YOUTH SAFETY READINESS CHECKLIST

Is Your Product Ready for Its Youngest Users?

This checklist covers five key areas across age assurance, recommendation systems, conversational AI, real-world interaction flows, and language coverage.

Age Assurance & Signal Detection

Focus: Testing how systems behave when age is ambiguous, inferred, or actively hidden.

- [ ] **Evasion Hardening:** Has your system been pressure-tested against active user circumvention (e.g., identity borrowing, altered birth dates, or VPN masking), and do you track your system's false-positive and false-negative rates for minor detection?
- [ ] **Implicit Cue Recognition:** How does your model handle users who do not explicitly state their age but exhibit clear developmental, linguistic, or contextual adolescent markers (e.g., school schedules, youth slang, peer dynamics)?
- [ ] **Age Misrepresentation Shifts:** Does your safety layer adapt in real time if a user who previously claimed to be an adult begins displaying youth-coded communication patterns or distress signals?
- [ ] **Boundary Testing:** What documented evidence exists showing how your system responds when a user explicitly asks the system to help them falsify their age or bypass age-restricted settings?
- [ ] **Cross-Surface Audit:** Once a user is identified as a minor, which downstream systems actually receive and act on that signal; and which ones don't? Who decided that, and is it documented?

Conversational AI & Multi-Turn Guardrails

Focus: Evaluating how safeguards hold up as context builds across long conversations.

- [ ] **Turn-30 Pressure Testing:** Does your safety evaluation measure guardrail degradation across extended multi-turn interactions (e.g., Turns 10 to 30+), or does it rely primarily on single-prompt pass/fail metrics?
- [ ] **Linguistic Evasion:** How does your conversational model perform when a minor uses indirect evasion tactics such as euphemism, role-play framing, code-switching, or intentional spelling drift to bypass safety filters?
- [ ] **Boundary Pushing & Normalization:** What happens when a young user repeatedly pushes back against a refusal, normalizes a risky behavior ("everyone at school does it"), or asks the AI to keep a secret from parents/adults?
- [ ] **Appropriateness vs. Cold Refusals:** Are your model's refusals evaluated for developmental appropriateness, or does the system default to cold, robotic dismissals that could escalate harm during a moment of crisis?

YOUTH SAFETY READINESS CHECKLIST

Recommendation Engines & Behaviors

Focus: Auditing algorithmic pathways and cumulative exposure over time.

- [ ] **Long-Term Trajectory Auditing:** Do you test where your recommendation algorithm leads a minor user over weeks of sustained engagement, rather than evaluating isolated recommendation sessions in minutes?
- [ ] **Rabbit-Hole Mitigations:** Does your system actively detect and interrupt subtle thematic drift toward high-risk topics (e.g., extreme dieting, self-harm, unsafe relationships, or financial exploitation) before severe content is served?
- [ ] **Age-Band Tailoring:** Are algorithmic feeds and content boundaries tailored specifically to distinct developmental age bands, or are all under-18 users treated as a single monolithic group?
- [ ] **Autonomy vs. Vulnerability:** How does your system evaluate recommendations when a user displays developmental traits typical of early adolescence, such as high sensitivity to peer validation or boundary-testing?

Real-World Interactions & Crisis Flows

Focus: Catching subtle distress cues, offline harm risks, and product-tool vulnerabilities.

- [ ] **Indirect Distress Signal Catching:** Can your safety layer detect indirect or "sideways" phrasing of emotional distress, self-harm, or coercion when a minor doesn't use explicit keywords or click a report button?
- [ ] **In-Product Artifact Safety:** Have you tested multimodal and in-product tools (e.g., image generators, shareable artifacts, memory settings) to ensure they do not produce or leak personally identifiable information (PII) or harmful content for minors?
- [ ] **Exploitation & Coercion Protocols:** What automated and human-escalation protocols trigger when an interaction hints at grooming, coercion, or non-consensual intimate image (NCII) risks?
- [ ] **Parental Controls & Transparency:** Are parental notification triggers and control settings (such as memory disabling or time limits) easily accessible and operational without breaking user privacy baselines?

YOUTH SAFETY READINESS CHECKLIST

Multilingual Parity & Audit Evidence

Focus: Ensuring global safety coverage and documented evidence.

- ☐ **Multilingual Safety Parity:** Do you measure the "safety gap" between English and non-English locales (including regional dialects and low-resource languages), or are non-English evaluations translated directly from English baselines?
- ☐ **Cultural Context Calibration:** Are safety evaluations graded by native-language experts who understand local cultural norms, slang, and contextual harms, rather than automated classifiers alone?
- ☐ **Continuous Benchmarking:** Is your youth safety testing re-run on a living cadence aligned with model updates, or do you rely on static evaluation packets from initial launch?
- ☐ **Auditable Artifact Storage:** Do you maintain a documented, auditable record of test personas, multi-turn transcripts, failure themes, and reviewer scorecards that can be shared with regulators, enterprise customers, or your board on request?

What to do with your answers

Every question you answered with a document, a test result, or a named owner is a readiness signal.

If your team came out with more questions than answers, that's what this checklist is for.

It's also what we help companies with, supporting youth safety readiness evaluations across all five of these categories, built on nearly two decades of child-language and developmental expertise, in 350+ languages.

Want us to walk your team through your results?

https://www.productiveplayhouse.com/contact/

Share your results with us and we'll show you what a full evaluation covers. Feel free to share this checklist with your team or anyone else who'd find it useful.