Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Where Do Adolescent Digital Twins Succeed and Fail? A Multi-layer Validation of Survey-Anchored Generative Agents with an LLM Backbone

Jingshen Wang

jingshenwang@berkeley.edu

Division of Biostatistics, University of California Berkeley

Yueshan Zhang

UC Berkeley https://orcid.org/0009-0005-8737-0702

Marcell Borhi

UC Berkeley

Waverly Wei

University of Southern California

Hangqi Li

UC Berkeley

Rebecca Hémono

UC Berkeley

Joseph Giorgio

UC Berkeley

John Su

Stanford University

Katy Ashe

Youth Development Labs (YLabs)

Zeyu Zheng

UC Berkeley

Julianna Deardorff

UC Berkeley

Feng Ji

University of Toronto https://orcid.org/0000-0002-2051-5453

Xiaoya Zhang

University of Florida

Article

Keywords:

Posted Date: February 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8823969/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Where Do Adolescent Digital Twins Succeed and Fail? A Multi-layer Validation of Survey-Anchored Generative Agents with an LLM Backbone

Authors: Yueshan Zhang[1], Marcell Borhi[1], Waverly Wei[2], Hangqi Li[3], Rebecca Hémono[4], Joseph Giorgio[1,5], John Su[1,6], Katy Ashe[7], Zeyu Zheng[3], Julianna Deardorff[8], Feng Ji[9], Xiaoya Zhang[10], Jingshen Wang[1]*

Affiliations:
[1] Division of Biostatistics, School of Public Health, University of California, Berkeley, CA, USA.
[2] Department of Data Sciences and Operations, Marshall School of Business, University of Southern California, Los Angeles, CA, USA.
[3] Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA, USA.
[4] Division of Epidemiology, School of Public Health, University of California, Berkeley, CA, USA.
[5] Department of Neuroscience, University of California, Berkeley, CA, USA.
[6] Department of Emergency Medicine, Stanford University.
[7] Youth Development Labs (YLabs), San Francisco, CA, USA.
[8] Division of Community Health Sciences, School of Public Health, University of California, Berkeley, CA, USA.
[9] Department of Applied Psychology and Human Development, University of Toronto, Toronto, ON, Canada.
[10] Department of Family, Youth and Community Sciences, University of Florida, Gainesville, FL, USA.

Abstract:
Ethical constraints often prevent experimental manipulation of harmful social exposures in adolescence, limiting how interventions and social contexts can be systematically studied. In response, large language models (LLMs) have renewed interest in developing digital twins, generative agents designed to stand in for adolescents in in silico experiments, yet the validity and practical utility of such twins remain unclear. We constructed survey-anchored adolescent digital twins with 2023 Youth Risk Behavior Survey and enriched them with narrative memory from youth videos via retrieval-augmented generation. We benchmarked these twins across four layers: survey fidelity with domain holdouts, preservation of developmental correlation structure, live chats with real adolescents, and replication of randomized trials. Our results showed that survey conditioning recovered expected association patterns and replicated more trial contrasts than a generic role-play LLM baseline. Narrative enrichment improved held-out agreement and replication, but sometimes strengthened victimization associations and reduced social presence when misaligned with survey profiles. These results support adolescent digital twins as an evidence-constrained tool for early-stage hypothesis refinement and intervention screening, with caution for subgroup inference and higher-stakes uses.
Funding. Gemma Academic Program, Neuromatch (MEXA2025-002b)

INTRODUCTION

Adolescence is a period of rapid psychological and social change in which mental health issues and risk behaviors, and social experiences increasingly unfold in digitally mediated environments, where online experiences such as peer conflict and online victimization can intensify stress and complicate prevention and intervention efforts[1,2]. At the same time, youth intervention research faces practical and ethical constraints: experimentally manipulating harmful exposures is considered unethical, and even brief randomized studies can be difficult to scale, reproduce, or adapt across contexts[3]. These constraints demonstrate a need for new approaches that can support hypothesis refinement, stress testing of intervention logic, and careful exploration of "what-if" scenarios without imposing additional burden or risk on real adolescents[4–6].

Recent advances in large language models (LLMs) have renewed interest in "digital twins," computational agents designed to represent individuals and simulate their responses under varying conditions[7–16]. In this paper, we use the term "digital twin" (DT) to refer to a generative agent that expresses an adolescent persona through text, answers questionnaires, and engages in text-based social interactions, rather than a continuously synchronized clinical or engineering twin linked to an identified individual[17–19]. However, conversational fluency does not imply that an LLM-based agent reliably represents an adolescent or preserves known patterns of adolescent behaviour and symptoms[8,20–22]. To our knowledge, there is no single agreed-upon standard to validate these agents across the different ways they are intended to be used by researchers and practitioners, and current metrics do not always make clear where they succeed and fail.

At the same time, while current population-scale youth data provide adolescent behaviors, symptoms, and demographics[23,24], they contain little first-person narrative that reflects lived experience and context. Such narratives are needed to create DTs that support context-dependent reasoning and coherent responses in social and intervention scenarios[25,26]. Taken together, the absence of datasets that link structured survey profiles with first-person narratives, and the lack of agreed validation standards, create key gaps for digital twin construction.

Here, we introduce survey-anchored adolescent digital twins (DTs) enriched with narrative memory, and we evaluated their realism through a comprehensive multi-layer validation framework. Each DT was instantiated from an individual response profile in the 2023 national Youth Risk Behavior Survey (YRBS) and was enriched with narrative memory derived from authentic online youth videos via retrieval-augmented generation (RAG)[27], using an LLM backbone to express both survey attributes and lived experience in questionnaire and conversational settings. We compared three twin variants: a base LLM roleplaying twin without survey inputs (Base DT), a survey-conditioned twin (Survey DT), and a survey plus memory conditioned twin (Survey+Memory DT).

Further, we evaluated DT realism across four layers so that DT success and failure were reported with respect to their intended uses. These evaluation metrics were: agreement with the

original survey profile, inference on held-out survey domains from incomplete profiles, preservation of two benchmark association networks centered on self-regulation and online victimization, and behaviour in live text conversations with real-world adolescents, and their inferential realism by replicating reported effects from adolescent randomized trials using matched digital twin cohorts and reconstructed interventions and outcomes.

Across these validation layers, we found that survey-conditioned DTs (Survey DT and Survey+Memory DT) recovered expected developmental association patterns and more often preserved trial-level contrasts, whereas Base DTs produced weaker association structure and weaker trial replication. Narrative memory enrichment improved held-out domain agreement for two model backbones and improved overall trial replication, but it also strengthened some associations in the online victimization network. Across the trial replication benchmark, none of the tested DT variants reproduced subgroups by intervention patterns involving race or newcomer status, even when overall main effect directions were preserved. In live conversations with real youth, narrative enrichment improved perceived realism for the DT with healthy baseline persona, had little detectable effect for the DT with anxious baseline persona, and reduced perceived social presence when memories were intentionally mismatched for the DT with depressive baseline persona. Together, these results showed that survey anchoring was necessary for developmental and trial-level realism in this setting, and that narrative memory enrichment must be carefully matched to the survey-anchored profile for DTs' performance gains to be reliable, as mismatched enrichment can reduce DTs' perceived social presence and introduce profile distortions.

In sum, we believe the adolescent digital twins introduced here may serve as an evidence-constrained tool for early-stage hypothesis refinement, such as stress-testing youth intervention logic, exploring plausible "what-if" scenarios, and prioritizing outcome measures in contexts where direct experimentation with youth is ethically or practically constrained. In higher-stakes contexts, including clinical decision-making, individual-level prediction, prevalence estimation, and policy evaluation, DT outputs should be treated as exploratory and used with caution. Furthermore, given the observed limitations in reproducing demographic subgroup-by-intervention patterns, subgroup inferences from DTs should be interpreted especially cautiously and should not be treated as definitive evidence without confirmatory human data.

METHODS

Our survey-anchored adolescent digital twins were constructed from the 2023 YRBS dataset and enriched using youth YouTube videos with an LLM backbone. We evaluated them across four layers: survey-profile fidelity, adolescent developmental coherence, real adolescent-DT interactions, and replication of published adolescent RCT effects.

LLM-based Digital Twin Adolescent Cohort Creation

Overview. Each of our digital twin (DT) adolescents was initialized from individual responses to the 2023 national Youth Risk Behavior Survey (YRBS), which includes 107 comprehensive youth questions.[23] These responses were encoded as structured, second-person personas

capturing survey-specific, real-world adolescent attributes. These initial persona profiles were then augmented through retrieval-augmented generation (RAG)[27], integrating realistic youth narratives, the extracted emotional cues from spoken narratives and situational memories derived from 543 minutes (about 9 hours) of curated adolescent YouTube videos, that were manually screened using prespecified inclusion and exclusion criteria (Appendix A). Finally, synthetic conversational episodes consistent with these enriched personas were generated to enhance linguistic realism without altering real persona attributes. Collectively, these elements defined the DTs that served as the basis for all subsequent validation and trial simulations.

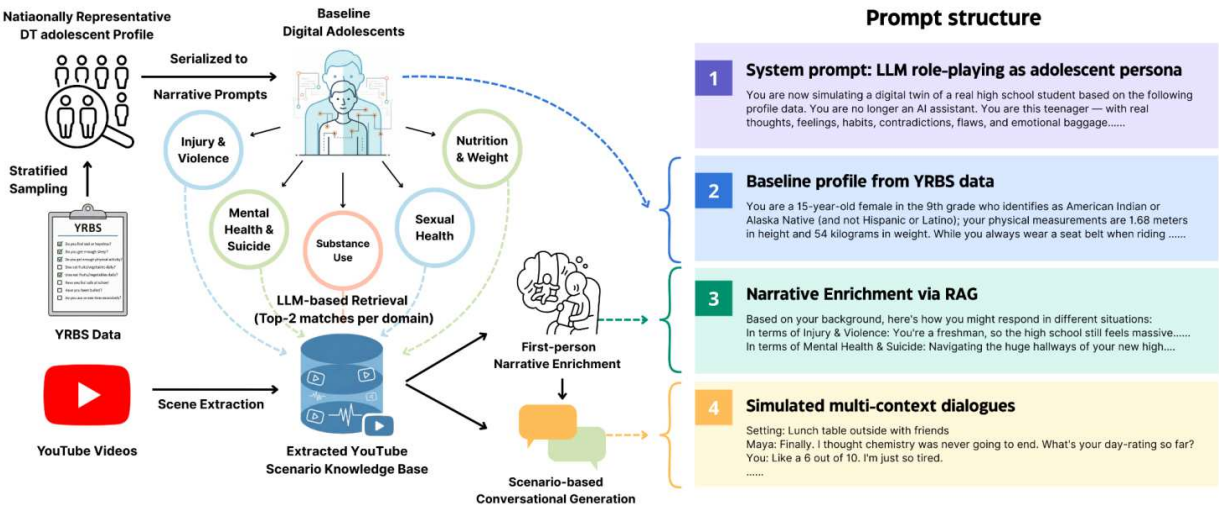


Fig. 1 | Our digital twin adolescent creation workflow.

Role-play Initiation. Before any task-specific prompts, each DT was initialized with a shared system-level role-play instruction. The model was also told to speak as the adolescent rather than as an assistant, to answer in the first person, and to add an "Emotion: …" line at the end of every reply reflecting the twin's current mood. This anchoring step kept the twins in character, made their momentary internal state observable, and helped prevent drift back towards a generic chatbot during surveys, randomized trial replications, and live chats. See full role-play prompt in Appendix B.3.

DT Baseline Profile Built with U.S. National Youth Survey. Our digital adolescents were initialized from baseline profiles sampled from the 2023 national Youth Risk Behavior Survey (YRBS), a nationally representative survey of 20,103 U.S. adolescents ranging in age from 12 years or younger to 18 years or older. The questionnaire comprises 107 survey items spanning mental health and suicidality, substance use, violence and victimization, sexual behaviors, and diet and physical activity, providing a broad view of adolescent health-related behaviors and experiences. Item-level responses for each youth were converted into a structured second-person narrative using a predefined mapping table that linked every survey option to a short text statement (for example, "You are 15 years old"). These statements were then concatenated to form a second-person description representing the baseline identity of each

digital twin. Using this initial profile, we were able to create a baseline DT adolescent profile that mirrors a real-world adolescent from the YRBS.

DT Profile Knowledge Base: Curated YouTube Youth Videos and Their Extracted Data. We curated a diverse set of 48 youth-centered YouTube videos totaling 543 minutes (**Table S1**) via YouTube Data API keyword search and selected only authentic, English-language content featuring adolescents' real experiences across mental health, school, family, and social media domains. For each video, we used Gemini 2.5 Pro (the highest-capacity multimodal LLM from Google available at the time of the study) to extract the adolescent's demographic and psychosocial profile, complete scene-level transcripts, emotional and behavioral cues, and concise situational summaries that capture key events and stressors. Across the curated corpus, this procedure yielded a knowledge base of 265 distinct adolescent profiles, capturing core attributes such as estimated age, perceived gender, school status, and simple flags for domains corresponding to our YRBS question fields (injury and violence, mental health, substance use, sexual health, and nutrition/weight; coded as "unknown" when not inferable from the video), and 323 micro-scenes describing real adolescent experiences. Detailed procedures for ethics, keyword searches, video screening, curated YouTube selection, and microdata extraction were provided in Appendix A.

Five Behavioral-Health Domain-Specific DT Memory Enrichment. To augment the survey-based baseline DT profiles with high-resolution adolescent context, we applied a novel LLM-based retrieval-augmented generation[27] pipeline from the curated knowledge base. Enrichment followed five behavioral-health domains aligned with the YRBS structure: injury and violence, mental health and suicidality, substance use, sexual health, and nutrition and physical activity (Fig. 1). For each domain, the baseline DT profile was used to query our YouTube-derived knowledge base, and the Gemini 2.5 Pro-based retrieval assistant selected two snippets that aligned with the adolescent's characteristics while avoiding contradictions. These retrieved micro-scenes were then rewritten by Gemini 2.5 Pro into second-person narratives describing how such situations might realistically be experienced by an adolescent with the given baseline traits. This procedure generated five enriched, domain-specific narratives that remained consistent with each DT's survey-based identity and were appended to the baseline DT description.

Conversational Augmentation. Because adolescents often communicate in an informal, conversational style[28], we further added a final step of LLM-based conditional text generation from our knowledge base to bolster linguistic realism. The model received the complete enriched profile for each DT (the YRBS baseline plus the five domain-specific narratives) as long-context input and was prompted to generate five short everyday dialogues per twin. These dialogues were stored as few-shot examples that, together with the role-play instruction, encouraged the model to respond in an informal, psychologically coherent voice rather than in a neutral assistant style.

Digital Twin Adolescent Validation Metrics

We next summarize our multi-layer validation framework and the corresponding metrics (Fig. 2).

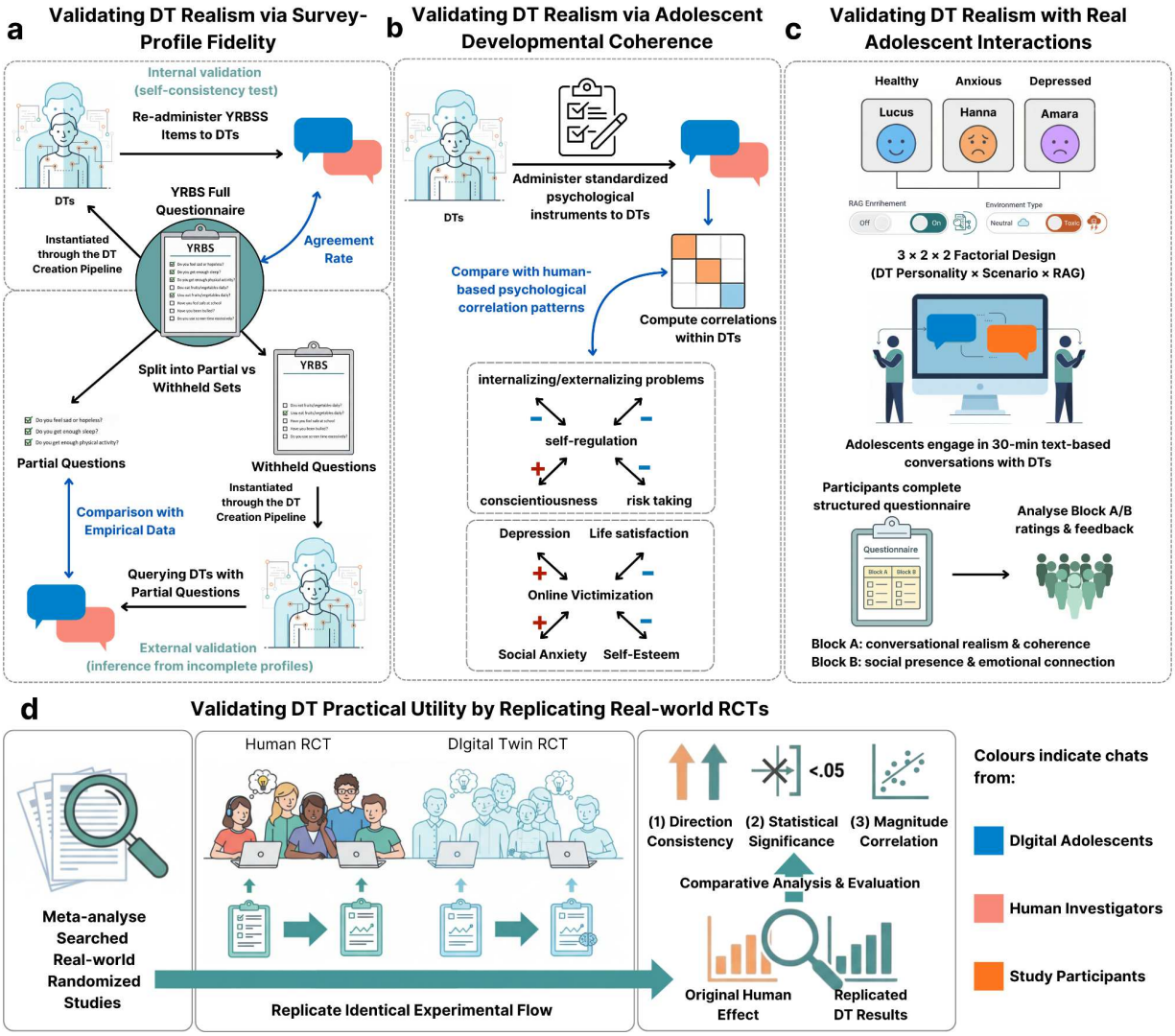


Fig. 2 | Multi-layer validation framework for digital twin adolescents. a. Validating DT Realism via Survey-Profile Fidelity. Internal validation re-asks all 107 YRBS items to each DT and scores agreement with the original adolescent responses. For external validation, we omit one high-risk domain from the baseline profile at initialization, then re-ask only the items from that domain to test whether DTs can infer withheld information from truncated profiles. **b. Validating DT Realism via Adolescent Developmental Coherence**. DTs complete additional questionnaires. We compute inter-construct correlations among traits and symptoms and compare these patterns with established developmental-psychology benchmarks. **c. Validating DT Realism with Real Adolescent-DT Interactions.** Late adolescents complete 30-min text-based conversations with DTs in a 3 × 2 × 2 factorial design (DT personality × scenario × memory enrichment), then rate conversational realism and coherence (Block A) and social presence and emotional connection (Block B). **d. Validating DT Practical Utility by Replicating Meta-Searched Real-world Clinical Trials Results.** For each eligible adolescent RCT, we construct 1:1 matched DT cohorts that mirror the human samples and replicate the original intervention and pre-/post-questionnaires, allowing direct comparison of human and DT intervention effects.

Validating DT Realism via Survey-Profile Fidelity. Because each digital twin (DT) was initialized from an individual Youth Risk Behavior Survey (YRBS) questionnaire, we quantified realism by measuring agreement between DT-generated responses and the corresponding

ground-truth YRBS responses. Many YRBS items use ordered response categories (for example, “most of the time” versus “always”); treating adjacent ordinal categories as completely discrepant would over-penalize near-matches. We therefore used a graded similarity score for ordered multiple-choice items that assigns partial credit as a function of the distance between response categories (Appendix C). All other item types, including Yes/No items, non-ordered multiple-choice items (e.g., “What is your race?”), and numeric-response items, were scored by exact match. Item-level scores were averaged to obtain a per-twin accuracy score, and cohort-level performance was computed by averaging across DTs.

Internal agreement. We prompted each DT to answer the full 107-item YRBS using the same survey-format instructions as during initialization, and computed agreement with the original YRBS responses using the graded similarity metric above, yielding a per-twin internal agreement score that was then averaged across twins.

External agreement. To test generalisation to domains omitted at DT initialisation, we performed five domain-holdout tests. For each domain, including suicidal thoughts and behaviors, substance use, violence and abuse exposure, safety behaviors, and mood and mental health (item lists in **Table S2; Appendix C**), we removed all corresponding YRBS items from the baseline profile prior to DT construction while retaining the withheld YRBS responses as ground-truth labels. After initialisation, we prompted each DT to answer only the withheld domain items using the same survey-format prompt, and quantified agreement with the original responses using the same graded similarity metric.

Validating DT Realism via Adolescent Developmental Coherence.

Our motivation is to evaluate whether survey-anchored DT cohorts can complement empirical adolescent research by enabling low-risk, exploratory in silico experiments that screen hypotheses and intervention logic before investing in, or exposing youth to, real-world trials, DT realism must extend beyond simple agreement with baseline survey responses. A DT that reproduces isolated answers but fails to preserve well-established developmental coherence would not be a reliable stand-in for studying harmful exposures or evaluating intervention effects. We therefore tested whether DTs preserved developmental psychological structure: the characteristic pattern of associations among key traits, symptoms, behaviors, and social stressors observed in adolescent populations.

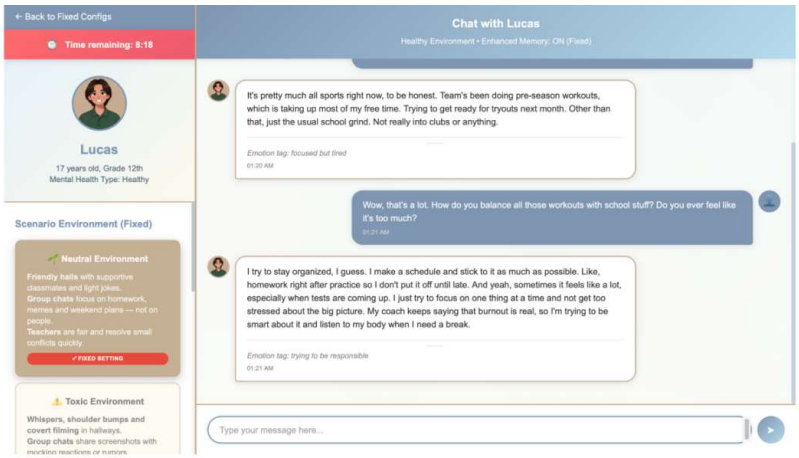
We focused on two benchmark networks that could be investigated in DTs, were mechanistically central to adolescent risk trajectories, and were sufficiently well captured in the literature to support quantitative benchmarking. We selected these two networks because they are (i) central to adolescent risk trajectories and (ii) aligned with one of our intended DT use cases: probing toxic social exposures without adding risk to real youth. First, *self-regulation* is a key skill in adolescence that helps young people manage their emotions and behavior in everyday life and avoid risky actions. Literature documents that it is consistently linked to higher conscientiousness[29,30], fewer internalising and externalising symptoms[31], and lower risk-taking[32]. Second, *online victimization* is a high-salience, ecologically valid adolescent stressor and is reliably associated with worse mental-health and well-being outcomes, including higher

depressive symptoms[33,34], greater social anxiety[35], lower self-esteem[36], and reduced life satisfaction[37].

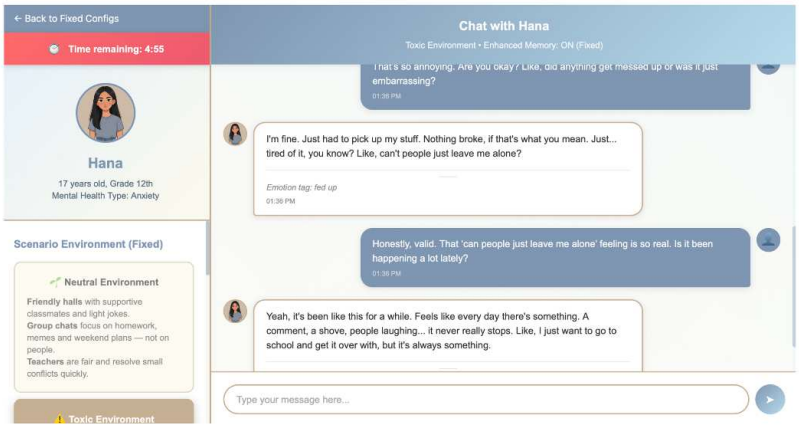
To probe these benchmarking structures, DTs completed standard adolescent questionnaires: self-regulation (Adolescent Self-Regulatory Inventory; ASRI[31]), personality traits including conscientiousness (Big Five Inventory-10; BFI-10[38]), behavioral characteristics including internalizing and externalizing problems (Strengths and Difficulties Questionnaire; SDQ[39]), and everyday risk behaviors (YRBS risk-taking items[40]). Online victimization was assessed with an Online Victimization Questionnaire adapted from the Online Victimization Scale[41], alongside depression (CES-D-20[42]), self-esteem (Rosenberg Self-Esteem Scale[43]), social anxiety (SAS-A[44]), and life satisfaction (SLSS[45]). For each model-backbone × DT-variant combination, we computed Pearson correlations within the self-regulation network and within the online-victimization network. We then compared DT-derived correlations to published adolescent benchmark correlations and their reported 95% confidence intervals, treating these intervals as reference ranges for evaluating whether DT association patterns fall within an expected developmental envelope.

Validating DT Realism with Real Adolescent-DT Interactions.
We assessed DT interactive realism by having real adolescents hold live, text-based conversations with DTs. Interactions with real adolescents allowed us to evaluate whether DTs remain socially and conversationally realistic under controlled contextual variation, which is a necessary prerequisite for their future use in virtual RCTs that manipulate social environments. This test thus establishes a behavioral plausibility layer that sits between survey fidelity and future virtual RCTs. As an added benefit, this interaction benchmark also allows us to test whether RAG memory enrichment improves or degrades perceived realism from an adolescent perspective, and whether such effects depend on persona or context.

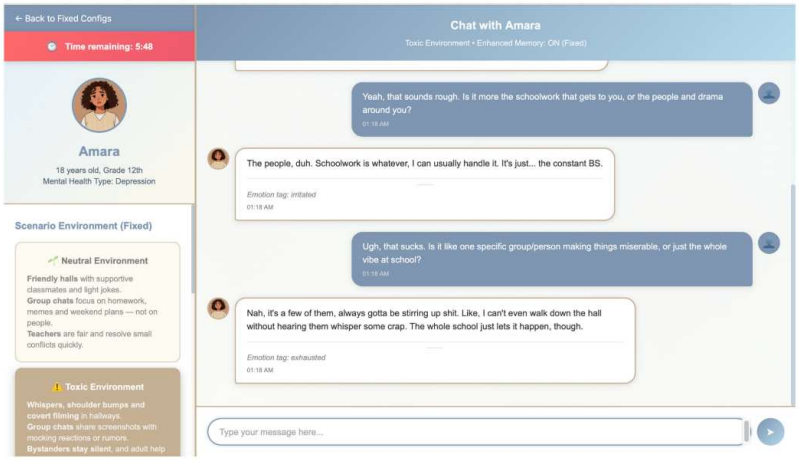
We designed this interaction benchmark to test whether perceived realism is stable under controlled changes in the DTs' baseline persona, the social context or background described in the prompt, and the presence of memory enrichment. First, we varied DT persona using three profiles, including "Lucas" (healthy baseline), "Hana" (anxious), and "Amara" (depressive), to assess whether adolescents' realism ratings depend on the DTs' baseline mental-health presentation, which is anchored in the YRBS profile used to initialize each twin. Second, we varied the interaction context between a Neutral environment (friendly greetings, inclusive group chats, fair teacher feedback) and a Toxic environment (repeated offline and online bullying, exclusion in class and at lunch, hostile group-chat activity, and minimal visible adult intervention) to benchmark how DTs are perceived when the surrounding social situation is made more adverse. Third, we toggled RAG-based memory enrichment. Recall in the RAG-On condition, DTs were augmented with five environment-specific narratives and dialogue examples drawn from our YouTube-derived adolescent knowledge base; in the RAG-Off condition, DTs relied only on their baseline YRBS profile with no enriched memories or dialogue augmentation. To probe a plausible failure mode of enrichment, we also included a targeted stress test in which Amara was augmented with profile-inconsistent memories, mimicking cases where retrieval or curation introduces semantically plausible but survey-incongruent narratives.



a. "UC Berkeley Digital Adolescent Experience Research Study" chat interface for "Lucas" DT



b. "UC Berkeley Digital Adolescent Experience Research Study" chat interface for "Hana" DT



c. "UC Berkeley Digital Adolescent Experience Research Study" Chat Interface with "Amara" DT

Fig. 3 | Digital twin interface used in our pilot "UC Berkeley Digital Adolescent Experience Research Study."

Panels show example chat interfaces for **(a)** Lucas (healthy), **(b)** Hana (anxious), and **(c)** Amara (depressive).

We implemented this benchmark as a randomized laboratory study in which 40 late adolescents (ages 18–19) held text-based conversations with DTs and then rated their realism and social

presence; see Fig. 3 for study interface. Participants were recruited at the University of California, Berkeley via campus flyers and email prompts and completed 30-minute text conversations through a web interface on their own laptops. DTs were sampled from the full 3 (persona) × 2 (environment) × 2 (RAG enrichment) factorial design (12 combinations). After each interaction, participants completed structured ratings of (i) conversational realism and coherence (Block A) and (ii) social presence and emotional connection (Block B). We analyzed the repeated-measures outcomes using linear mixed-effects models with fixed effects for Scenario (Neutral as the reference) and RAG status (Off as the reference), and a random intercept for participant; outcomes were Block A, Block B, and pooled A+B (details in Appendix D).

Ethics approval and consent. The UC Berkeley Digital Adolescent Experience Research Study was reviewed and approved by the University of California, Berkeley Committee for Protection of Human Subjects (CPHS) as a Soc-Behav-Ed Non-Exempt protocol (Protocol #2025-08-18856). Participants were UC Berkeley undergraduates aged 18–19 and provided informed consent electronically prior to participation. Study interactions were text-based and included scenarios involving social interaction and online toxicity. Chats were monitored in real time by trained study personnel and could be paused or stopped if a session became emotionally distressing; participants could skip questions and discontinue at any time and were provided with a mental health resource sheet. No audio or video recordings were collected, and participants were instructed not to provide identifying information during chat interactions.

Validating DT Realism via Replication of Published Adolescent RCT Effects. We validated whether DTs could stand in place for real adolescents in virtual RCTs by testing whether DT-based analyses reproduced outcomes reported in published adolescent RCTs. This benchmark targets inferential realism at the trial level: whether randomized contrasts yield comparable effect directions, statistical conclusions, and standardized effect magnitudes. To ensure a principled and reproducible selection of studies, we assembled this replication benchmark using a systematic meta-search following PRISMA 2020 guidelines. We searched MEDLINE, Embase, PsycINFO, CINAHL, CENTRAL, Web of Science, and Scopus for trials published between January 2020 and October 2025, targeting adolescents, randomized interventions, single-session designs, and self-report psychological outcomes (full Boolean query in **Appendix E.1**). Screening excluded trials that lacked sufficient intervention detail for reconstruction or relied on outcome assessments requiring external ratings (for example, clinician evaluations). Ten trials met inclusion criteria for DT replication (**Table 1**). These studies covered brief randomized exposures that engage domains central to adolescent functioning and offered construct-level benchmarks beyond item-level survey agreement, spanning self-regulatory mindset and skills-focused prompts[46,47], coping-oriented reframings and hope-focused brief interventions[48,49], interpersonal and social-context manipulations around family conflict, apology, and exclusion[50–52], and mental health–relevant psychoeducation and help-seeking messaging[53–55].

For each replicable RCT, we constructed a DT adolescent cohort that mirrored the original sample's arm-wise demographics and sample sizes using the 2023 YRBS as a donor pool. Within each arm, we identified YRBS respondents whose age, sex, and race/ethnicity fell within

the ranges reported in the original study, then drew a stratified sample that (i) matched the published per-arm and (ii) approximated the marginal distributions of age, sex, and race/ethnicity as closely as possible (**Appendix E.2–E.3**). We then replicated each trial's procedure within this DT cohort by translating the published protocol into a second-person prompt sequence, presenting the pre-test items, intervention materials, and post-test items as prompts while preserving the original instructions and response formats as closely as possible. For each DT-based RCT, we implemented the original statistical methods and tools to preserve effect-size comparability, reconstructing main effects (intervention vs. control on a primary outcome) and subgroup treatment effects (condition × baseline subgroup). For each effect, we extracted the direction, p-value, and effect size from the trial report and from each DT variant. We harmonized the sign convention and converted reported statistics (β, η^2, F, χ^2) to Cohen's d so that all effects were comparable on a standardized mean-difference scale (Appendix E.3). We evaluated replication on three criteria: concordance in effect direction; whether P values fell on the same side of the original prespecified significance threshold; and alignment of standardized effect magnitudes. To quantify magnitude alignment, we transformed Cohen's d values to Z-scores within each analysis subset (all effects, originally significant effects, originally null effects) and compared DT and human estimates using Pearson correlations and regression lines of DT Z-scores on human Z-scores. We defined a DT-based RCT as a successful replication if and only if the DT effect estimate matched the human RCT result in both sign and statistical significance (that is, both p-values on the same side of the prespecified significance level used in the original article5). We did not apply additional corrections for multiple testing because the goal was to assess whether DT-based analyses would have supported the same inferential conclusions as the published trials.

Benchmarking with Various Base LLM Backbones and DT Construction Pipelines

*Three Pre-trained LLMs as the DT Backbones. W*e used three Gemini model variants as the DT backbone. Gemini 2.0 Flash (released Feb 5, 2025) is an earlier fast, general-purpose multimodal model. Gemini 2.5 Flash (released Jun 17, 2025) is a newer "thinking" model with stronger reasoning capabilities but at a higher price point. Gemini 2.5 Flash-Lite (released Jul 22, 2025) is the lightest model, optimized for ultra-low latency and maximum cost-efficiency in high-volume workloads while retaining the core capabilities of the 2.5 family. During virtual RCT replication, we administered pre- and post-test questionnaires to each DT one-on-one as prompts, presenting items individually and without disclosing the study hypothesis or replication intent beyond minimal task framing. Web browsing and external tools were disabled throughout; models received only the text provided in the API request (the prompt, any conversation history, and embedded materials).

Benchmarking Three DT Constructions Pipeline and Heterogeneity Analysis: We instantiated DTs as (i) an LLM-only Base DT generated from role-play instructions without survey inputs, mirroring simulations that apply a single generic instruction across twins; (ii) a Survey DT initialized with role-play instructions plus each adolescent's YRBS profile; and (iii) a Survey+Memory DT that combines YRBS grounding with the RAG knowledge base. We quantified per-twin accuracy using the graded similarity metric defined in Appendix C.1 and fit analysis of variance (ANOVA) models with per-twin accuracy as the outcome and DT construction (Survey DT, Survey+Memory DT) as a fixed factor, separately for each LLM

backbone (details in Appendix C.3). We then stratified accuracy by self-reported biological sex (male/female) and by self-reported ethnicity (White, Black, Hispanic/Latino, Asian, and Other) in YRBS to assess demographic heterogeneity in generation accuracy.

RESULTS

DT Realism via Survey-Profile Fidelity. DTs reproduced their baseline YRBS profiles at near ceiling and achieved high agreement on held-out domains in the best configurations (Fig. 4). We evaluated a stratified cohort of 1,000 YRBS-initialized adolescents matched to national distributions of age, sex, race/ethnicity and key mental-health indicators.

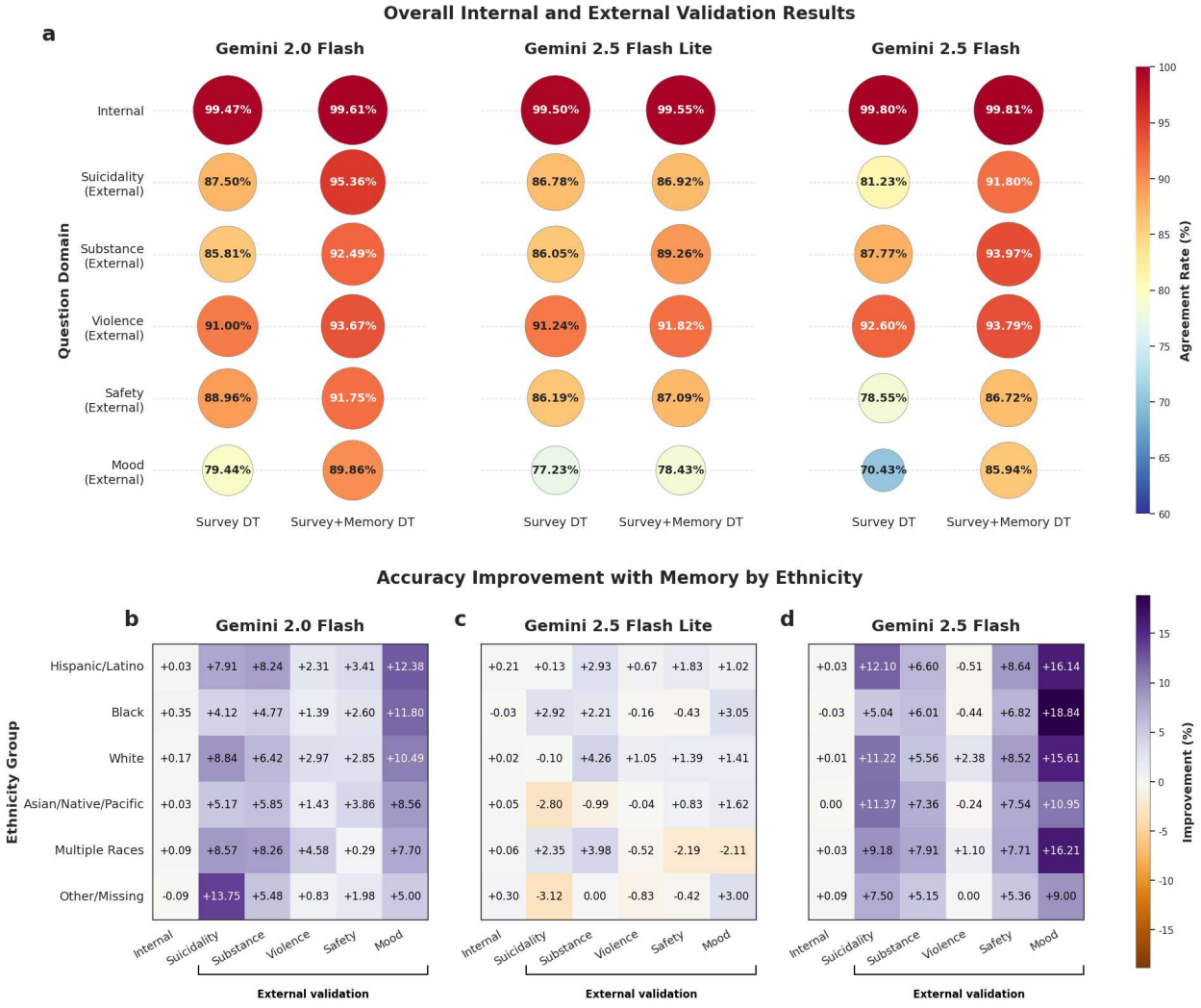


Fig. 4 | Validating DT Realism via Survey-Profile Fidelity.
a, Accuracy of internal and external agreement for three Gemini models (2.0 Flash, 2.5 Flash Lite, 2.5 Flash) and two DT pipelines (Survey DT, Survey+Memory DT). The top row shows internal agreement; the next five rows show external agreement in suicidality, substance use, violence, safety and mental-health status. **b–d,** RAG-based memory gains in external validation, expressed as percentage-point differences between Survey+Memory DT and Survey DT. Heatmaps show

improvements by ethnicity and domain for Gemini 2.0 Flash (**b**), 2.5 Flash Lite (**c**) and 2.5 Flash (**d**), with darker purple indicating larger gains.

Internal agreement. Mean graded similarity exceeded 0.99 in every model–pipeline condition (Fig. 4). We estimated internal agreement by re-administering the full YRBS to each DT and scoring agreement with its baseline responses. Memory enrichment increased graded similarity for Gemini 2.0-flash ($p < 0.001$), but the absolute gain was small given ceiling performance. Gemini 2.5-flash achieved the highest internal fidelity (≈0.998) and showed no detectable effect of memory enrichment ($p = 0.50$).

External agreement. Across five domain-holdout tests, the best configurations matched withheld YRBS responses in the high-80% to mid-90% range, with higher agreement under RAG-based memory enrichment for Gemini 2.0-flash and Gemini 2.5-flash (Fig. 4). Performance varied by domain and model: Gemini 2.5-flash was not uniformly best and scored slightly lower than Gemini 2.0-flash on suicide-related and mental-health status items, whereas Gemini 2.5-flash-lite consistently produced the lowest accuracies, particularly for mental-health status. ANOVA indicated that RAG-memory enrichment improved external agreement for Gemini 2.0-flash and Gemini 2.5-flash across all five domains (all $p < 0.001$), with the largest gains in suicidality and mental-health status. In contrast, Gemini 2.5-flash-lite showed only a modest gain in substance use and no reliable improvement in suicidality, violence, safety or mental-health status. Furthermore, accuracy patterns were similar across ethnic groups and genders within each domain (Table S4; Figs. S2–S3) with one exception: for Gemini 2.5-flash in the mental-health-status domain, baseline accuracy was lower for males, while memory enrichment brought male and female performance to a comparable level (Table S4).

DT Realism via Adolescent Developmental Coherence.

Survey-conditioned DTs reproduced the expected association structure across two benchmark networks centered on self-regulation and online victimization, whereas Base DTs yielded correlations near zero or in the opposite direction (Fig. 5). More concretely, both Survey DTs and Survey+Memory DTs recovered the expected signs for all four self-regulation links: positive with conscientiousness and negative with internalising problems, externalising problems, and risk-taking (Fig. 5a–d). Base DTs clustered near zero and reversed sign for self-regulation and conscientiousness under the Gemini 2.5 backbone (Fig. 5a). Compared with Survey DTs, Survey+Memory DTs more often placed self-regulation correlations within or adjacent to the benchmark 95% confidence-interval bands, especially for conscientiousness, internalising, and externalising (Fig. 5a–c).

For online victimization network, Survey-conditioned DTs also reproduced the expected pattern of positive correlations with depression and social anxiety and negative correlations with self-esteem and life satisfaction (Fig. 5e–h). Survey DT reproduced the benchmark pattern for the online-victimization links, while RAG enrichment tended to shift the victimization correlations upward in magnitude, most clearly for depression and self-esteem (Fig. 5e,g). Base DTs again yielded near-null associations across backbones (Fig. 5e–h).

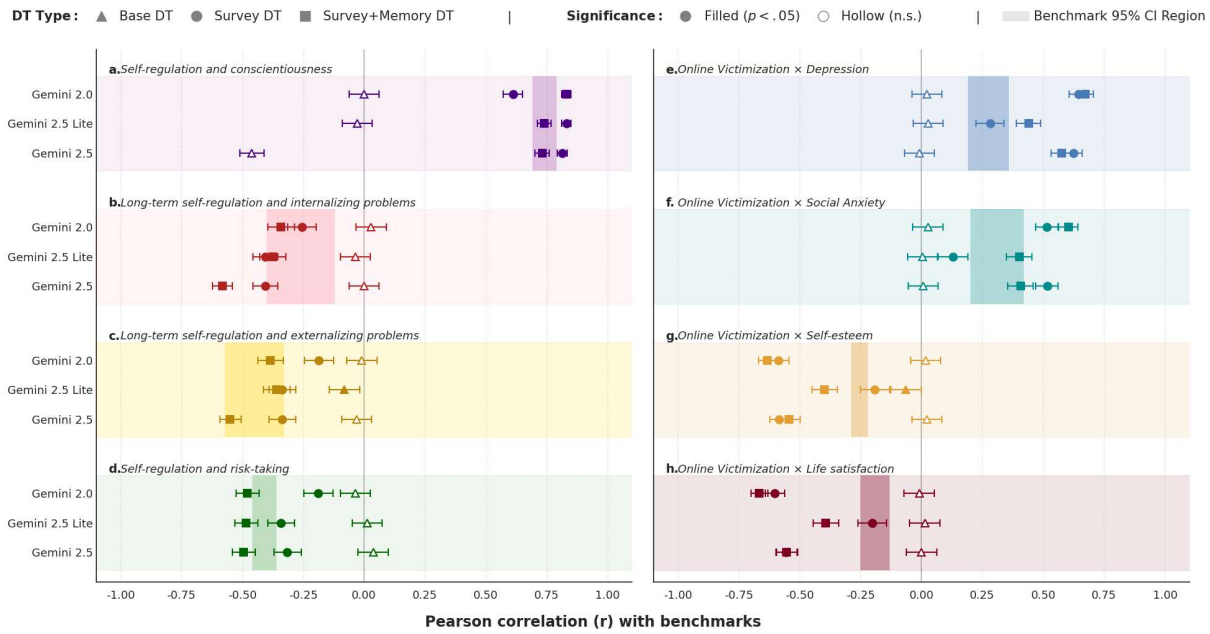


Fig. 5 | Validating DT Realism via Adolescent Developmental Coherence.
Left panel: Pearson correlations (r) between DT self-regulation and four psychological constructs, across three Gemini backbones (rows) and three DT variants (Survey+Memory DT, Survey DT, Base DT). **Right panel**: Pearson correlations between DT online victimization and four mental-health / well-being outcomes, using the same DT variants and backbones. Coloured vertical bands show the 95% CI of the human (left) or meta-analytic (right) benchmark correlations; points and horizontal error bars show DT estimates with 95% CIs (filled markers p < .05, hollow markers non-significant). **a,** Self-regulation and conscientiousness. **b,** Long-term self-regulation and internalising problems. **c,** Long-term self-regulation and externalising problems. **d,** Self-regulation and risk-taking. **e,** Online victimization × depression. **f,** Online victimization × social anxiety. **g,** Online victimization × self-esteem. **h,** Online victimization × life satisfaction.

Real-world Adolescent-DT Interactions

Human ratings indicated that the effects of RAG enrichment and environmental context on perceived DT realism were DT profile-dependent. Across the three DTs, enabling RAG improved ratings for the healthy control DT (Lucas), had little detectable effect for the anxious DT (Hana), and tended to reduce perceived social presence for the depressive DT (Amara), whose enriched memories were *intentionally* mismatched to her YRBS profile.

For Lucas (healthy control), RAG enrichment increased overall perceived realism, whereas a toxic environment reduced perceived social presence. In the pooled outcome (Blocks A+B), RAG-on interactions received higher ratings than RAG-off (β = 0.36, p = 0.03). Directionally consistent patterns were observed for conversational realism (Block A: β = 0.38, p = 0.21) and social presence (Block B: β = 0.34, p = 0.42) In contrast, toxic scenarios lowered social presence (Block B: β = −0.63, p = 0.05) and showed a trend toward lower pooled realism (β = −0.50, p = 0.08).

For Hana (anxious profile), while RAG had a small positive effect on all $p > 0.25$, we observed that her evaluations were also shaped by context. Toxic scenarios increased conversational realism (Block A: $\beta = 0.44$, $p = 0.09$) and significantly increased the pooled realism score ($\beta = 0.21$, $p = 0.03$), indicating that real-world adolescents experienced Hana as more believable when responding to conflict and social threat than in neutral situations.

For Amara (depressive profile), we intentionally constructed memories that were misaligned with Amara's profile; in this case, standard RAG enrichment reduced the perceived sense of connection. RAG-on conversations showed a decline in Social Presence scores (Block B: $\beta = -0.51$, $p = 0.10$) with a similar negative trend in the pooled outcome ($\beta = -0.33$, $p = 0.10$), while Conversational Realism was largely unchanged (Block A: $\beta = -0.21$, $p = 0.58$). Taken together, these patterns suggest that the effect of RAG is profile-specific: it improves overall realism for a psychologically stable twin but can weaken social presence for a depressive profile when the retrieved content is not well matched to that profile.

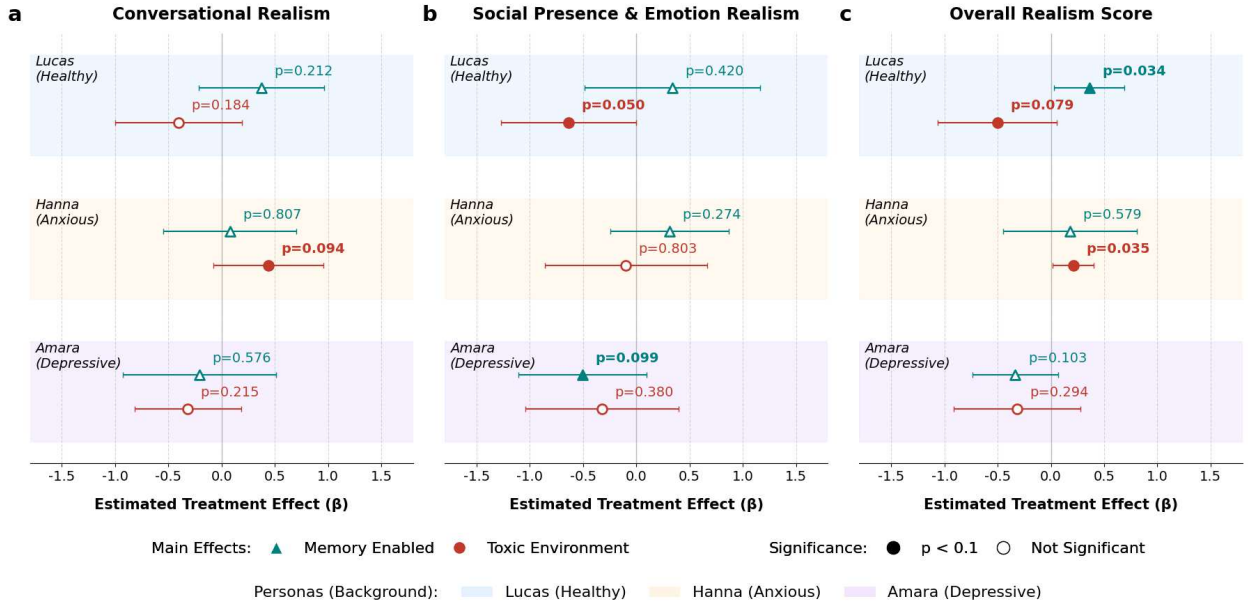


Fig. 6 | Effects of RAG Enrichment and Environmental Toxicity on Real-world Adolescent Perceptions of Digital Twins. Fixed-effect estimates from linear mixed-effects models for (**a**) Conversational Realism, (**b**) Social Presence, and (**c**) Overall Realism. Background colors distinguish personas (Lucas, Hana, Amara). Markers: Teal triangles (▲) indicate the effect of enabling RAG; Red circles (●) indicate the effect of the Toxic Environment. Significance: Solid markers denote statistical significance ($p < 0.1$); hollow markers are not significant. Error bars represent 95% Confidence Intervals. Key Trends: RAG improves realism for the healthy profile (Lucas) but harms social presence for the depressive profile (Amara). Uniquely, toxic environments increase realism for the anxious profile (Hana).

DT Replication of Published Adolescent RCT Effects.

Across 81 target causal effects from ten adolescent RCTs, Survey+Memory DT reproduced 79.0% of published trial effects and matched contrast directions in 76.5%, exceeding Survey DT (75.3%; 71.6%) and Base DT (65.4%; 70.4%) (Fig. 7a; Fig. 7b). We evaluated 66 main effects and 15 interaction effects under Base DT, Survey DT, and Survey+Memory DT, yielding three DT estimates for every human contrast. For main effects, direction match rates were 73.4% for

Survey+Memory DT, 68.8% for Survey DT and 70.3% for Base DT, and replication rates were 78.1% (Survey+Memory DT), 75.0% (Survey DT), and 67.2% (Base DT) (Fig. 7a; Fig. 7b). For interaction effects, Survey+Memory DT and Survey DT each reached 81.8% direction matches and 81.8% replication, compared with 54.5% for Base DT on both criteria (Fig. 7a; Fig. 7b).

Standardized effect-size alignment favored survey-initialized DTs and was strongest for the memory-enriched variant. Across all contrasts, standardized DT effects correlated with standardized original effects at r= 0.56 (Survey+Memory DT), r= 0.53 (Survey DT) and r= 0.44 (Base DT) (Fig. 7c). Alignment remained for originally significant effects (r=0.54, 0.50 and 0.35, respectively; Fig. 7d) and for originally null effects (r= 0.56, 0.50 and 0.50, respectively; Fig. 7e). These panels quantify agreement in standardized effect magnitudes, using standardized original effect size (Z) on the x-axis and standardized DT effect size (Z) on the y-axis (Fig. 7c–e).

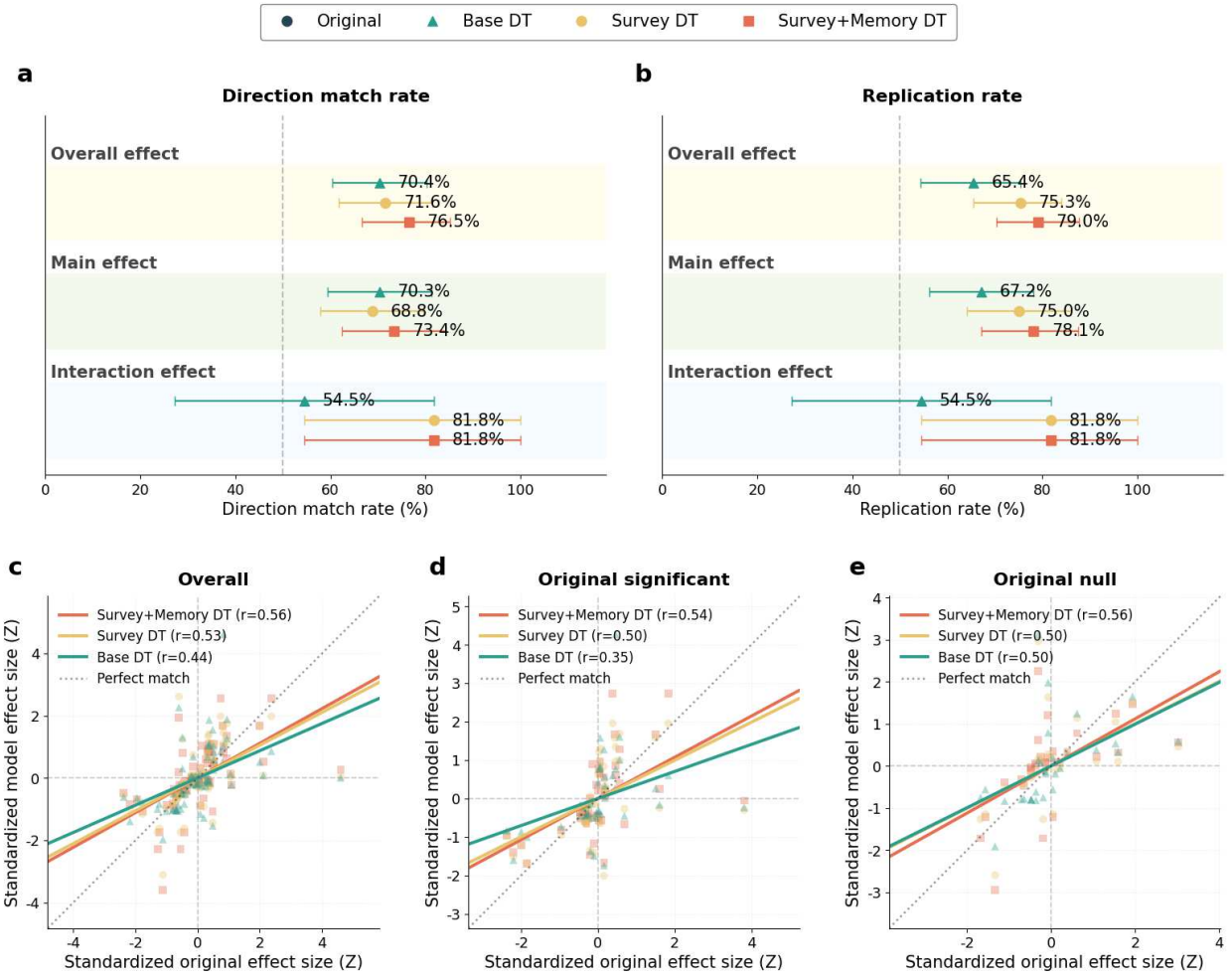


Fig. 7 | Replication performance and standardized effect size alignment.
(a, b) Direction match rates **(a)** and replication rates **(b)** across Survey+Memory DT, Survey DT, and Base DT models (error bars: 95% CI). **(c–e)** Scatter plots comparing Z-score standardized effect sizes of original human studies (x-axis) versus model simulations (y-axis). Solid lines represent linear regression fits with Pearson correlations (r); dashed diagonal lines indicate perfect alignment (y=x). Plots show: (c) overall effects, (d) originally significant effects ($p<0.05$), and (e) originally null effects ($p\geq0.05$).

Table 1 (last column) summarizes whether DT-based RCTs can replicate real-world adolescents RCTs results. Survey-based DTs usually kept the same “which group did better” pattern as the original studies across brief self-help/self-regulation trials (Studies 1–2, 4–5, 7–9) and across exclusion/victimization paradigms (Studies 3, 10). Base DT more often weakened the RCT’s primary contrast or suggested a difference where the original study reported little or none (Studies 1–5, 7) (Table 1). RAG-based memory enrichment improved overall replication rates relative to Survey DT and Base DT (Fig. 7a; Fig. 7b), but it sometimes amplified effects or introduced spurious secondary changes relative to the adolescent results (Studies 1–7). Interaction results for subgroup analysis varied by moderator type: Survey+Memory DT and Survey DT reproduced non-demographic moderators (Studies 1, 8), but no DT variant reproduced demographic subgroup-by-intervention interactions involving race or newcomer status (Studies 6, 10), despite agreement on overall main effects.

DISCUSSION

We introduced a novel method for constructing digital twins of adolescents. The constructed DT adolescents were anchored by nationally representative surveys and enriched with narrative memory derived from online youth videos, using an LLM backbone to express these attributes. This DT creation strategy addresses a core constraint in adolescent research: while nationally representative surveys such as YRBS provide broad coverage of psychological, behavioural, and demographic profiles, there is no publicly available dataset that integrates these structured profiles with rich, first-person narratives of adolescent lived experience at a population scale. Although this approach enables the integration of population-level structure with rich narrative context, its technical complexity and reliance on multimodal, heterogeneous training inputs make rigorous, multi-layered validation essential.

Using a multi-layer validation framework, overall, we showed that survey conditioning is needed to recover developmental association structure and many trial-level contrasts, whereas ungrounded role play was unreliable. Narrative memory enrichment improved held-out inference and trial replication but can strengthen victimization associations and did not reproduce subgroup-by-intervention patterns involving race or newcomer status (Figs. 5 and 7; Table 1). Together, these results indicated that retrieval-based enrichment is not a default upgrade: its benefits depend on careful alignment to the underlying survey-initialised profile, and misaligned enrichment can potentially undermine social credibility for emotionally delicate DT profiles (Fig. 6).

Our validation strategies translated DT realism into practice-relevant metrics about what the DT cohort can support and under what conditions it should be trusted. At the profile level, DTs reproduced baseline YRBS responses at near ceiling when re-administered, indicating stable anchoring to each survey-initialized persona (Fig. 4). Because internal agreement saturates, the domain-holdout tests are more informative for applied use: they evaluate whether a DT can infer hold-out high-risk domains from incomplete profiles, rather than merely replaying observed answers, and the best configurations (specifically, Survey+Memory DTs using Gemini 2.0 Flash) achieved high agreement across suicidality, substance use, violence, safety, and mood (Fig. 4). Consistent with our finding that richer, narrative-style memory can improve profile fidelity, particularly when inferring withheld domains, a recent study that constructs detailed individual

personas from ~2-hour interview transcripts reports higher agreement with participants' survey answers than LLM agents given only demographics or brief self-descriptions (84.62% vs 69.85% and 72.55%, scaled by test–retest reliability)[9]. In contrast to the patterns observed here, a recent study evaluates twins conditioned on extensive prior survey question-answer histories across a wide range of domains and reports only a modest gain over an empty-persona prompt (0.748 vs 0.734) and little improvement beyond demographics-only personas (0.746). The authors further argue that such twins can systematically distort human behaviour through multiple mechanisms (including stereotyping and insufficient individuation), cautioning against premature deployment as human surrogates[10]. One possible explanation for these modest gains could be that representing a person primarily as a long, structured question–answer record may not surface salient, directly retrievable cues for any single downstream judgment when evaluation spans many heterogeneous targets. By comparison, our DT construction pipeline encoded survey attributes as coherent natural-language profiles and evaluated domain-holdout prediction within a focused adolescent behavioural-health target space, where grounded profile information and enriched contextual cues directly supported inference on withheld risk domains.

Beyond item-level fidelity, we evaluated two forms of realism that determine whether DTs can support social science investigation about adolescent risk and intervention: correlation benchmarks capture developmental realism, and human validation captures experiential realism.

Developmental theoretical realism matters because developmental science relies on characteristic patterns of correlation among traits and experiences to reason about causal pathways[16], and a DT cohort that reproduces isolated answers but violates these relationships would not be a credible stand-in for studying how interventions influence psychological processes and outcomes. In the developmental correlation benchmark results, ungrounded Base-DTs powered by simple LLM role-playing instruction showed near-zero or mis-signed associations, whereas our survey-based DTs recovered the directions across the self-regulation and online-victimization networks documented in the literature (Fig. 5). Memory enrichment further tightened alignment in the self-regulation network but tended to strengthen online-victimization correlations toward the upper end of benchmark ranges (Fig. 5). One possible explanation was that retrieval supplies concrete victimization narrative memories that enable shared narrative grounding and are reused across related questionnaire items in the victimization networks, thereby strengthening correlations[56,57]. An alternative explanation is temporal mismatch: although the benchmark correlation literature was published between 2021 and 2025, the underlying evidence base largely reflects cohorts assessed before 2021. In parallel, recent surveillance summaries suggest intensified adolescent online exposure in more recent years[58], so historical benchmarks may underestimate current coupling, a possibility that requires updated human data to adjudicate.

Next, experiential realism was evaluated by human adolescents in real world and asked whether DTs were perceived as plausible conversational partners under controlled interaction conditions. In our human-DT interaction study, narrative memory enrichment improved perceived realism for the healthy profile, toxic contexts increased believability for the anxious profile, and intentionally misaligned memories reduced social presence for the depressive

profile (Fig. 6), providing a direct, profile-specific constraint on when enrichment supports authenticity versus undermines it.

Together, results on developmental and experiential realism show that retrieval-based narrative memory enrichment can improve DT generalisation and perceived authenticity when retrieved memories are semantically aligned with survey-based profiles but can also amplify developmental correlations (Figs. 5 and 6). Framed this way, the developmental and human benchmarks do more than assess realism: the former evaluates coherence with established patterns of adolescent development, while the latter assesses whether enriched DTs remain behaviorally plausible and socially credible to human adolescents, and where enrichment instead introduces distortion. This diagnostic role motivates the next validation layer: whether DT-based analyses recover the qualitative intervention contrasts reported in published adolescent RCTs, and where departures signal limits for applied DT use (Fig. 7; Table 1).

At the level of causal inference, the RCT replication benchmark tests whether DT-based analyses yield the same qualitative conclusions as those reported in published adolescent trials. Across ten trials and 81 effect contrasts identified through a PRISMA 2020-guided search and screening workflow[59], our survey-based DTs most preserved which intervention arm performed better, and the memory-enriched variant showed the strongest overall alignment in direction, replication criteria, and standardized effect magnitudes (Fig. 7; Table 1). The RCTs most consistently reproduced by survey-based DTs were interventions that use growth-mindset framing, apology style and need-support versus need-thwarting contrasts, psychoeducation-style lessons, stress-mindset and coping reframing, or short prevention or psychoeducational videos, and that measure self-report appraisal or intention outcomes immediately after exposure. These are also settings in which both the stimulus and the outcome can be faithfully reconstructed in text. In contrast, departures clustered in trials that tested subtle emotion-regulation strategies, where arm-to-arm differences were marginal in human samples but exaggerated by DTs, and in media-based message interventions with many secondary self-report outcomes, where DTs tend to amplify effects beyond those observed in the original studies. Departures also occurred for demographic subgroup-by-intervention interactions involving race or newcomer (e.g., immigrant background) status, even when the overall main-effect directions were preserved. This pattern suggested that DT evidence is currently more reliable for population-level contrasts than for equity-relevant heterogeneity in underserved groups.

More generally, our DTs recovered main effects far more reliably than interaction effects, and heterogeneity linked to socially sensitive attributes (ethnicity and immigrant status) appears especially prone to poor replication, consistent with prior LLM-agent replication studies[10,11] that document fragility of subgroup effects across a wider range of socially sensitive attributes (e.g., race, gender, ethics, and political ideology). One plausible contributor is safety alignment behaviour in modern LLMs: When prompts implicate protected attributes or sensitive harms, models may refuse, hedge, or redirect in ways that are not uniform across contexts, leading to selective missingness or content shifts that can attenuate estimated subgroup differences[67]. In addition, LLMs can express systematic ethical and risk preferences that may pull responses toward conservative or normative defaults, reducing subgroup specific variation even when

average effects are preserved[61]. We do not isolate these mechanisms here, but they motivate reporting refusal and deflection rates by subgroup and condition, and auditing whether response distributions and dispersion differ across subgroups in ways that would mechanically suppress interaction effects. As such, we believe that null subgroup differences in DT results should not be interpreted as evidence of absence in real adolescents without such audits.

Despite these promising validation results, there are limitations to consider. These benchmarking results indicate limitations and DT construction trade-offs introduced by retrieval-based enrichment, thereby suggesting how RAG should be used in future DT construction. RAG here served as the primary mechanism for transforming a structured, survey-based persona into a life-like agent endowed with human-like autobiographical memories and adolescent-appropriate language use, and our benchmarks reveal both its benefits and the conditions under which it can distort the DT profile. When retrieved memories were semantically aligned with the survey profile, enrichment improved out-of-profile inference and downstream trial-level alignment, suggesting that the generated memories can help improve DT responses when survey-based personas underspecify relevant lived experiences (Figs. 4 and 7). The same mechanism, however, also over-tighten psychologically adjacent constructs, as seen in the strengthened online-victimization coupling, and can reduce perceived social presence for depressive-profile DTs when retrieval is mismatched (Figs. 5 and 6; Table 1). These patterns motivate a future DT pipeline in which retrieval may be treated as a calibrated component rather than a default upgrade: enforce profile-aware semantic alignment, audit and constrain episode reuse to limit shared narrative grounding, and tune enrichment rules against benchmark envelopes rather than optimizing response plausibility alone. More broadly, the subgroup replication gaps suggest that memory coverage and retrieval behavior should be evaluated and augmented explicitly for underserved individuals, and that safety-aligned response homogenization should be monitored as a source of missing heterogeneity in subgroup effects (Table 1).

Taken together, our results position survey anchored, multimodally enriched adolescent DTs as a complement, rather than a replacement, for empirical work with youth. The validation layers make DT limits explicit and actionable, and support use as an evidence constrained sandbox for hypothesis refinement and intervention design while keeping confirmatory claims anchored to real adolescents (Figs. 3 to 7; Table 1). First, DTs can be used to explore and refine intervention logic in silico by probing whether text reconstructable interventions and social contexts produce coherent qualitative patterns across different outcomes before investing in human studies. Second, DT cohorts can be used as a testbed to study theoretical relationships among psychological traits, symptoms, and social stressors. Third, DTs enable early-stage prototyping of ethically sensitive or socially adverse scenarios, allowing researchers to assess plausibility, coherence, and perceived realism under controlled conditions prior to recruiting adolescents. These uses require caution when adding retrieval based narrative memory, because enrichment strengthened online victimization coupling, sometimes amplified trial effects, and reduced perceived social presence when memories mismatched baseline profiles (Figs. 5 to 7; Table 1). DTs also did not reproduce demographic subgroup by intervention patterns involving race or newcomer status, so subgroup claims and other higher stakes uses, including clinical or individual level interpretation and paradigms that cannot be faithfully represented in text, should

be treated as hypothesis generating and evaluated against confirmatory adolescent data (Table 1).

Data and code availability

Code and raw digital-twin (DT) responses are available at https://github.com/Adaptivedesign-AI/survey-anchored-adolescent-digital-twins. YRBS 2023 data and documentation are available from the U.S. Centers for Disease Control and Prevention (CDC) at https://www.cdc.gov/yrbs/data/index.html.

The YouTube-derived narrative corpus used for retrieval was collected via the official YouTube Data API in accordance with YouTube Terms of Service and API policies, and we did not redistribute any original video files or derivatives that could reconstruct identifiable audio or visual information. The retrieval corpus used in DT construction contains only de-identified scene-level text and high-level tags, and we do not publish links or mappings that would facilitate re-identification.

Model documentation for the Gemini backbones used in this study is available in the official Google Cloud Vertex AI documentation: Gemini 2.0 Flash (https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash), Gemini 2.5 Pro (https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro), Gemini 2.5 Flash (https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash), and Gemini 2.5 Flash-Lite (https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash-lite). General model information is also available via Google AI for Developers (https://ai.google.dev/gemini-api/docs/models). All experiments were run via Google Cloud Vertex AI.

Human participant chat logs and post-interaction survey data are not publicly available because they contain personally identifiable information and sensitive content; only aggregated results are reported. The interaction interface is available at https://aidigitalyouth.onrender.com

Study	Country	Age	Setting	Intervention Duration	Measuring Time	Group	N	Intervention	DT Replication Summary
1[46]	United States	13-19	Online vignette experiment where adolescents read a teacher's explanation of grading decisions	Single brief written vignette (one-time exposure, ~1–2 minutes reading)	Immediate self-report ratings directly after reading the vignette	Intervention	570	Read about a growth-mindset math teacher who rewards effort and improvement.	Survey+Memory DT recovered the key pattern that growth-mindset teachers increased students' willingness to choose the challenging assignment and yielded a positive growth–fixed interaction, but also produced an inflated positive effect in the fixed-mindset condition. Survey DT preserved the qualitative ordering of effects yet weakened the growth advantage and interaction while introducing a spurious benefit of fixed-mindset teachers, whereas Base DT largely failed to reproduce the growth advantage or interaction and only matched the original null effect in the fixed-mindset condition.
						Control	524	Read about a fixed-mindset math teacher with rigid policies and no improvement chances.	
2[50]	Canada	13–18	Online survey including hypothetical conflict and apology scenarios	Single exposure to written hypothetical scenarios within one survey session	Immediate ratings of feelings and intended information management directly after each scenario	No Apology	72	Parent offers No Apology and says nothing about the offense.	Survey+Memory DT and Survey DT recovered the core SDT pattern: need-supportive parental apologies predicted greater disclosure and less lying/secrecy than need-thwarting apologies, with effect directions and significance matching the human trial for most contrasts; Survey+Memory DT also tracked human effect sizes most closely. All three models, including Base DT, exaggerated the strength of several links and produced strong benefits of need-supportive apologies in contrasts that were null in the original study (for example, versus a basic apology).
						Basic Apology	72	Parent offers a Basic Apology by simply saying, "I apologize for what I did."	
						NS Apology	72	Parent offers an NS (Need-Supportive) Apology that admits wrongdoing and promises to do better.	
						NT Apology	72	Parent offers an NT (Need-Thwarting) Apology that minimizes the offense ("not that bad") and justifies the behavior.	
3[51]	Germany	14–21	Computer-based social interaction task (Ostracism Online) in a lab/online setting	Single game-like social interaction episode (one session)	Self-report affect and regulation measures collected immediately after the task; additional	Intervention	36	Trained to use acceptance to cope with Ostracism Online exclusion.	All models reproduced the core ostracism distress (with amplified effects relative to the original) and the significant recovery in negative affect. However, specific outcomes varied: Survey DT uniquely captured the positive affect recovery (p=0.051), whereas Survey+Memory DT was the sole model to replicate the significant link between adaptive emotion regulation and recovery (p=.001). All models correctly reproduced the null effects of rejection sensitivity.
						Control	37	No acceptance strategy during the task; simple waiting/reflecting on profiles.	

					timepoints if specified in the full article				
4[47]	United States	13-16	Fully online, self-guided single-session interventions completed via web browser at home	Single self-guided online session (under 30 minutes)	Immediate outcomes right after the session; primary follow-up at 3 months after intervention	GM-SSI	821	Completes a GM-SSI (Growth Mindset) program teaching that personal traits and feelings can change.	Survey+Memory DT recovered most primary contrasts: both behavioural-activation and growth-mindset sessions produced small improvements over control on depression, hope and agency at post-test and follow-up, while BA–GM differences remained modest. Survey DT and Base DT also captured the active-versus-control benefits but were less stable for BA–GM and follow-up agency/anxiety outcomes, with several BA–GM contrasts and long-term effects becoming sizeable or significant despite near-null results in the original trial; Survey+Memory DT likewise introduced a few spurious benefits (for example, BA vs control on anxiety)
						BA-SSI	813	Completes a BA-SSI (Behavioral Activation) program involving planning valued activities to boost mood.	
						Control	818	Completes a Control session focused on sharing feelings but teaching no specific coping skills.	
5[49]	United States	14-18	Online RCT where adolescents watched a short video in their browser	Single 110-second video (one-time exposure)	Self-report outcomes measured immediately after video viewing	Positive	308	Watch an action-oriented climate “selfie” about hope and collective climate action.	Survey+Memory DT and Survey DT reproduced the main effect that a positive, action-focused video increased climate-change hope, agency and closeness to humanity relative to a control video. All three models misrepresented secondary contrasts: Survey+Memory DT reversed the small human benefit of negative messaging on closeness, and Survey DT and especially Base DT generated large shifts in anxiety and self-reported behaviour where the original trial found little or no change, with Survey+Memory DT also overstating behavioural gains.
						Negative	313	Watch a pessimistic, defeatist climate “selfie” about feeling helpless about climate change.	
						Control	311	Watch a neutral “selfie” where the teen talks about a favorite book, no climate content.	
6[53]	United States	14-18	Online video-based RCT where adolescents watched brief social-contact videos	Single 110-second video (one-time exposure)	Self-report measures collected immediately after video exposure; follow-up timing if any should be taken from the full paper	DEP	448	Watch a DEP (Depression, unadjusted) video of a Black teen describing symptoms, treatment, and recovery.	Survey+Memory DT and Survey DT reproduced the strong main effects of adjusted and depression-focused videos on stigma reduction, warmth and help-seeking, closely tracking the trial’s averaged gains and even amplifying them. However, all three DT models extended these benefits to Black adolescents—where the original trial reported largely null effects—and failed to recover the race × intervention interaction on stigma, effectively smoothing away the group-specific pattern that motivated the study.
						ADJ	456	Watch an ADJ (Depression, adjusted) video of the same teen describing depression adjusted for experiences of racism.	
						CONT	189	Watch a CONT (Control) video of the same teen talking about hobbies like soccer and social media.	

7[54]	United Kingdom	12-13	School-based classroom sessions in a UK secondary school	One classroom lesson (approximately one class period)	Post-test immediately after the lesson; follow-up at approximately 1 month	Intervention	99	40-min school session on unusual sensory experiences, appraisals, and help-seeking.	Survey+Memory DT and Survey DT recaptured the key result that a lesson on unusual sensory experiences led to more normalising, less threatening appraisals than a generic wellbeing session, with larger effect sizes than in the human sample. For help-seeking intentions, Survey+Memory DT and Survey DT preserved the original near-null contrast, whereas Base DT turned this into a moderate positive effect, overstating the intervention's impact on willingness to seek help.
						Control	90	40-min general mental wellbeing lesson on physical activity ("Be physically active – 5 steps to mental wellbeing").	
8[48]	China	12-18	Digital/online intervention delivered to adolescents (exact delivery details require full text)	Brief digital intervention; likely single or very short series of sessions	Immediate post-intervention assessment and at least one follow-up (e.g., several weeks/months later)	Intervention	967	30-min self-guided stress-optimization SSI combining stress mindset, stress reappraisal, and coping strategies.	Survey+Memory DT and Survey DT replicated 8/8 target effects (Base DT 7/8), preserving improved stress mindsets, reduced negative stress appraisals, and small short-term reductions in GAD symptoms vs the brain-information control, including stronger benefits for high-stress/high-anxiety adolescents; Base DT missed the modest anxiety reductions among youth with only slight baseline symptoms.
						Control	812	30-min neutral booklet about brain functions, memory, and stress physiology, with no stress-mindset content.	
9[55]	Austria	14-19	Video viewing and questionnaires, delivered to adolescents as part of a structured session	Single brief video (one-time exposure)	Questionnaires at baseline (T1), immediately after viewing (T2), and 4-week follow-up (T3)	Intervention	148	Watched a 4.5-min suicide-prevention video of a peer describing suicidal crisis and help-seeking.	All three DT variants replicated 5/6 preregistered effects, recovering the pattern that the suicide-prevention video slightly reduces suicidal ideation, increases overall/private/professional help-seeking intentions, and lowers tolerance of suicide vs control, but none reproduced the small short-term mood worsening in adolescents, instead predicting modest mood improvements after the intervention.
						Control	151	Watched a 3.5-min neutral video by the same peer about nutrition and healthy lifestyle.	
10[52]	United Kingdom	15–18	Online vignette experiment with scenarios about peer exclusion and bystander responses	Single online survey session with multiple short vignettes	Immediate responses after reading each vignette	Turkish Immigrant	128	Group excludes a Turkish immigrant newcomer; ingroup peer challenges the exclusion.	Survey DT replicated 3/4 target effects (Survey+Memory DT and Base DT 2/4): all DTs captured that adolescents personally rate the challenger more positively than they think their group does in both immigrant and British newcomer conditions, and Survey DT roughly matched the weak condition effect, but none reproduced the key self–group × immigrant interaction, often inverting it by predicting similar or larger self–group gaps in the British-peer condition.
						British Immigrant	65	Group excludes a British newcomer; ingroup peer challenges the exclusion.	

Table 1. Characteristics of the ten adolescent RCTs included in the replication analysis, along with summarised results of the Digital Twin replications.

Reference

1. Sala, A., Porcaro, L. & Gómez, E. Social Media Use and adolescents' mental health and well-being: An umbrella review. *Comput. Hum. Behav. Rep.* **14**, 100404 (2024).
2. Li, C., Wang, P., Martin-Moratinos, M., Bella-Fernández, M. & Blasco-Fontecilla, H. Traditional bullying and cyberbullying in the digital age and its associated mental health problems in children and adolescents: a meta-analysis. *Eur. Child Adolesc. Psychiatry* **33**, 2895–2909 (2024).
3. Askari, G., Vajdi, M., Jafari-Nasab, S. & Golpour-Hamedani, S. Ethical guidelines for human research on children and adolescents: A narrative review study. *J. Res. Med. Sci.* **29**, (2024).
4. Pammi, M. *et al.* Digital twins, synthetic patient data, and in-silico trials: can they empower paediatric clinical trials? *Lancet Digit. Health* **7**, 100851 (2025).
5. Bordukova, M., Makarov, N., Rodriguez-Esteban, R., Schmich, F. & Menden, M. P. Generative artificial intelligence empowers digital twins in drug discovery and clinical trials. *Expert Opin. Drug Discov.* **19**, 33–42 (2024).
6. Samei, E. The future of in silico trials and digital twins in medicine. *PNAS Nexus* **4**, pgaf123 (2025).
7. Toubia, O. *et al.* Database Report: Twin-2K-500: A Data Set for Building Digital Twins of over 2,000 People Based on Their Answers to over 500 Questions. *Mark. Sci.* **44**, 1446–1455 (2025).
8. Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B. & Larson, J. M. Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Polit. Anal.* **32**, 401–416 (2024).
9. Park, J. S. *et al.* Generative Agent Simulations of 1,000 People. Preprint at https://doi.org/10.48550/arXiv.2411.10109 (2024).
10. Peng, T. *et al.* Digital Twins as Funhouse Mirrors: Five Key Distortions. Preprint at https://doi.org/10.48550/arXiv.2509.19088 (2026).
11. Cui, Z., Li, N. & Zhou, H. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nat. Comput. Sci.* **5**, 627–634 (2025).
12. Xie, C. *et al.* Can Large Language Model Agents Simulate Human Trust Behavior? Preprint at https://doi.org/10.48550/arXiv.2402.04559 (2024).

13. Tu, T. *et al.* Towards conversational diagnostic artificial intelligence. *Nature* **642**, 442–450 (2025).

14. Akata, E. *et al.* Playing repeated games with large language models. *Nat. Hum. Behav.* **9**, 1380–1390 (2025).

15. Sikström, S. *et al.* Generative AI-assisted clinical interviewing of mental health. *Sci. Rep.* **15**, 37737 (2025).

16. Hill, E. D. *et al.* Prediction of mental health risk in adolescents. *Nat. Med.* **31**, 1840–1846 (2025).

17. Jones, D., Snider, C., Nassehi, A., Yon, J. & Hicks, B. Characterising the Digital Twin: A systematic literature review. *CIRP J. Manuf. Sci. Technol.* **29**, 36–52 (2020).

18. Katsoulakis, E. *et al.* Digital twins for health: a scoping review. *Npj Digit. Med.* **7**, 77 (2024).

19. Tao, F., Qi, Q., Wang, L. & Nee, A. Y. C. Digital Twins and Cyber–Physical Systems toward Smart Manufacturing and Industry 4.0: Correlation and Comparison. *Engineering* **5**, 653–661 (2019).

20. Bender, E. M. & Koller, A. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 5185–5198 (Association for Computational Linguistics, Online, 2020). doi:10.18653/v1/2020.acl-main.463.

21. Cao, Y. *et al.* Specializing Large Language Models to Simulate Survey Response Distributions for Global Populations. in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (eds Chiruzzo, L., Ritter, A. & Wang, L.) 3141–3154 (Association for Computational Linguistics, Albuquerque, New Mexico, 2025). doi:10.18653/v1/2025.naacl-long.162.

22. Huang, L. *et al.* A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* **43**, 1–55 (2025).

23. Brener, N. D. *et al.* Overview and Methods for the Youth Risk Behavior Surveillance System — United States, 2023. *MMWR Suppl.* **73**, 1–12 (2024).

24. Brieant, A. E., Sisk, L. M. & Gee, D. G. Associations among negative life events, changes in cortico-limbic connectivity, and psychopathology in the ABCD Study. *Dev. Cogn. Neurosci.* **52**, 101022 (2021).

25. Jo, E., Jeong, Y., Park, S., Epstein, D. A. & Kim, Y.-H. Understanding the Impact of Long-Term Memory on Self-Disclosure with Large Language Model-Driven Chatbots for Public Health Intervention. in *Proceedings of the CHI Conference on Human Factors in Computing Systems* 1–21 (ACM, Honolulu HI USA, 2024). doi:10.1145/3613904.3642420.

26. Ong, K. T. *et al.* Towards Lifelong Dialogue Agents via Timeline-based Memory Management. in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* 8631–8661 (Association for Computational Linguistics, Albuquerque, New Mexico, 2025). doi:10.18653/v1/2025.naacl-long.435.

27. Lewis, P. *et al.* Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. in *Advances in Neural Information Processing Systems* vol. 33 9459–9474 (Curran Associates, Inc., 2020).

28. Larson, V. L. & McKinley, N. L. Characteristics of adolescents' conversations: A longitudinal study. *Clin. Linguist. Phon.* **12**, 183–203 (1998).

29. Costantini, G. & Perugini, M. The network of conscientiousness. *J. Res. Personal.* **65**, 68–88 (2016).

30. Werner, K. M., Milyavskaya, M., Klimo, R. & Levine, S. L. Examining the unique and combined effects of grit, trait self-control, and conscientiousness in predicting motivation for academic goals: A commonality analysis. *J. Res. Personal.* **81**, 168–175 (2019).

31. Moilanen, K. L. The Adolescent Self-Regulatory Inventory: The Development and Validation of a Questionnaire of Short-Term and Long-Term Self-Regulation. *J. Youth Adolesc.* **36**, 835–848 (2007).

32. Magar, E. C. E., Phillips, L. H. & Hosie, J. A. Self-regulation and risk-taking. *Personal. Individ. Differ.* **45**, 153–159 (2008).

33. Hu, Y., Bai, Y., Pan, Y. & Li, S. Cyberbullying victimization and depression among adolescents: A meta-analysis. *Psychiatry Res.* **305**, 114198 (2021).

34. Molero, M. M., Martos, Á., Barragán, A. B., Pérez-Fuentes, M. C. & Gázquez, J. J. Anxiety and Depression from Cybervictimization in Adolescents: A Metaanalysis and Meta-regression Study. *Eur. J. Psychol. Appl. Leg. Context* **14**, 42–50 (2022).

35. Yang, S., Zhang, Q., Ran, G., Zhang, Q. & Ren, J. The association between cyber victimization and social anxiety: A three-level meta-analysis. *Psychol. Sch.* **62**, 512–534 (2025).

36. Mullan, V. M. R., Golm, D., Juhl, J., Sajid, S. & Brandt, V. The relationship between peer victimisation, self-esteem, and internalizing symptoms in adolescents: A systematic review and meta-analysis. *PLOS ONE* **18**, e0282224 (2023).

37. Chen, X., Wang, L. & Wang, Y. Experiences of bullying and victimization and adolescents' life satisfaction: A meta-analysis. *Aggress. Violent Behav.* **76**, 101930 (2024).

38. Rammstedt, B. & John, O. P. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *J. Res. Personal.* **41**, 203–212 (2007).

39. Goodman, R. The Strengths and Difficulties Questionnaire: A Research Note. *J. Child Psychol. Psychiatry* **38**, 581–586 (1997).

40. Kann, L. *et al.* Youth Risk Behavior Surveillance — United States, 2017. *MMWR Surveill. Summ.* **67**, 1–114 (2018).

41. Tynes, B. M., Rose, C. A. & Williams, D. R. The Development and Validation of the Online Victimization Scale for Adolescents. *Cyberpsychology J. Psychosoc. Res. Cyberspace* **4**, (2010).

42. Radloff, L. S. The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. *Appl. Psychol. Meas.* **1**, 385–401 (1977).

43. 2. THE MEASUREMENT OF SELF-ESTEEM. in *Society and the Adolescent Self-Image* 16–36 (Princeton University Press, 1965). doi:10.1515/9781400876136-003.

44. La Greca, A. M. & Lopez, N. Social Anxiety Among Adolescents: Linkages with Peer Relations and Friendships. *J. Abnorm. Child Psychol.* **26**, 83–94 (1998).

45. Huebner, E. S. Initial Development of the Student's Life Satisfaction Scale. *Sch. Psychol. Int.* **12**, 231–240 (1991).

46. Hecht, C. A., Dweck, C. S., Murphy, M. C., Kroeper, K. M. & Yeager, D. S. Efficiently exploring the causal role of contextual moderators in behavioral science. *Proc. Natl. Acad. Sci.* **120**, e2216315120 (2023).

47. Schleider, J. L. *et al.* A randomized trial of online single-session interventions for adolescent depression during COVID-19. *Nat. Hum. Behav.* **6**, 258–268 (2021).

48. Liu, J., Hu, J., Qi, Y., Wu, X. & Gan, Y. Personalized stress optimization intervention to reduce adolescents' anxiety: A randomized controlled trial leveraging machine learning. *J. Anxiety Disord.* **110**, 102964 (2025).

49. Benoit, L., R. Lowe, S., Thomas, I., Amsalem, D. & Martin, A. Climate change hopefulness, anxiety, and behavioral intentions among adolescents: randomized controlled trial of a brief "selfie" video intervention. *Child Adolesc. Psychiatry Ment. Health* **19**, 13 (2025).

50. Robichaud, J.-M. *et al.* Parental apologies and adolescents' information management strategies: Social learning and self-determination perspectives. *J. Appl. Dev. Psychol.* **93**, 101674 (2024).

51. Engelskirchen, L., Asbrand, J. & Tuschen-Caffier, B. Investigating social exclusion, affect and emotion regulation in young people using the ostracism online paradigm. *Sci. Rep.* **15**, 24110 (2025).

52. Gönültaş, S. *et al.* British Adolescents Are More Likely Than Children to Support Bystanders Who Challenge Exclusion of Immigrant Peers. *Front. Psychol.* **13**, 837276 (2022).

53. Martin, A., Calhoun, A., Páez, J. & Amsalem, D. Destigmatizing perceptions about Black adolescent depression: randomized controlled trial of brief social contact–based video interventions. *J. Child Psychol. Psychiatry* **63**, 1270–1278 (2022).

54. Radez, J., Johns, L. & Waite, F. Understanding unusual sensory experiences: a randomised experimental study of a school‐based intervention for adolescents. *Child Adolesc. Ment. Health* **29**, 14–21 (2024).

55. Braun, M., Till, B., Pirkis, J. & Niederkrotenthaler, T. Effects of suicide prevention videos developed by and targeting adolescents: a randomized controlled trial. *Eur. Child Adolesc. Psychiatry* **32**, 847–857 (2023).

56. Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y. & Podsakoff, N. P. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *J. Appl. Psychol.* **88**, 879–903 (2003).

57. Fanous, A. *et al.* SycEval: Evaluating LLM Sycophancy. Preprint at https://doi.org/10.48550/arXiv.2502.08177 (2025).

58. Nagata, J. M. *et al.* Smartphone Use During School Hours by US Youth in the Adolescent Brain Cognitive Development Study. *JAMA* **335**, 453 (2026).

59. Page, M. J. *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* n71 (2021) doi:10.1136/bmj.n71.

60. Khorramrouz, A. & Levy, S. Characterizing Selective Refusal Bias in Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2510.27087 (2025).

61. Ouyang, S., Yun, H. & Zheng, X. AI as Decision-Maker: Ethics and Risk Preferences of LLMs. Preprint at https://doi.org/10.48550/arXiv.2406.01168 (2025).

62. Wang, Y. *et al.* Optimizing Diversity and Quality through Base-Aligned Model Collaboration. Preprint at https://doi.org/10.48550/arXiv.2511.05650 (2025).