accountability are central to identity (*C6.8*). Concerns about originality and displacement also shape acceptance thresholds, conditioning the willingness to adopt rather than denying access outright (*C6.6*, *C6.10*). Studio cases further show that friction in integrating outputs into pipelines can slow or stall adoption, even where initial enthusiasm is high (*C6.14*–*C6.16*). The combined picture is one of conditional expansion: entry barriers fall, but sustained adoption depends on alignment with role requirements, pipeline compatibility, practitioner expertise, and evolving perceptions.

Implications and Recommendations. Democratisation should be treated as a conditional process rather than a guaranteed outcome. Analyses should distinguish between access and sustained adoption, measuring uptake at the granularity of role, asset class, and pipeline stage. On-ramps such as those identified in the uses above require targeted investment in expertise development and integration fit if they are to translate into durable practice. Perception trajectories should be monitored and addressed, with attention to content qualities such as verbosity that may erode utility. Strategies for adoption need to incorporate and address originality and labour concerns so that widened access converts into stable and legitimate participation, rather than remaining a transient increase in experimentation (see **C9** on authorship and **C7** on legitimacy and labour protections).

11.7 C7: Risks, ethics, and labour precarity

Synthesis. Accounts of opportunity are accompanied by recognition of ethical uncertainty and the exposure of creative labour. The primary studies identify ethical and legal ambiguity, concerns about originality, and the possibility of labour displacement as salient systemic issues (*C7.1*–*C7.3*). Practitioner perspectives translate these into risk categories that guide human oversight and the establishment of authorship boundaries. In solo and independent practice, these include career growth-, personal investment-, creativity-, and intellectual ownership risk. Each of these frames potential losses in professional advancement, the erosion of authorship, the disputability of ownership and credit, and the narrowing of net benefits through sunk costs in skill acquisition and verification (*C7.4*–*C7.7*). Studies in studio settings reveal parallel concerns in the form of job security anxieties, legal uncertainties, and disputes over data ownership (*C7.8*–*C7.10*). Ethnographic evidence further specifies how such risks are enacted and sustained in practice. Anticipatory displacement appears as a lived experience, shaping community-level narratives and leading some practitioners to strategically avoid integrating tools into workflows where job loss is perceived as a credible threat (*C7.11*). In solo contexts, personal investment risk manifests in the transfer of costs from task execution to configuration, verification, and skill development (*C7.5*), which in turn determines the realisation of nominal efficiency gains (**C3**). Across settings, uncertainty about originality and attribution function as a gate on asset inclusion, with adoption dependent on whether provenance and credit can be formalised (*C7.2*, *C7.7*; see **C9** for authorship policy). The corpus contains no scope-matched evidence that contradicts the existence of ethical or labour risk. Where counterpoints appear, they operate at different scopes: augmentation framings in **C5** focus on the mechanics of collaboration without disputing the need for legitimacy measures, while adoption optimism in **C6** captures attitudes that coexist with ownership, legal, and job-security constraints. It therefore seems reasonable to say that without safeguards for authorship and role protection, the benefits of GenAI in game development remain contested and fragile.

Implications and Recommendations. Risk management should be treated as a primary design and governance concern. Provenance capture and attribution policies need to be embedded in workflows to address originality concerns (*C7.2*), alongside the codification of licensing and data-ownership practices (*C7.9*–*C7.10*). Role and asset-specific acceptance criteria should be defined to preserve the visibility of human authorship and protect the stability of junior labour (*C7.4*, *C7.8*). Anticipated personal investment costs should be budgeted explicitly, both in training provision and in evaluation time (*C7.5*). Such practices create the conditions under which technical benefits can be realised without normalising precarity or undermining the creative legitimacy of the work. These safeguards underpin the collaboration stances in **C5**, influence adoption patterns described in **C6**, and align with medium-aware authorship measures in **C9**.

11.8 C8: Governance of practice and workflow friction

Synthesis. Across settings, realised value from GenAI is mediated by explicit workflow design, evaluation criteria, and organisational infrastructure rather than generic model capability. In game jams and rapid prototyping, organisers position LLMs inside the process, with guidance that formalises when to prompt, how to iterate, and where to hand off so that outputs enter subsequent work coherently (*C8.1*). Industry evidence suggests the same conclusion at organisational scale, diagnosing immature or absent infrastructure that constrains institutional adoption even where local enthusiasm is high (*C8.2*). Studio reflections consolidate governance as a first-class requirement by specifying the responsibilities that make integration possible in practice, including methods for evaluating AI at integration gates, model selection policies, acceptance thresholds per asset class, regression checks across pipeline stages, and quality assurance hand-offs that preserve accountability (*C8.3*). Where these structures are weak or absent, professional ethnography records friction in workflow that suppresses day-to-day uptake despite perceived promise, indicating that organisational design is a condition for routine use rather than a post hoc enhancement (*C8.4*). These accounts link directly to the pipeline and artefact constraints identified earlier in C4. Governance provides the organisational means by which misfits are managed and variation is controlled. Process placement prevents unreal expectations of end-to-end automation and clarifies the social and technical roles involved in prompting, critique, selection, and inclusion. Evaluation harnesses and acceptance criteria translate abstract capability into fit-for-purpose judgements at the level of asset class and pipeline stage. Regression checks and provenance capture make model and

parameter changes auditable over time, which is necessary for stable collaboration across teams. In this framing, positive local results are not counter-examples to the need for governance; rather, they are instances where strong process design has already been installed.

Implications and Recommendations. Governance should be treated as a design object in its own right. Teams should specify role and asset-class acceptance criteria, implement evaluation harnesses and provenance capture at hand-off, and invest in integration infrastructure and team learning so that know-how is not confined to a few experts. Success should be measured not only in artefact quality and throughput but also in observed reductions in workflow friction and in the stability of inclusion decisions over time (*C8.1* - *C8.4*). In short, governance is the condition under which GenAI can be incorporated accountably and at scale, rather than an optional layer added after capability has been demonstrated.

11.9 C9: Authorship, materiality, and aesthetic consequence

Synthesis. GenAI is framed not only as a tool or collaborator but as a medium whose adoption redistributes authorship and shapes style. Industry-facing analysis frames the system as material, locating agency in craft decisions about how the medium is manipulated (*C9.1*). At the field level, review evidence identifies aesthetic flattening, in which reliance on generative outputs risks stylistic convergence that diminishes expressive distinctiveness (*C9.2*). Professional discourse situates these concerns within questions of ethics, agency, and ownership, making clear that authorship and credit become contested when AI-generated material is incorporated into final artefacts (*C9.3*). Ethnography underlines this contestation, with disagreement over whether prompting, selection, and post-editing constitute creative labour (*C9.4*). Divergence also appears across role boundaries: artists and programmers display different epistemic commitments to authorship, which shape their attribution practices and policy preferences (*C9.5*). Practitioners also articulate futures in which legitimate, medium-aware use is possible without eroding distinctiveness. "Imagining beyond resistance" describes design-oriented strategies that reconcile adoption with stylistic diversity, including provenance-aware workflows, dataset curation, and editorial constraint setting (*C9.6*). Interactional practices such as anthropomorphising the system can support collaboration by making its behaviour more predictable in dialogue, although such personification does not resolve the underlying questions of authorship or credit (*C9.7*). The combined evidence indicates that treating GenAI as a medium directly affects attribution norms and aesthetic outcomes. The dataset contains no similarly scoped evidence that denies the link between medium and authorship or the reality of style effects. Points of tension occur in other domains: collaboration framings in **C5** address the interpersonal stance adopted towards the system, while risk and labour discussions in **C7** address legitimacy and protection. In both cases, these do not contradict the central claim that medium choice and use conditions directly impact creative credit and stylistic consequence.

Implications and Recommendations. Authorship policies must be medium-aware. Definitions of creative contribution in AI-mediated pipelines should be explicit, specifying thresholds for authorship claims across roles and artefact types, as well as mandating provenance capture and attribution procedures (see **C7**). Evaluation of GenAI should extend beyond throughput to include measures of expressive distinctiveness, with style-diversity and convergence diagnostics used to detect and mitigate aesthetic flattening (*C9.2*). Where AI is incorporated, medium-aware practices such as those described in *C9.6* should guide dataset selection, constraint setting, and documentation of human editorial labour. While interactional personification can support workflow fluency (*C9.7*), it must not substitute for clear, enforceable authorship and credit rules that protect the integrity of creative work. Where these policies meet pipeline realities, practical arrangements are discussed in **C4** and **C8**.

12 Discussion

We briefly summarise our findings w.r.t. our four research questions (Sec. 1), referring back to the original sections where additional summarisation would create too much redundancy.

RQ1. Our overview of included studies (Sec. 8) and the corresponding Table 3–4 inform **"how existing qualitative studies differ"**. We identified two broad study contexts: educational and professional/industry-related. Within the first, researchers have studied the exploration of GenAI as part of game development classroom exercises and professional training. The latter in contrast comprised studies of game professionals in studios from indie to AAA, but also developers not clearly affiliated with industry as present in game jams. The education-focused studies involve students or early-career interns (US/UK/Denmark) and mixed student teams (US), while the professional studies report data from broad, cross-regional industry samples and, more rarely, with a national (Finland, Austria) focus. The majority of works studies entertainment games wit more than half of all studies focused on industry professionals. The identified studies were methodologically diverse with interviews and online surveys only accounting for half the studies. Crucially though, more than half of studies show medium to strong methodological weaknesses as articulated through CASP weightings.

RQ2. It is the function of our synthesis (Sec. 11) and the corresponding map (Supplementary Materials) to identify how findings **"characterise the ways in which GenAI is adopted, used, and perceived within game development workflows"** (RQ2) in the primary literature. We do so by integrating the studies' 2nd-order interpretations into 3rd-order, interpretive constructs that examine how reported concepts qualify, extend, or sit in tension with one another and what shared patterns they jointly imply [11, 40]. We identify nine overarching themes (3rd-order interpretations) in total, spanning refinement practices, ideation and efficiency, pipeline and artefact constraints, socio-technical positioning, access and adoption, labour and legitimacy, governance, and authorship (**C1–C9**). Across **C1–C4**, the synthesis characterises the use of GenAI as human-in-the-loop and front-loaded within the workflow: prompting operates as a form of progressive specification work, generative outputs function as provisional artefacts that require expert curation and editing before inclusion, and efficiency appears as a

conditional, phase- and asset-dependent property that is most credible in upstream ideation and prototyping and least so at the point of integration, where refinement and reconciliation with engine and pipeline constraints are concentrated. It is tempting to first conceive the human-GenAI co-creative act as *task-divided creativity*, where "partners take specific roles within the co-creative process [e.g. prompting, generating], producing new concepts satisfying the requirements of one party" [26]. However, the evidence suggests that it is better understood as *alternating co-creativity*, in which the "co-creative partners take turns in creating a new concept satisfying the requirements of both parties". Here, we understand GenAI's "requirements" in the sense of generative bias or output constraints. Across **C5**, **C6** and **C8**, adoption emerges as a form of conditional democratisation rather than straightforward diffusion. Entry barriers fall and new on-ramps into game making are created, yet sustained uptake is anchored where GenAI can be cast as an assistant that fits existing craft identities, role responsibilities, asset classes and locally available integration competence, instead of displacing these outright. Read through Schön's distinction between reflection-on-action and reflection-in-action, the material reveals both modes and clarifies how such assistantship is maintained in practice. Participants' retrospective accounts of game jams, classroom projects and studio work describe how using GenAI led them to re-evaluate design and development decisions, including whether its involvement had been beneficial and how it ought to be positioned in future projects (reflection-on-action). Their descriptions of iterative prompting, diagnosis and revision in situ, by contrast, exemplify reflection-in-action as they adjust prompts, constraints and inclusion thresholds in response to the system's behaviour and to pipeline-specific breakdowns [66]. Taken together, these observations indicate, first, that reflection-in-action has become a routine part of the micro-practices through which practitioners manage generative bias and keep provisional outputs compatible with engine and artefact requirements (**C1**–**C4**); and second, that reflection-on-action underpins the longer-term calibration of when, where and for whom GenAI is considered appropriate, feeding into role-specific judgements about assistantship, adoption and governance (**C5**, **C6**, **C8**). Across **C7**–**C9**, perceptions are structured by questions of legitimacy and medium: ethical and labour risks, concerns about originality and ownership, and anxieties about stylistic convergence act as gates on acceptable practice and are negotiated through provenance and attribution arrangements, workflow-level governance and medium-aware authorship policies. At the level of RQ2, the synthesis thus portrays GenAI in contemporary game development as a provisional, assistant-like technology whose principal value lies in human-steered ideation and exploratory work, and whose longer-term integration is conditioned by pipeline fit, organisational design and contested settlements around labour, authorship and aesthetic distinctiveness.

RQ3. Our reciprocal translation (Sec. 9) and the final synthesis map (Supplementary Materials) show **"in what ways findings across studies converge, complement, or contradict one another"**. Across the corpus, findings converge on four points: (1) GenAI's present value is upstream (ideation) rather than end-to-end authorship,(2) human intervention remaining necessary for output integration, (3) efficiencies are conditional rather than general, and (4) realised benefit is dependant on pipeline fit and existing governance. They complement one another by situating these claims in different environments: classrooms and game jams illustrate how GenAI is taken up for experimentation and team creativity, while professional accounts highlight asset-specific constraints, organisational dependencies, and risks to authorship and labour. Direct contradictions are rare; when they do appear, they mark boundaries or limitations: reports of prototyping speed-ups contrasting with evidence of slowdowns during asset inclusion, or classroom framings of GenAI as a "teammate" versus professional accounts that position it as a tool when authorship or responsibility is at stake. These divergences refine rather that overturn the shared conclusions by underlining the conditions under which claims about value and use can be sustained.

Tensions. Several productive tensions emerged from our translation across settings and analytic scopes. These did not constitute refutational contradictions in the sense of Noblit and Hare, as they rarely addressed the same claim at the same analytic level. We therefore do not claim to have conducted a refutational analysis and present these tensions here in a concentrated fashion and as complement to our reciprocal translation. Overall, the tensions reveal how assumptions about GenAI shift as practices move between exploratory, educational, and professional environments. First, tensions around *efficiency* signal that speed is not a stable property of GenAI use but depends on where in the development arc the system is applied. Reports of rapid prototyping and early-stage acceleration (**C2**; *C3.1*; *C3.5*; *C3.11*) sit alongside evidence that downstream inclusion slows once refinement and debugging are required (**C1**; *C3.6*–*C3.8*; **C4**). While these findings do not contradict each other directly, they underline that efficiency is phase-dependent. What appears as acceleration in exploratory phases becomes mitigated by downstream integration costs, thus creating a boundary between speed of ideation and production throughput. Second, we find tensions in the *social positioning* of GenAI. Classroom and game jam settings describe the technology as a teammate or partner supporting their collective momentum (*C5.6*; *C2.13*–*C2.15*), whereas several professional settings revert to a tool-based framing when authorship or craft identity are at stake (*C5.7*; *C7*). This reframing is not simply linguistic. It reveals that the social ontology of GenAI is shaped by local norms of responsibility. When learning and experimentation are the primary goals, relational framings flourish. When reputation or liability are implicated, boundaries around authorship become more guarded. Third, studies diverge in expectations of what GenAI can *legitimately produce*. Educational contexts emphasise breadth and exploratory flexibility (*C2.4*-*C2.6*), while professional accounts constrast this with the practical realities of asset preparation and integration (**C4**; *C1.8*–*C1.11*). The tension is therefore not a disagreement about capability, but a contextual difference in what counts as acceptable. In exploratory contexts, rough concepts are sufficient; in production, viability is governed by specific constraints that sharply limit which outputs can progress. Finally, contrasts in *adoption narratives* show that GenAI is interpreted through the lens of local labour conditions. Independent and student developers often characterise GenAI as a means of lowering barriers to experimentation (*C6.1*–*C6.2*; *C6.18*–*C6.23*). Industry accounts, by contrast, foreground concerns around authorship and

the precarious position of junior practitioners (*C7.4*–*C7.10*). This friction illustrates that enthusiasm and caution are not opposites but reflections of different forms of exposure to risk. For some, GenAI expands opportunities; for others, it complicates the conditions under which creative work is made accountable. Taken together, these disagreements, whilst not sufficient to perform refutational translation, offer insight into where claims about GenAI are stable and where they are contingent. They show that speed, collaboration, capability, or adoption are not intrinsinc qualities of the systems but relational outcomes that depend on the structure of the environment in which they are embedded. They also highlight opportunities for future HCI research to focus on the situated arrangements that shape what GenAI becomes in practice.

RQ4. Our comparison of studies, translation and synthesis all inform **"which conceptual, empirical, or methodological gaps remain"** as starting point for informing future research priorities. The following gaps and recommendations complement those identified in Sec. 11 and 13.

Conceptually, studies often conflate short-term access with durable adoption (**C6**), and pipeline-level efficiency with task-level speed-ups (**C3-C4**). Refinement is widely observed (**C1-C3**) but rarely theorised in its own right, making it difficult to distinguish basic prompting from the broader competence required to diagnose or steer outputs. Similarly, framings of GenAI as assistant, colleague, or tool (**C5-C7**) remain context-dependent but underspecified, making it unclear under which conditions each stance is most productive.

Empirically, many of our primary studies draw on only small groups of game makers. Even taken together, the combined sample remains very limited and poorly representative of the global game industry [27]. Most existing studies capture the early(ish)-adoption phase of this transition (Fig. 2), and we still lack a clear view of the large-scale structural changes that GenAI pipelines may introduce to game development. Related, empirical coverage is uneven: 2D imagery and code are relatively well evidenced, while 3D, animation, audio, and mixed-pipeline artefacts remain under-studied (**C1, C3, C4**). Engine-level integration (**C4**) and organisational governance (**C8**) are widely discussed but rarely examined in situ with direct observation or instrumentation. No studies so far are of longitudinal nature, which could e.g. inform insights on adoption w.r.t. retention, abandonment, or skill accumulation (**C6**). While insights into industry studies are of high relevance, these could be complemented with studies on hobbyists and game-jams which are presently rare and "underpowered". Surprisingly, studies with a focus on US industry as important game development ecosystem so far only focus on small studios or solo developers. Ideally, this should be complemented with insights from bigger studio and compared to findings from other local economies to contrast local industry practices.

Methodologically, reporting practices vary substantially, with some studies clearly separating pipeline stages and asset types, while others aggregate them, limiting comparability (**C1, C3, C4**). In addition to this, details of model choice and configurations are not consistently provided, making it difficult to trace how specific claims were grounded (**C7, C9**). Effects on ideation are usually described narratively rather than evaluated with explicit criteria. Related to the identified gaps on empirical coverage, researchers must become more transparent about who was interviewed and the exact timing of the data collection and analysis. In the case of GenAI adoption, even a few months can lead to very different findings. Game industry professionals are accustomed to highly flexible ways of working, often planning only two months ahead [29] and being extremely sensitive to timing and to trends circulating in the field. More broadly, sample sizes, durations, and units of analysis vary markedly across studies which impacts comparability (Fig. 2). This particularly held for case studies and postmortems; we consider these valuable, but only if complemented with future studies of similar design and potentially more participants. Diary studies have so far only been employed to capture the observation of solo academic designers; we highlight these as excellent vehicle to gain the longitudinal insights promoted earlier and believe that deployment with professionals and hobbyists could yield particularly valuable insights. More generally, short data collection intervals for distinct research questions in Fig. 2 illustrate the need for more longitudinal research. Our detailed CASP rating (Supplementary Materials) informs further methodological improvements. Here, we only note the arguably most prominent shortcoming: in several studies, qualitative coding was done at an insufficient standard, e.g. omitting descriptions of 2nd-order interpretations or grounding in the raw data. This makes consolidation efforts such as ours particularly hard and should be urgently avoided.

13 Contextualisation in Game Production Trends

In game development, a five-year observation window is still relatively short. We are working with a phenomenon in flux, where new tools and pipelines are tested and adopted while their impacts unfold. At the same time, AI is by no means new to the industry. From early procedural content generation to machine learning for character control, games have long experimented with automation [78]. GenAI represents a significant step forward, yet its integration into production pipelines will take time. Procedural methods (Sec. 2) illustrate this: despite decades of research and demonstrations, their use remains uneven and often inefficient within mainstream development environments.

Studying GenAI in isolation risks misinterpretation, since many of the phenomena currently discussed align with long-standing dynamics in game development. Asset production, for example, has always required varying rounds of iteration [29, 30], whether assets are produced in-house or outsourced. Generative tools may alter the tempo or form of these loops, but they may not eliminate the fundamental need for revision and reworking. Similarly, the apparent democratising potential of GenAI continues decades of progress in the accessibility of development environments. Unity's shift towards free licensing models already broadened participation and changed industry practices [27, 79], while low-code tools such as GameMaker, Bitsy, and Ren'Py have enabled game-making for those unfamiliar to programming. Yet all platforms also encode biases in the kinds of games they most easily produce [e.g. 13]. However, practitioners also often subvert these conventions by misusing or stretching tools [23]. GenAI similarly risks reproducing existing norms rather than enabling radical departures, though practitioners may also find ways to push against its constraints.

Another important dimension is that the central problems of game development are increasingly less about technology and more about human resources. Post-mortem studies have shown that difficulties in production often stem from soft skills, team coordination, and organisational dynamics rather than mere technical obstacles [54]. In this respect, GenAI may intersect with challenges of communication and collaboration as much as with hard-skills related efficiency or automation. The industry's structural precarity must also be acknowledged: long before and in parallel to GenAI, volatility was created by shifting business models and changing technologies [8, 30], keeping up precarity and uncertainty. The rise of free-to-play, for instance, forced many premium studios to adopt new pipelines and revenue strategies [27]. The recent post-pandemic wave of lay-offs continues this pattern of instability. Certain creative roles, such as concept artists, writers, or voice actors, among other roles have always been fragile. GenAI therefore enters a labour landscape already marked by insecurity, and its effects need to be considered against this broader background.

Seen in this light, GenAI appears less a radical rupture than a continuation of existing trajectories. Its use so far remains largely conservative, producing incremental variations in ideas and styles. While it can multiply the quantity of assets, characters, levels, or environments, "more" content does not necessarily translate into more engaging or meaningful experiences. Audience tastes are themselves conservative, reinforced by industry consolidation and mass-market imperatives, although counter-trends such as the rise of K-pop and other non-Hollywood media suggest that shifts in cultural consumption are always possible. The studies synthesised here highlight that a significant impact is underway, but sustainable insights on GenAI's long-term impact require more time, further study and stronger contextualisation in the broader cultural, economic, and technological dynamics of the industry.

14 Reflexivity & Study Limitations

As an interdisciplinary team working across game HCI, computational creativity, and design practice, our backgrounds inevitably informed how we interpreted the primary studies. We approached the topic from a broadly agnostic stance toward GenAI, recognising its potential value when used to complement rather than replace human creative practice, but also acknowledging potential threats, e.g to environment, individuals and society. To complement individuality in interpretation, all stages of our meta-ethnography were conducted collaboratively, with iterative comparison of alternative readings and repeated reference back to the primary sources. This process ensured that our 3rd-order interpretations remained grounded in the evidence and not in our preconceptions. We present a full account of reflexive procedures, including stage-by-stage details of our collaborative workflows and positionality statements in Appx. B.

While we emphasise the transformative impact of GenAI, it is important to recognise that several other forces are reshaping daily work and production in game industry. Hybrid work models, the current geopolitical climate, and declining consumer purchasing power all influence how studios can structure and sustain the workforce. In addition, the long-running rise of third-party tools and game engines continues to democratise production and accelerate change, and this trend predates GenAI. While we contextualise our findings in empirically grounded game production trends (Sec. 13), our synthesis does not integrate the interaction of the identified GenAI phenomena with these wider developments at the same level as within GenAI. Research and policy must take a holistic stance.

Moreover, it would have been desirable to ground our findings game development theory specifically. Crucially though, such theory is largely absent: we lack a comprehensive understanding of game development practices and the regional differences in how developers think and work [9, 27, 28]. When studying game development, it is essential to situate the data not only in time but also in place. Practices vary from project to project, but they also differ across local cultures, companies, and levels of experience [30]. The present synthesis cannot account for these differences in sufficient detail to directly support theory building, but our overview of studies (Sec. 8; 9) and identified research gaps (Sec. 12) point at specific contexts to investigate further with the appropriate methods and epistemologies to further the development of such theory. At present, we ground our findings in theory that is applicable beyond games.

The quality of our synthesis rests heavily on the quality of its underlying primary studies. The evidence base remains modest and uneven across contexts, with primary studies varying in conceptual depth and methodological quality. This asymmetry shapes the density of our translation networks and limits the precision with which comparisons can be made across the corpus. Poor 2nd-order interpretations in some studies required us to augment such interpretations to foster comparability across studies. This introduces an additional interpretative layer.

Fig. 2 illustrates that the qualitative studies underpinning our synthesis are overall modest in numbers and scattered over more than three years with typically short data collection intervals (Panchanadikar and Freeman [52] exceptionally collected social media data over a longer timeframe) and distinct research questions. This situation did not allow for our synthesis to track changes over time. Instead, it combines several points in time, each representing potentially transient phenomena. Therefore, the synthesis must not be understood as a snapshot of the present. Moreover, interactions with other developments in games (see above) matter and the present moment is unusually opaque – many practitioners describe the current landscape as "dark", shaped by overlapping and rapidly changing pressures: data regulation, declining consumer spending, AAA instability, consolidation, and intensifying global competition, among others. This uncertainty makes it difficult to forecast how GenAI pipelines will actually take hold.

On a similar note, while our translation explicitly compares different contexts, the synthesis may obfuscate that that game development is by no means a unified or standardised practice. Game products vary dramatically in complexity, scope, and form, and so do game development contexts and communities. Communities of practice (e.g. AAA, mobile and hyper-casual, arthouse, academic, hobbyist, etc.) overlap, but they also remain distinct in their norms, expectations, and structural conditions. Additionally, each sector operates under different constraints and (e.g. commercial) incentives. A consequence of abstracting from evidence drawn from

various development contexts, our synthesis highlights mutual support and tensions in identified phenomena, e.g. between production and learning environments, but we remind our audience of the diversity in practices from which our evidence was sourced.

While we hope for our findings to provide guidance for e.g. practice or policy, it is for the reasons above that we strongly recommend to not use the synthesis insights in isolation; instead, we suggest to use them as a starting point which, complemented with the identified research gaps and quality appraisal, can inform the collection of further in-the-moment data to inform definite decisions or forecast future developments. Sensitive to the decision task, this research may have to balance an holistic approach with sensitivity to e.g. diversity in practice or local characteristics.

15 Conclusion

We conducted the first qualitative research synthesis (QRS) on the impact of GenAI on videogame production. We employed PRISMA-S to systematically search the literature, identifying a corpus of 10 primary literature items to then expose to a meta-ethnography as popular QRS method, guided by the eMERGe framework and supported by CASP quality appraisal. Through reciprocal translation, line of argument synthesis and visual mapping, we identified and summarise nine overarching themes in existing qualitative work. Together with our detailed comparison of qualitative studies along various axes such as research goals, settings, demographic and methodology, we highlight gaps in research and provide recommendations for future work. Since the impact of GenAI does not happen in a vacuum, we contextualise our findings within broader trends in game production. We hope that these insights will benefit industry stakeholders, researchers and policymakers with grounded and concentrated insights to guide future practice, research and support governance.

Acknowledgments

We highlight individual contributions using the Contributor Roles Taxonomy [CRediT, 51]: Funding acquisition (CG), Conceptualisation (CG, AD, JC), Methodology (AT, CG), Investigation (AT, AD, JC, CG), Formal analysis (AT, AD, CG), Visualisation (AT, JC), Validation (AT, AD, JC, CG), Supervision (CG, AD), Writing - original draft (ALL).

Author JC was partially funded by national funds through FCT – Foundation for Science and Technology, I.P., within the scope of the research unit UID/00326; and by the Portuguese Recovery and Resilience Plan through project C645008882-00000055, Center for Responsible AI.

References

[1] Stability AI. 2022. *Stable Difusion Launch Announcement.* https://stability.ai/blog/stable-difusion-announcement

[2] Sultan Alharthi. 2025. Generative AI in Game Design: Enhancing Creativity or Constraining Innovation? *Journal of Intelligence (J. Intell.)* 13 (May 2025). doi:10.3390/jintelligence13060060

[3] Alpha3D. n.d.. Alpha3D: Transform Text and 2D Images into 3D Assets. https://alpha3d.io/. Accessed: 2025-09-12.

[4] Andrew Begemann and James Hutson. 2024. Empirical insights into AI-assisted game development: A case study on the integration of generative AI tools in creative pipelines. *Metaverse* 5, 2 (July 2024), 2568. doi:10.54517/m.v5i2.2568 Number: 2.

[5] Josiah D Boucher, Gillian Smith, and Yunus Doğan Telliel. 2024. Is Resistance Futile?: Early Career Game Developers, Generative AI, and Ethical Skepticism. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24).* Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3613904.3641889

[6] Nicky Britten, Rona Campbell, Catherine Pope, Jenny Donovan, Myfanwy Morgan, and Roisin Pill. 2002. Using meta ethnography to synthesise qualitative research: a worked example. *Journal of health services research & policy* 7, 4 (2002), 209–215.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.

[8] Ergin Bulut. 2020. *A precarious game: The illusion of dream jobs in the video game industry.* Cornell University Press.

[9] Loïc Cadin, Francis Guérin, and Robert DeFillippi. 2006. HRM Practices in the Video Game Industry:: Industry or Country Contingent? *European Management Journal* 24, 4 (2006), 288–298.

[10] Mairead Cahill, Katie Robinson, Judith Pettigrew, Rose Galvin, and Mandy Stanley. 2018. Qualitative synthesis: A guide to conducting a meta-ethnography. *British Journal of Occupational Therapy* 81, 3 (March 2018), 129–137. doi:10.1177/0308022617745016 Publisher: SAGE Publications Ltd STM.

[11] R. Campbell, P. Pound, M. Morgan, G. Daker-White, N. Britten, R. Pill, L. Yardley, C. Pope, and J. Donovan. 2011. Evaluating meta-ethnography: systematic analysis and synthesis of qualitative research. *Health Technology Assessment (Winchester, England)* 15, 43 (Dec. 2011), 1–164. doi:10.3310/hta15430

[12] Harold Cohen and Pamela McCorduck. 1991. *Aaron's code: meta-art, artificial intelligence, and the work of Harold Cohen.* W. H. Freeman & Co.

[13] Mia Consalvo and Dan Staines. 2021. Reading Ren'Py: game engine affordances and design possibilities. *Games and Culture* 16, 6 (2021), 762–778.

[14] Shuyao Dai, Yang Li, Kazjon Grace, and Anastasia Globa. 2023. Towards human-AI collaborative architectural concept design via semantic AI. In *International Conference on Computer-Aided Architectural Design Futures.* Springer, 68–82.

[15] Mary Dixon-Woods, Alex Sutton, and Rachel Shaw. 2008. Appraising qualitative research for inclusion in systematic reviews: A quantitative and qualitative comparison of three methods (Journal of Health Services Research and Policy (2007) 12, (42-47)). *Journal of Health Services Research & Policy* 13 (Jan. 2008). doi:10.1258/jhsrp.2007.000023

[16] James W Drisko. 2020. Qualitative research synthesis: An appreciative and critical introduction. *Qualitative Social Work* 19, 4 (2020), 736–753.

[17] ElevenLabs Inc. n.d.. ElevenLabs: AI Audio Research and Deployment Company. https://elevenlabs.io/. Accessed: 2025-09-12.

[18] Ahmed Elgammal, Marian Mazzone, et al. 2020. Artists, Artificial Intelligence and Machine-Based Creativity in Playform. *Artnodes* 26 (2020), 1–8.

[19] Emma F. France, Maggie Cunningham, Nicola Ring, Isabelle Uny, Edward A. S. Duncan, Ruth G. Jepson, Margaret Maxwell, Rachel J. Roberts, Ruth L. Turley, Andrew Booth, Nicky Britten, Kate Flemming, Ian Gallagher, Ruth Garside, Karin Hannes, Simon Lewin, George W. Noblit, Catherine Pope, James Thomas, Meredith Vanstone, Gina M. A. Higginbottom, and Jane Noyes. 2019. Improving reporting of meta-ethnography: the eMERGe reporting guidance. *BMC Medical Research Methodology* 19, 1 (Jan. 2019), 25. doi:10.1186/s12874-018-0600-0

[20] Fiona French, David Levi, Csaba Maczo, Aiste Simonaityte, Stefanos Triantafyllidis, and Gergo Varda. 2023. Creative Use of OpenAI in Education: Case Studies from Game Development. *Multimodal Technologies and Interaction* 7, 8 (Aug. 2023), 81. doi:10.3390/mti7080081 Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[21] April M. Grow and Foaad Khosmood. 2023. ChatGPT GameJam: Unleashing the power of Large Language Models for Game Jams. In *Proceedings of the 7th International Conference on Game Jams, Hackathons and Game Creation Events (ICGJ '23).* Association for Computing Machinery, New York, NY, USA, 51–54. doi:10.1145/3610602.3610605

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

[23] Andrew Hugill and Hongji Yang. 2013. The creative turn: new challenges for computing. *International Journal of Creative Computing* 1, 1 (2013), 4–19.

[24] Nanna Inie, Jeanette Falk, and Steve Tanimoto. 2023. Designing Participatory AI: Creative Professionals' Worries and Expectations About Generative AI. In *Extended Abstracts of the Conference on Human Factors in Computing Systems.* ACM, Article 82, 8 pages. doi:10.1145/3544549.3585657

[25] Anna Kantosalo and Tapio Takala. 2020. Five C's for Human–Computer Co-Creativity: An Update on Classical Creativity Perspectives. In *International Conference on Computational Creativity.* Association for Computational Creativity, 17–24.

[26] Anna Kantosalo and Hannu Toivonen. 2016. Modes for creative human-computer collaboration: Alternating and task-divided co-creativity. In *Proceedings of the seventh international conference on computational creativity.* 77–84.

[27] Aphra Kerr. 2017. *Global games: Production, circulation and policy in the networked era.* Routledge.

[28] Rilla Khaled and Pippin Barr. 2023. Generative Logics and Conceptual Clicks: A Case Study of the Method for Design Materialization. *Design Issues* 39, 1 (2023),

55–69.
[29] Annakaisa Kultima. 2015. Developers' perspectives on iteration in game development. In *Proceedings of the 19th International academic Mindtrek conference*. 26–32.
[30] Annakaisa Kultima. 2018. *Game design praxiology*. Ph. D. Dissertation. University of Tampere.
[31] Annakaisa Kultima. 2025. Trends in Gaming - Views of Industry Professionals and Game Scholars. In *Proceedings of the 20th International Conference on the Foundations of Digital Games (FDG '25)*. Association for Computing Machinery, New York, NY, USA, Article 10, 7 pages. doi:10.1145/3723498.3723743
[32] Ichiro Lambe. 2025. The Surprising NEW Number of GenAI Games on Steam. https://www.totallyhuman.io/blog/the-surprising-new-number-of-genai-games-on-steam. Accessed September 6, 2025.
[33] M. Lankes and A. Stockl. 2023. AI-Powered Game Design: Experts Employing ChatGPT in the Game Design Process, Vol. 24. 1–9. doi:10.55549/epstem.1406194
[34] Ruihuang Li, Caijin Zhou, Shoujian Zheng, Jianxiang Lu, Jiabin Huang, Comi Chen, Junshu Tang, Guangzheng Xu, Jiale Tao, Hongmei Wang, et al. 2025. Hunyuan-Game: Industrial-grade Intelligent Game Creation Model. *arXiv preprint arXiv:2505.14135* (2025).
[35] Hannah A Long, David P French, and Joanna M Brooks. 2020. Optimising the value of the critical appraisal skills programme (CASP) tool for quality appraisal in qualitative evidence synthesis. *Research Methods in Medicine & Health Sciences* 1, 1 (Sept. 2020), 31–42. doi:10.1177/2632084320947559 Publisher: SAGE Publications Ltd STM.
[36] Luma AI, Inc. n.d.. Luma AI. https://lumalabs.ai/. Accessed: 2025-09-12.
[37] Gunver Majgaard. 2024. A Pilot Study: Engineering Students Use Generative AI to support the development of playful educational technology. *European Conference on Games Based Learning* 18, 1 (Oct. 2024), 590–597. doi:10.34190/ecgbl.18.1.2872 Number: 1.
[38] Umair Majid and Meredith Vanstone. 2018. Appraising Qualitative Research for Evidence Syntheses: A Compendium of Quality Appraisal Tools. *Qualitative Health Research* 28, 13 (Nov. 2018), 2115–2131. doi:10.1177/1049732318785358 Publisher: SAGE Publications Inc.
[39] Mahdi Farrokhi Maleki and Richard Zhao. 2024. Procedural content generation in games: A survey with insights on emerging llm integration. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 20. 167–178.
[40] Alice Malpass, Alison Shaw, Debbie Sharp, Fiona Walter, Gene Feder, Matthew Ridd, and David Kessler. 2009. "Medication career" or "Moral career"? The two sides of managing antidepressants: A meta-ethnography of patients' experience of antidepressants. *Social Science & Medicine* 68, 1 (Jan. 2009), 154–168. doi:10.1016/j.socscimed.2008.09.068
[41] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. 2015. Generating Images From Captions With Attention. *arXiv preprint arXiv:1511.02793* (2015).
[42] Meshy LLC. n.d.. Meshy AI: Transform Text and Images into 3D Models. https://www.meshy.ai/. Accessed: 2025-09-12.
[43] Meta AI. 2023. AudioCraft: A Simple One-Stop Code Base for Generative Audio. https://ai.meta.com/resources/models-and-libraries/audiocraft/. Accessed: 2025-09-12.
[44] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. 2022. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13492–13502.
[45] Midjourney. 2022. *Midjourney*. https://www.midjourney.com/home/
[46] Jewoong Moon, Unggi Lee, Junbo Koh, Yeil Jeong, Yunseo Lee, Gyuri Byun, and Jieun Lim. 2025. Generative artificial intelligence in educational game design: Nuanced challenges, design implications, and future research. *Technology, Knowledge and Learning* 30, 1 (2025), 447–459.
[47] George W. Noblit and R. Dwight Hare. 1988. *Meta-Ethnography: Synthesizing Qualitative Studies*. SAGE Publications, Inc., Newbury Park, CA. doi:10.4135/9781412985000
[48] OpenAI. 2021. *DALL·E: Creating Images from Text*. https://openai.com/blog/dall-e/
[49] OpenAI. 2022. *DALL·E 2*. https://openai.com/dall-e-2/
[50] Jonas Oppenlaender. 2022. The Creativity of Text-to-Image Generation. In *Proc. International Academic Mindtrek Conference*. 192–202.
[51] National Information Standards Organization. 2022. *Contributor Roles Taxonomy (CRediT)*. http://credit.niso.org/
[52] Ruchi Panchanadikar and Guo Freeman. 2024. "I'm a Solo Developer but AI is My New Ill-Informed Co-Worker": Envisioning and Designing Generative AI to Support Indie Game Development. *Proceedings of the ACM on Human-Computer Interaction* 8, CHI PLAY (Oct. 2024), 1–26. doi:10.1145/3677082 Number: CHI PLAY.
[53] Jack Parker-Holder, Shlomi Fruchter, and Google DeepMind. 2025. Genie 3: A New Frontier for World Models. Blog post, Google DeepMind. https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/
[54] Cristiano Politowski, Fabio Petrillo, Gabriel C Ullmann, and Yann-Gaël Guéhéneuc. 2021. Game industry problems: An extensive analysis of the gray literature. *Information and Software Technology* 134 (2021), 106538.
[55] Producer.ai (formerly Riffusion). 2025. Producer.ai — Create the Music You Imagine. https://www.producer.ai/. Invite-only generative AI instrument for creating, remixing, and sharing studio-quality songs from prompts. Accessed: 2025-09-12.
[56] Promethean AI. n.d.. Luma AI. https://www.prometheanai.com/. Accessed: 2025-11-17.
[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.
[58] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).
[59] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation With Clip Latents. *arXiv preprint arXiv:2204.06125* (2022).
[60] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-To-Image Generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.
[61] Jay Ratican and James Hutson. 2024. Adaptive worlds: Generative AI in game design and future of gaming, and interactive media. *ISRG Journal of Arts, Humanities and Social Sciences* 2, 5 (2024).
[62] Melissa L. Rethlefsen, Shona Kirtley, Siw Waffenschmidt, Ana Patricia Ayala, David Moher, Matthew J. Page, Jonathan B. Koffel, Heather Blunt, Tara Brigham, Steven Chang, Justin Clark, Aislinn Conway, Rachel Couban, Shelley de Kock, Kelly Farrah, Paul Fehrmann, Margaret Foster, Susan A. Fowler, Julie Glanville, Elizabeth Harris, Lilian Hoffecker, Jaana Isojarvi, David Kaunelis, Hans Ket, Paul Levay, Jennifer Lyon, Jessie McGowan, M. Hassan Murad, Joey Nicholson, Virginia Pannabecker, Robin Paynter, Rachel Pinotti, Amanda Ross-White, Margaret Sampson, Tracy Shields, Adrienne Stevens, Anthea Sutton, Elizabeth Weinfurter, Kath Wright, Sarah Young, and PRISMA-S Group. 2021. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. *Systematic Reviews* 10, 1 (Jan. 2021), 39. doi:10.1186/s13643-020-01542-z
[63] Rafael Ribeiro, Alexandre Valle de Carvalho, and Nelson Bilber Rodrigues. 2024. Image-based video game asset generation and evaluation using deep learning: a systematic review of methods and applications. *IEEE Transactions on Games* (2024), 1–10. doi:10.1109/TG.2024.3487054
[64] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proc. Conference on Computer Vision and Pattern Recognition*. IEEE, 10684–10695.
[65] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. 2022. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18603–18613.
[66] Donald A Schön. 1983. *The reflective practitioner: How professionals think in action*. Basic Books.
[67] Samuel Shields, Alexander Calderwood, Shi Johnson-Bey, Noah Wardrip-Fruin, Michael Mateas, and Edward Melcer. 2024. Generating Together: Lessons Learned from Developing an Educational Visual Novel with AI Collaboration. In *2024 IEEE Conference on Games (CoG)*. 1–8. doi:10.1109/CoG60054.2024.10645553 ISSN: 2325-4289.
[68] Karl Sims. 1991. Artificial evolution for computer graphics. In *Proceedings of the 18th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1991*, James J. Thomas (Ed.). ACM, 319–328. doi:10.1145/122718.122752
[69] Soundraw. n.d.. SOUNDRAW — AI Music Generator – Royalty Free Beats. https://soundraw.io/. Accessed: 2025-09-12.
[70] Penny Sweetser. 2024. Large language models and video games: A preliminary scoping review. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. 1–8.
[71] Unity Technologies. 2023. Unity Muse: AI Platform for Accelerated Real-time 3D Creation. https://unity.com/blog/engine-platform/introducing-unity-muse-and-unity-sentis-ai. Accessed: 2025-09-12.
[72] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems* 30 (2017).
[73] Veera Vimpari, Annakaisa Kultima, Perttu Hämäläinen, and Christian Guckelsberger. 2023. "An Adapt-or-Die Type of Situation": Perception, Adoption, and Use of Text-to-Image-Generation AI by Game Industry Professionals. *Proceedings of the ACM on Human-Computer Interaction* 7, CHI PLAY (Sept. 2023), 131–164. doi:10.1145/3611025 Number: CHI PLAY.
[74] Roosa Wingström, Johanna Hautala, and Riina Lundman. 2022. Redefining Creativity in the Era of AI? Perspectives of Computer Scientists and New Media Artists. *Creativity Research Journal* (2022), 1–17.
[75] Zhenhua Wu, Zhuohao Chen, Di Zhu, Christos Mousas, and Dominic Kao. 2025. A Systematic Review of Generative AI on Game Character Creation: Applications, Challenges, and Future Trends. *IEEE Transactions on Games* (2025), 1–15. doi:10.1109/TG.2025.3564869

[76] Daijin Yang, Erica Kleinman, and Casper Harteveld. 2024. GPT for Games: A Scoping Review (2020-2023). In *2024 IEEE Conference on Games (CoG)*. 1–8. doi:10.1109/CoG60054.2024.10645548
[77] Daijin Yang, Erica Kleinman, and Casper Harteveld. 2025. GPT for Games: An Updated Scoping Review (2020-2024). *IEEE Transactions on Games* (2025), 1–16. doi:10.1109/TG.2025.3563780
[78] Georgios N Yannakakis and Julian Togelius. 2018. *Artificial intelligence and games*. Vol. 2. Springer.
[79] Chris J Young. 2021. Unity production: Capturing the everyday game maker market. *Game production studies* (2021), 141.

A Search Strategy

Search strings were constructed using Boolean operators, with terms within each block combined using OR and blocks joined using AND. Wildcards (e.g., *) were used to capture morphological variation, and quotation marks were used to ensure semantic specificity in multi-word terms.

Finalised multi-block query (general form):

```
("game dev*" OR "game creat*" OR "game prod*" OR "game design" OR "game studio*" OR "game pipeline*" OR "game jam*" OR "game hackathon*")
AND
("gen* AI" OR genAI OR "AI gen*" OR "generative machine-learning" OR "generative model*" OR "foundation model*" OR "large language model*" OR LLM* OR ChatGPT OR "text-to-image generat*" OR TTIG OR LTGM OR "transformer-based generat*" OR "transformer model*" OR "diffusion-based generat*" OR "diffusion model*" OR "Stable Diffusion" OR Midjourney OR DALL*E OR DALL-E* OR DALL·E* OR DALLE* OR "multimodal generative model*")
AND
(qualitative OR interview* OR survey* OR "focus group*" OR questionnaire* OR "case stud*" OR "thematic analysis" OR "practice-based" OR "design research" OR "research through design" OR ethnograph* OR autoethnograph* OR "grounded theory" OR "diary study")
```

Scopus query:

```
( TITLE-ABS-KEY ( "game dev*" OR "game creat*" OR "game prod*" OR "game design"
OR "game studio*" OR "game pipeline*" OR "game jam*" OR "game hackathon*" ) )
AND
( TITLE-ABS-KEY ( "gen* AI" OR "genAI" OR "AI gen*" OR "generative machine-learning"
OR "generative model*" OR "foundation model*" OR "large language model*" OR llm*
OR "ChatGPT" OR "text-to-image generat*" OR "TTIG" OR "LTGM"
OR "transformer-based generat*" OR "transformer model*"
OR "diffusion-based generat*" OR "diffusion model*" OR "Stable Diffusion"
OR "Midjourney" OR "DALLE*" OR "DALL-E*" OR "DALL E*" OR "DALL·E*"
OR "multimodal generative model*" ) )
AND
( TITLE-ABS-KEY ( qualitative OR interview* OR survey* OR "focus group*"
OR questionnaire* OR "case stud*" OR "thematic analysis" OR "practice-based"
OR "design research" OR "research through design" OR ethnograph*
OR autoethnograph* OR "grounded theory" OR "diary study" ) )
```

Comparative Query Evaluation

To assess the relevance and impact of the fourth, more abstract block (conceptual framings and interaction modalities), we compared the output of a three-block query (blocks 1–3) against the full four-block query. The comparison revealed that while the fourth block filtered out non-relevant technical papers, it occasionally excluded conceptually relevant but less explicitly framed studies. Consequently, we included results from both configurations, manually reviewing overlaps to ensure a balanced and inclusive study set that captured both technically and conceptually grounded work.

B Reflexive Positioning and Analytic Collaboration

This appendix expands on how reflexivity was embedded throughout the synthesis, including (1) an overview of collaboration across all stages of the meta-ethnography and (2) brief positionality statements from each author. It complements the brief reflexivity notes in Sec. 14 and seeks to foster transparency in interpretive qualitative synthesis.

B.1 Collaborative Analytic Process Across Meta-Ethnography Stages

Our collaborative workflow followed the seven stages of meta-ethnography (Sec. 4.2), with reflexive discussion integrated throughout. We met weekly or biweekly basis with additional asynchronous exchanges via Slack and shared documents. Author D contributed additional contextualisation in game industry practices (Sec. 1 and 13) post-synthesis.

(1) Selecting meta-ethnography and defining focus (Sec. 1;4). The initial scope, research questions, and methodological rationale were drafted by authors B, C, and E and refined in group discussions. Reflexivity centred on clarifying our assumptions about GenAI's role in creative practice and ensuring that these would not prestructure the synthesis.

(2) Determining what is relevant (Sec. 5.1; 5.2). Eligibility criteria were jointly developed at the project outset. Authors A-C and E participated in paper screening, with disagreements resolved consensually in group discussions. Agreement on inclusion was sought at both abstract/title screening and full-text review stages. Iterative reflections concerned our scoping of 'game development' and how boundaries around creative roles may influence interpretation.

(3) Reading included studies (Sec. 5.3). Author A read all included studies and authors B, C and E subsets. Notes and preliminary observations were shared in group meetings. Reflexive discussion focused on how our disciplines shaped what we each attended to in the text (e.g., workflow details, creative reasoning, organisational conditions).

Table 6: Search terms grouped by query domain for structured literature search (transposed).

Game domain	GenAI domain	Concept domain	Methodology domain
game dev*	gen* AI	creativ*	qualitative
game creat*	genAI	computational creativity	interview*
game prod*	AI gen*	creative AI	case stud*
game design	generative machine-learning	AI creativity	thematic analysis
game studio*	generative model*	AI for creativ*	practice-based
game pipeline*	foundation model*	collaborat*	design research
game jam*	large language model*	co-creat*	research through design
game hackathon*	LLM*	mixed-initiative*	ethnograph*
	ChatGPT	ideat*	autoethnograph*
	text-to-image generat*	feedback loop*	grounded theory
	TTIG	enhanced creativ*	diary study
	LTGM		
	transformer-based generat*		
	transformer model*		
	diffusion-based generat*		
	diffusion model*		
	Stable Diffusion		
	Midjourney		
	DALL·E (all variants)		
	multimodal generative model*		

(+) Qualitative Research Appraisal (Sec. 6). The appraisal of each paper was done collaboratively between authors A and B on a shared spreadsheet with the modified 11-question CASP checklist [35]. The appraisers rated each criterion and provided a detailed justification resting on evidence from the source material (Supplementary Materials). Reflexivity concerned the tipping-point criteria and fair treatment of studies with different goals and methodology.

(4) Determining how studies are related (Sec. 7). The relationships between concepts were mapped collaboratively. Authors A and E extracted the interpretations from the primary sources and drafted initial affinity structures, which were then reviewed and reworked jointly with authors B and C. Regular meetings were used to challenge interpretive assumptions and explore alternative conceptual groupings.

(5) Translating studies into one another (Sec. 9). Reciprocal translations were developed iteratively. Semantic connectors were initially assigned individually by authors A-C and E and then validated collaboratively until consensus was reached. Cases of uncertainty or disagreement were discussed in group review sessions and 1st- and 2nd-order interpretations were consulted when discrepancies arose. This stage involved explicit reflexive questioning about what was being emphasised or omitted in our comparisons, and how CASP ratings influence a study's coverage.

(6) Synthesising translations (Sec. 10). Authors A-C and E created 3rd-order interpretations over several synthesis meetings. Competing interpretations were discussed and alternative framings (e.g., authorship, accountability, reflective practice) tested against the primary studies. Decisions were grounded in textual evidence, not majority opinion.

(7) Expressing the synthesis (Sec. 11). The line-of-argument was drafted collaboratively and iteratively revised by authors A-C and E. Reflexive attention was given to the risks of overgeneralisation, the influence of our shared agnostic stance toward GenAI, and the need to clearly surface tensions and boundary conditions within the literature.

B.2 Individual Author Positionality

The following brief statements outline the epistemic and disciplinary perspectives each author brought to the synthesis. These statements are intentionally anonymised, focusing on intellectual positioning rather than biographical detail.

Author A is a designer and researcher, with a background in product and service design, currently working within HCI, with a focus on how creative practitioners think, decide, and collaborate with technology. They approach GenAI as both a design material and a workplace actor, interested in how its integration changes expertise, authorship, and everyday design work.

Author B works in game HCI, with a focus on player and developer experiences, and the dynamics of tool adoption in game production. They approach GenAI pragmatically, attending to how it reorganises labour, coordination, and decision-making within teams.

Author C specialises in computational creativity and interactive media research. Their work examines how creators engage with computational tools, and how such tools foster exploration, iteration, and creative identity. They adopt a broadly constructivist stance, viewing creativity as emerging from interactions among people, artefacts, and technological systems.

Author D has done extensive research on game-development practices, particularly the lived experiences of developers working in an evolving landscape. They have engaged in conversations with hundreds of developers from a wide range

of nationalities and roles, from leaders in AAA studios to independent creators and hobbyists with diverse design values. Focusing on Northern Europe, the author's interactions also include developers from across East and South Asia, the Americas, Australia, and South Africa. They bring sensitivity to organisational dynamics and the lived realities of creative work.

Author E researchers computational creativity, human-AI collaboration, and game AI, with a longstanding interest in the boundaries between human and machine creative agency. They seek to support a societally and culturally sustainable future of AI via empirical work, and by letting the latter guide AI research. Their perspective foregrounds reflective practice and the ethical and epistemic implications of integrating AI into creative processes. Here, they employed an interpretative stance and a critical mindset toward GenAI research and adoption practices.

While individual perspectives differ slightly in criticality, we share a broadly agnostic stance toward GenAI: recognising its potential value when used to complement, rather than replace, human creative work, and cautious of framings that overstate automation or diminish human judgement and creative expression. These shared commitments informed our interpretive focus on creative negotiation, reflective engagement, authorship, and accountability.