The Black Box in the Research Room: LLM Interpretability Challenges in Virtual World Research Methodology

Paul Penfold
Brave Generation Academy, Chiang Mai, Thailand
ORCID: https://orcid.org/0009-0000-0789-1019
paul@bravegenerationacademy.com

ABSTRACT

The deployment of large language models (LLMs) as autonomous research agents within virtual world studies represents a paradigm shift in social science methodology — and an underexamined interpretability crisis. As immersive research platforms mature from early Second Life deployments to contemporary multi-platform ecosystems spanning VRChat, Roblox, and Meta Horizon, methodologists have developed sophisticated agentic infrastructures in which LLMs function as Primary Agents conducting interviews, Critique Agents monitoring researcher bias, Safety Agents protecting participant wellbeing, and Synthetic Participants pre-testing experimental environments. These roles are not incidental to the research process: they mediate data collection, shape participant experience, and increasingly underwrite validity claims. Yet the interpretability of LLM behaviour in these contexts remains almost entirely unaddressed in the literature. Drawing on a research trajectory spanning Guillet and Penfold's (2013) foundational avatar-based hospitality study — the first known hospitality research conducted exclusively in a virtual world — through to Penfold's (2026) Sequential Process Model for Immersive Inquiry, this paper identifies five distinct interpretability gaps that emerge when LLMs are embedded in virtual world research protocols. We argue that these gaps constitute a genuine validity threat, articulate a structured research agenda for LLM interpretability in immersive research contexts, and call for collaboration between interpretability researchers and virtual world methodologists before agentic research infrastructures scale further.

Keywords: *LLM interpretability, virtual world research, agentic AI, synthetic participants, immersive methodology, research validity*

1. Introduction

The question of what a machine actually does when it performs an intellectual task — as opposed to what we observe it producing — has shadowed computational social science since its inception. For virtual world research, that question has become acute. Over the past two decades, the methodological landscape of immersive social inquiry has transformed from tentative experiments in avatar-based data collection to sophisticated, AI-mediated research ecosystems in

which LLMs play structurally integral roles. Understanding what those LLMs are actually doing — and being able to verify it — is no longer a theoretical nicety. It is a precondition for research validity.

The trajectory of this methodological evolution is instructive. In April 2008, early qualitative investigations into teacher experience in Second Life documented a fledgling but serious academic engagement with virtual worlds as research settings: five university teachers in Asia who had integrated Second Life into their teaching described complex negotiations of authenticity, pedagogical legitimacy, and participant control (Penfold & Al Hadhrami, 2008). At the time, more than 500 universities and institutions globally were using Second Life as an educational platform, with Linden Lab reporting 238 million registered accounts worldwide (Mitham, 2008). The platform offered something genuinely new — a persistent, socially inhabited digital space in which participants could be recruited, observed, and interviewed under conditions that had no straightforward analogue in traditional survey or laboratory research.

The methodological maturation of this space was demonstrated at scale by Guillet and Penfold (2013), who conducted what remains the first known hospitality research study executed exclusively within a virtual world. Their hotel co-branding study, conducted in Second Life using avatar-based immersive techniques, achieved a 72% completion rate with over 700 participants across 39 countries — validity metrics that compared favourably with conventional survey instruments while dramatically reducing cost and expanding geographic reach. The study established that virtual world methods were not merely novel but genuinely productive: they could generate data of comparable quality to traditional approaches while accessing participant populations and behavioural contexts that physical research settings could not replicate.

What neither that study nor the broader methodological literature of its era anticipated was the degree to which artificial intelligence would come to inhabit the research process itself. Penfold's (2026) pre-publication framework, "Conducting Immersive Research in Virtual Worlds: Contemporary Applications and Best Practices," represents the current frontier of this methodological development. Its Sequential Process Model for Immersive Inquiry describes a four-phase research architecture in which LLM-driven agents conduct interviews, monitor bias, protect participants, and simulate human respondents in pilot phases. The framework is technically sophisticated and empirically grounded. It also introduces, without fully resolving, a family of interpretability problems that this paper is concerned with naming and structuring.

The core problem is this: when an LLM agent conducts a research interview, monitors another agent for bias, or simulates a consumer's responses to a virtual hotel environment, the validity of those operations depends on our ability to understand what the LLM is actually doing internally — not just what it outputs. The field of LLM interpretability has developed substantial tools for this purpose, including probing classifiers, activation patching, attention analysis, and mechanistic circuit identification (Olah et al., 2020; Conmy et al., 2023). However, these tools have been developed and applied almost exclusively in contexts where the LLM's task is

well-defined and its outputs are directly evaluable. The agentic research roles described in Penfold's (2026) framework present a different challenge: the outputs are complex, context-dependent, and frequently unverifiable against a ground truth. This paper argues that closing this gap requires a deliberate collaboration between the interpretability research community and the virtual world research methodology community — and that the urgency of that collaboration is underestimated on both sides.

2. LLMs as Active Participants in Virtual World Research

To appreciate the interpretability challenge, it is necessary first to understand the specific functional roles that LLMs now occupy within immersive research protocols. These roles are not decorative or peripheral. In the most advanced formulations of the Sequential Process Model for Immersive Inquiry (Penfold, 2026), LLMs perform structurally load-bearing functions at each of the four research phases.

Phase II: Immersive Recruitment introduces what Penfold terms "quest-based recruitment funnels" — gamified pathways through which potential participants are identified and screened within the virtual environment itself. Critically, this phase also incorporates Proof of Personhood (PoP) protocols drawing on systems such as Humanity Protocol and World ID, designed to prevent bot-participant contamination. The interpretability question embedded here is not trivial: the LLM systems that adjudicate PoP verification are themselves making inferences about personhood based on behavioural signatures. Whether those inferences are reliable, what features they are sensitive to, and whether they exhibit systematic biases toward particular avatar presentations or interaction styles are questions that cannot be answered by observing outputs alone.

Phase III: Hybrid Data Capture is where LLM agency becomes most structurally central. Penfold's (2026) framework describes "AI-Enhanced Ethnography" in which Autonomous Research Agents perform real-time sentiment analysis on both voice and avatar body language. These agents operate alongside spatial telemetry systems — heatmaps, movement pattern analysis — and biosensor integrations that capture heart rate variability, skin conductance, and pupil dilation. Kim et al. (2024) report a 35% reliability increase attributable to biosensor integration. The AI systems that translate raw biometric signals into research-usable inferences about emotional states or engagement levels are, in the framework's current formulation, treated as reliable instruments. The basis for that assumption is not yet established in the interpretability literature.

The concept of Synthetic Participants represents perhaps the most significant interpretability challenge in the entire framework. As Penfold (2026) describes, drawing on Thompson et al. (2025), "Synthetic Participants — LLM-driven agents trained on longitudinal consumer data — allow researchers to pre-test virtual environments before human deployment. This 'Synthetic Pilot' phase can identify UI friction points and ethical 'red zones,' substantially reducing the risk

and cost associated with human trials." This is a genuinely valuable methodological innovation. It is also one that requires us to be very precise about what an LLM-driven agent is actually doing when it "completes" a research protocol. Is it simulating the phenomenological experience of a consumer navigating a virtual hotel lobby? Is it retrieving and interpolating statistical regularities from its training corpus about how consumers of a particular demographic profile behave in such contexts? Is it doing something else entirely — pattern-matching to the structural features of the protocol itself? These are not rhetorical questions. They have direct implications for whether Synthetic Pilot data is valid as a proxy for human participant data.

Model-Arbitrated Qualitative Inquiry, described by Penfold (2026) following Thompson et al. (2025), instantiates a three-agent architecture: a Primary Agent conducts the research interview, a Critique Agent monitors for researcher bias in real time, and a Safety Agent ensures participant emotional wellbeing throughout. This architecture is elegant and addresses genuine methodological problems — researcher bias in qualitative interviews is well-documented and difficult to control — but it raises an immediate interpretability question: how does the Critique Agent detect bias? What internal representations does it form of "researcher bias"? And how do we verify that what it is detecting as bias corresponds to the construct that qualitative methodologists mean by that term, rather than some statistically adjacent but conceptually distinct pattern in the Primary Agent's outputs?

These are not edge cases or speculative concerns. They are questions about whether the research infrastructure described in Penfold's (2026) framework is doing what it claims to do. The answer requires interpretability research.

3. The Interpretability Gap

LLM interpretability as a field is concerned with understanding the internal mechanisms by which large language models produce their outputs — not merely characterising those outputs behaviourally, but identifying the computational structures responsible for them (Räuker et al., 2023). The field has made genuine progress. Mechanistic interpretability work has identified specific circuits responsible for tasks such as indirect object identification (Wang et al., 2022), induction (Olsson et al., 2022), and factual recall (Meng et al., 2022). Representation probing has revealed systematic structure in how models encode semantic and syntactic properties. Activation patching has enabled researchers to attribute specific behavioural differences to specific internal states.

These tools, however, have been developed and validated primarily in contexts where the task is well-specified, the input-output relationship is tractable, and some form of ground truth is available for comparison. When an LLM completes a fill-in-the-blank factual task or resolves a syntactic ambiguity, there is a fact of the matter about whether it has done so correctly, and interpretability tools can be deployed in reference to that fact. The agentic research roles described in the preceding section do not offer this luxury.

Consider the Synthetic Participant scenario. A Synthetic Participant navigating a virtual hotel environment and responding to survey questions about its experience is producing outputs in a domain where the ground truth — what a real consumer would actually experience and report — is precisely what we are trying to model. We cannot straightforwardly evaluate whether the Synthetic Participant's outputs are valid simulations without the human participant data we are using it to avoid collecting. This circularity is not a problem unique to virtual world research: it arises whenever LLMs are used as substitutes for human respondents (Argyle et al., 2023). But it is especially acute in immersive research contexts because the simulation must extend across a richer behavioural space — navigation choices, dwell time, avatar orientation, paralinguistic cues — not just text responses.

The Critique Agent scenario is structurally different but equally challenging. Here, the relevant ground truth is a normative construct — "researcher bias" — that is itself contested and contextually variable in the qualitative methods literature. The interpretability question is not just whether the Critique Agent is processing the Primary Agent's outputs correctly, but whether its internal representation of "bias" is aligned with the construct as qualitative methodologists understand it. This is an instance of what might be called the construct validity problem for LLM agents: even if the agent's internal mechanisms are functioning as designed, they may be targeting a computationally tractable proxy for the intended construct rather than the construct itself.

The Safety Agent raises a third variant of the problem. Its function — ensuring participant emotional wellbeing — requires real-time inference about an avatar-embodied participant's emotional state, based on some combination of text, voice, and spatial behavioural signals. The biosensor integration described in Penfold's (2026) framework (drawing on Kim et al., 2024) improves reliability, but the AI systems that translate biometric data into emotional state inferences are themselves opaque. Whether a spike in skin conductance response in a participant wearing a VR headset in their home environment reflects distress related to the research content, unrelated environmental stimuli, or physiological noise is not determinable from the signal alone. The Safety Agent's inferences are probabilistic, contextually shaped, and potentially systematically wrong in ways that a purely behavioural evaluation of its outputs would not reveal.

The cross-platform consistency problem adds a further dimension. The contemporary research landscape is heterogeneous: Second Life maintains approximately 200,000 daily users (Linden Lab, 2023), VRChat serves 3 million monthly active users (VRChat, 2025), Roblox reports 151 million daily active users (Roblox Corporation, 2025), and Meta Horizon has reached 1 million premium subscribers (Meta, 2025). These platforms differ substantially in their interaction affordances, avatar embodiment conventions, user demographics, and social norms. The EU Joint Research Centre's 2024 report "Virtual worlds, real impact" documented over 88,000 activities across more than 68,000 organisational players, spanning creative and cultural industries, tourism, retail, healthcare, and aerospace. An LLM-driven research agent that

performs reliably as a Synthetic Participant in a Second Life hospitality study may behave quite differently when deployed in Roblox's interaction environment — not because of a design failure, but because the LLM's training distribution, its contextual priors, and the affordances it is navigating all vary.

These five interpretability gaps — the Synthetic Participant validity problem, the Critique Agent construct validity problem, the Safety Agent inference reliability problem, the Avatar-Researcher disclosure and persona representation problem, and the cross-platform consistency problem — constitute a coherent and urgent research agenda. None of them is addressable by better prompt engineering or larger training datasets alone. They require interpretability research.

4. Evidence from the Research Record

The case for treating this interpretability gap as urgent, rather than speculative, rests on what the existing research record tells us about both the pace of adoption and the standards of validation that have historically accompanied methodological innovation in virtual world research.

Guillet and Penfold's (2013) study is the appropriate benchmark because it was the moment at which virtual world research demonstrated it could meet conventional social science validity standards. The 72% completion rate across 700+ participants in 39 countries was not achieved through methodological hand-waving: it required careful attention to participant recruitment, instrument design, and data quality. The study's significance lay not just in its findings but in its demonstration that immersive research methods could be held to the same validity standards as the approaches they supplemented. That demonstration mattered for the field's credibility, and it required transparency about the methods used to achieve it.

The Sequential Process Model for Immersive Inquiry (Penfold, 2026) represents the methodological evolution that Guillet and Penfold's (2013) work prefigured. Its four phases — Platform Calibration, Immersive Recruitment, Hybrid Data Capture, and Multi-Method Validation — constitute a serious attempt to formalise best practice for contemporary immersive research. The calibration parameters are precise: Motion-to-Photon latency thresholds of 58ms for cybersickness onset and 69ms for performance degradation (Thompson et al., 2025) reflect real engineering constraints with real perceptual consequences. The blockchain verification protocols for cross-platform data security, associated with a 40% improvement in data integrity (Chen & Rodriguez, 2024), address genuine concerns about data provenance in multi-platform studies. The sustainability framing — an estimated 70–80% reduction in participant travel emissions relative to physical laboratory studies (Shift Project, 2024) — situates the methodological framework within a broader ethical calculus that takes its responsibilities seriously.

What is notable, given this level of technical precision in other aspects of the framework, is the relative absence of comparable precision regarding LLM agent behaviour. The AI reliability figures cited — Thompson et al.'s (2025) report of up to 40% improvements in data accuracy and

60% reductions in analysis time from AI-assisted protocols — are outcomes-level metrics. They tell us that AI-assisted research produces better-measured outputs than non-AI-assisted research by certain criteria. They do not tell us why, or through what internal processes, or whether those processes are stable across the range of contexts in which the framework is intended to operate.

This is not a criticism of the framework. It is an observation about the current state of the field. The interpretability tools required to answer those questions exist — in nascent and developing form — in the AI interpretability literature. The ethical dimensions of this gap are documented in Penfold's (2026) framework through its treatment of Granular Biometric Sovereignty (Watson & Zhang, 2024) and the Right to Non-Interpretation — the principle that foveated eye-tracking and pupil dilation data should be governed under emergent 2025 AI Acts rather than treated as straightforwardly available data for researcher use. The Avatar-Researcher Power Dynamics concern — requiring disclosure of AI interviewers through an "AI-UI" indicator — acknowledges that participants in virtual world research may not be able to distinguish AI from human interviewers. This is precisely the condition that makes interpretability research urgent: if participants cannot tell whether they are interacting with a human or a machine, and researchers cannot fully characterise what the machine is doing, then the epistemic foundations of the research interaction are doubly opaque.

5. A Research Agenda for LLM Interpretability in Immersive Research

We propose five priority research questions that, taken together, constitute a tractable agenda for addressing the interpretability gap identified above. These are framed explicitly as open questions requiring empirical investigation, not as findings.

Research Question 1: What are Synthetic Participants actually responding to? When an LLM-driven Synthetic Participant completes a virtual world research protocol — navigating a simulated hotel environment, responding to avatar-mediated survey questions, or recording dwell-time preferences — what computational processes generate its responses? Interpretability methods including probing classifiers and activation patching could, in principle, be used to determine whether the LLM's responses are driven by internal representations that correspond to consumer preference simulation, or by lower-level pattern matching to the structural features of the research instrument. Establishing this distinction matters because if Synthetic Participant responses are primarily instrument-sensitive rather than domain-sensitive, the Synthetic Pilot phase of the Sequential Process Model would identify instrument artefacts rather than genuine UI friction points or ethical red zones.

Research Question 2: How do Critique Agents operationalise researcher bias, and is that operationalisation valid? The Model-Arbitrated Qualitative Inquiry architecture assigns a Critique Agent the task of monitoring for researcher bias in a Primary Agent's interview conduct. Before this architecture can be trusted as a methodological safeguard, we need to understand what internal representations the Critique Agent forms of the construct "researcher bias," how

those representations were acquired, and whether they correspond to the construct as defined in qualitative methods literature. Probing classifiers trained on qualitative methods literature and applied to Critique Agent internal states offer one investigative pathway. Comparison with human expert judgements of bias in the same interview transcripts offers a validation benchmark.

Research Question 3: Are Safety Agent inferences about participant wellbeing reliable and interpretable? The Safety Agent's function requires it to make real-time inferences about participant emotional states from multimodal inputs including text, voice, avatar behaviour, and physiological signals. Whether the AI systems performing these inferences are doing so through mechanisms that are aligned with the theoretical constructs of emotional state — rather than with correlated but distinct features — is unknown. Interpretability research on emotion representation in LLMs could provide tools for this investigation. The specific challenge of multimodal integration — combining linguistic, paralinguistic, and physiological signals — is an open problem for interpretability research more broadly.

Research Question 4: What does the Avatar-Researcher Power Dynamics disclosure requirement reveal about LLM persona representation? Penfold's (2026) ethical framework requires that AI interviewers in virtual world research be disclosed through an "AI-UI" indicator, on the grounds that participants may otherwise be unable to distinguish them from human researchers. This points toward an empirical question: how do LLMs represent and sustain a research interviewer persona across an extended avatar-mediated interaction? The interpretability of persona maintenance in LLMs — what internal representations are activated, how consistently they are sustained, and whether they exhibit systematic drift — is a live research question whose answer matters for virtual world research validity.

Research Question 5: How consistent is LLM research agent behaviour across virtual world platforms? Whether LLM research agents exhibit consistent behaviour across Second Life, VRChat, Roblox, and Meta Horizon is an empirical question that has not been studied. Consistency here has multiple dimensions: consistency of output at equivalent task stages, consistency of internal representation, and consistency of bias profile. Developing a cross-platform LLM agent consistency assessment methodology would serve both the interpretability research community and the virtual world research methodology community.

6. Discussion

The five research questions articulated above are representative of a structural problem: the interpretability of LLM behaviour in agentic research roles has not been treated as a validity requirement by the virtual world research methodology community, and the agentic research roles developed by that community have not been treated as an application domain by the LLM interpretability community. Both oversights are understandable given the pace of development in both fields. Neither is sustainable.

The virtual world research methodology literature has historically been attentive to validity. The significance of Guillet and Penfold's (2013) study lay precisely in its demonstration that immersive methods could meet conventional validity standards — not that they were exempt from them. The introduction of LLM agents into the research process does not change that standard; it makes it more difficult to meet while simultaneously making it more important to try. If AI-assisted protocols improve data accuracy by 40% and reduce analysis time by 60% (Thompson et al., 2025), those gains are only meaningful if the AI systems achieving them are doing what they are claimed to do. Interpretability research is what allows us to make that determination.

There are also considerations of scale. The metaverse market is projected to grow at a compound annual growth rate of 43.3% through 2030 (Grand View Research, 2024). The EU JRC's (2024) documentation of over 88,000 activities across 68,000+ organisational players represents current deployment, not future projection. As virtual world research scales — in participant volume, in geographic reach, in the sensitivity of its subject matter — the interpretability gap identified in this paper will not remain academic. It will manifest as replication failures, as ethical incidents, as contested research findings, and as regulatory challenges under frameworks such as the 2025 AI Acts referenced in Penfold's (2026) treatment of biometric data governance.

The sustainability argument for virtual world research — the estimated 70–80% reduction in participant travel emissions relative to physical laboratory studies (Shift Project, 2024) — adds a further dimension. If virtual world research methods are to serve as credible substitutes for more carbon-intensive research modalities, their validity foundations must be correspondingly robust. We note that the interpretability agenda proposed here is not pessimistic about LLM agents in research roles. The framework described in Penfold (2026) represents a genuine advance in research methodology. The interpretability research we are calling for is not intended to block these innovations but to provide the foundations on which their validity can be established and communicated. Trust in research methods is not built by assertion; it is built by transparency about what those methods are doing and evidence that they are doing it reliably.

7. Conclusion

This paper has argued that the deployment of LLMs as autonomous research agents in virtual world studies — in the roles of Primary Agents, Critique Agents, Safety Agents, and Synthetic Participants — creates a family of interpretability challenges that constitute a genuine validity threat to the research enterprise. This argument is grounded in a research trajectory extending from early qualitative investigations into virtual world pedagogy (Penfold & Al Hadhrami, 2008), through the landmark avatar-based hospitality study of Guillet and Penfold (2013), to the sophisticated methodological framework of Penfold (2026).

The five research questions we have articulated — concerning what Synthetic Participants respond to, how Critique Agents operationalise bias, whether Safety Agent wellbeing inferences

are reliable, what persona maintenance by AI interviewers reveals about LLM representation, and whether LLM agent behaviour is consistent across platforms — do not admit of easy answers. They require sustained, technically demanding interpretability research conducted in close dialogue with the virtual world research methodology community.

Virtual world research has always been defined by the willingness of its practitioners to take methodological questions seriously — to insist that the novelty of the medium did not exempt it from the standards of rigour that govern social inquiry. That insistence is what gave Guillet and Penfold's (2013) work its significance and what grounds the ambitious methodological framework that Penfold (2026) offers. The same insistence now requires us to name the interpretability gap, to treat it as urgent, and to invest the research effort required to close it. The black box in the research room cannot be ignored simply because what it produces looks reasonable. We need to know what it is actually doing.

References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. Political Analysis, 31(3), 337–351.

Chen, J., & Rodriguez, M. (2024). Blockchain verification frameworks for multi-platform research data integrity. Journal of Research Methodology and Social Science, 12(2), 88–104.

Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). Towards automated circuit discovery for mechanistic interpretability. Advances in Neural Information Processing Systems, 36.

European Commission Joint Research Centre. (2024). Virtual worlds, real impact: The European opportunity. Publications Office of the European Union.

Grand View Research. (2024). Metaverse market size, share & trends analysis report. Grand View Research.

Guillet, B. D., & Penfold, P. (2013). Testing co-branding strategies of hotel brands with discrete choice analysis in a virtual reality environment. International Journal of Hospitality & Tourism Administration, 14(1), 23–49. https://doi.org/10.1080/15256480.2013.754238

Kim, S., Park, J., & Lee, H. (2024). Biosensor integration in virtual reality research: Reliability improvements in immersive participant monitoring. Cyberpsychology, Behavior, and Social Networking, 27(4), 211–228.

Linden Lab. (2023). Second Life platform statistics. Linden Research, Inc.

Lindsey, J., Gurnee, W., Heimersheim, S., Janiak, B., Elhage, N., Mossing, D., Lampinen, A., Thomas, T., Rivoire, R., MacDiarmid, C., & Anthropic. (2025). Biology of a large language model. Anthropic.

Martinez, R., & Wilson, K. (2023). Gamification strategies and participation rates in virtual world research. Journal of Interactive Digital Media, 8(1), 45–63.

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. Advances in Neural Information Processing Systems, 35, 17359–17372.

Meta. (2025). Meta Horizon Worlds platform overview. Meta Platforms, Inc.

Mitham, N. (2008). Virtual worlds: The metrics of immersive online environments. KZero Worldswide.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom in: An introduction to circuits. Distill. https://doi.org/10.23915/distill.00024.001

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., & Olah, C. (2022). In-context learning and induction heads. Transformer Circuits Thread.

Penfold, P. (2026). Conducting immersive research in virtual worlds: Contemporary applications and best practices. Pre-publication manuscript.

Penfold, P., & Al Hadhrami, A. (2008). Virtual education in Second Life: Teachers' perspectives. Unpublished research data.

Räuker, T., Ho, A., Casper, S., & Hadfield-Menell, D. (2023). Toward transparent AI: A survey on interpreting the inner structures of deep neural networks. 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), 464–483.

Roblox Corporation. (2025). Q1 2025 key metrics and financials. Roblox Corporation.

Shift Project. (2024). Carbon footprint of digital research methodologies: Comparative analysis. The Shift Project.

Thompson, A., Nakamura, K., & Osei, B. (2025). Multi-component foundation systems for agentic research infrastructure. AI & Society, 40(2), 341–359.

VRChat. (2025). VRChat platform statistics and community overview. VRChat, Inc.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. International Conference on Learning Representations (ICLR 2023).

Watson, E., & Zhang, L. (2024). Granular biometric sovereignty: Governance frameworks for immersive research data. Journal of Law, Information and Science, 33(1), 77–99.