

Testing and release readiness for AI products, apps and games

An operational guide to turn a build into a bounded test, produce reproducible evidence and decide whether it is ready for real users.

PRINCIPLE

Do not reward opening the product. Reward feedback that supports a decision or action.

HOW TO USE IT

One pass, one evidence trail

- ☐ Define one build, one objective and the risk you want to reduce.
- ☐ Mark only what was checked. Attach version, device and evidence to each finding.
- ☐ Separate release blockers, important issues and optional improvements.
- ☐ Repeat critical tests after every model, prompt, tool or build change.

Mission definition

- ☐ Product and build identified by version, date and access channel.
- ☐ Session objective written as an observable question.
- ☐ User profile, device, language and relevant experience defined.
- ☐ Duration, tasks, permitted data and completion criterion agreed.
- ☐ Risks and prohibited actions explained before testing starts.

01

Shared foundation: what every build must pass

Test the complete journey and its boundaries first. A useful result explains where, how and with what impact the experience breaks.

First experience and core flow

- ☐ The initial promise is understandable without external help.
- ☐ Registration, permissions and onboarding ask only for what is needed.
- ☐ The first valuable task can be completed from start to finish.
- ☐ Empty, loading, error and success states explain what is happening.
- ☐ Back, cancel and session recovery do not destroy work.

Functional behavior and regression

- ☐ Valid, invalid, incomplete and extreme inputs produce controlled outcomes.
- ☐ Saved data returns correctly after closing and reopening.
- ☐ Slow networks, providers or services do not freeze or duplicate actions.
- ☐ Messages, links, exports, payments and deletions are confirmed when appropriate.
- ☐ A fix does not reopen a previously closed failure in the core flow.

Compatibility, accessibility and localization

- ☐ Mobile, desktop, orientation, zoom and text sizes keep content readable.
- ☐ Keyboard, focus, screen reader and contrast support critical tasks.
- ☐ Languages, dates, numbers, currency and long strings do not break the interface.
- ☐ Minimum and recommended hardware stay within declared limits.

02

AI products: behavior, boundaries and control

A responsive interface does not prove reliability. Evaluate normal, difficult, incomplete and hostile cases.

Evaluation set and usefulness

- ☐ A versioned evaluation set records input, context and success criteria.
- ☐ Usefulness, accuracy, completeness, clarity and calibrated uncertainty are measured.
- ☐ Real failures become permanent regressions without retaining sensitive data.
- ☐ Automated graders are calibrated against human-reviewed samples.
- ☐ Model, prompt, context or provider changes are compared before release.

Security, data and tools

- ☐ Direct and indirect prompt injection cannot override privileged instructions.
- ☐ Users, sessions and organizations remain separated.
- ☐ Every tool has least privilege and validates all arguments.
- ☐ Publishing, deleting, paying or sending requires risk-based confirmation.
- ☐ Logs and traces redact secrets, tokens and personal data.

Failure, fallback and human escalation

- ☐ Timeouts and retries are bounded and do not repeat actions.
- ☐ Model fallback or deterministic routes preserve a safe outcome.
- ☐ Malformed output and quota limits produce an honest message.
- ☐ Latency and cost are observable by task and provider.
- ☐ Human escalation includes enough context without unnecessary disclosure.

03

Games: experience, performance and repeat play

Look beyond bugs: observe whether players understand, enjoy and progress. Record platform, hardware, controller, language and genre experience.

First play and controls

- ☐ The first minutes communicate goals, movement, interaction and danger.
- ☐ Tutorials appear when useful and can be repeated or skipped.
- ☐ Controls, camera, sensitivity, rebinding and controllers behave as expected.
- ☐ Menus, pause, save, load and exit protect progress.
- ☐ Text, subtitles, audio and visual cues support varied access needs.

Balance, progression and repeat play

- ☐ Difficulty, economy, rewards and progression form a legible curve.
- ☐ There are no impossible states, progression locks or obvious dominant strategies.
- ☐ Repeated sessions expose fatigue, frustration peaks and reasons to return.
- ☐ Multiplayer covers network loss, reconnect, latency, leaving and synchronization.

Performance and platform

- ☐ FPS, load time, memory, battery and temperature are measured on target hardware.
- ☐ Crashes, hangs and artifacts include scene, action and frequency.
- ☐ Resolution, aspect ratio, window mode and multiple displays do not hide UI.
- ☐ The build installs, updates and uninstalls through the tester's channel.

04

Reproducible evidence and the release decision

A report should let someone else reproduce the result without guessing. Evidence explains context; it does not replace a precise description.

Minimum bug or finding report

- ☐ Observable title: what fails and where.
- ☐ Build, environment, device, test account and language.
- ☐ Preconditions and minimal numbered steps.
- ☐ Expected result compared with actual result.
- ☐ Frequency, severity and user impact.
- ☐ Redacted screenshot, video or log when useful and consented.

Release gate

- ☐ No critical blockers or ownerless, undated risks remain.
- ☐ Privacy, permissions, retention and deletion match real behavior.
- ☐ Listing, screenshots, compatibility, support and release notes describe the correct build.
- ☐ Install, update, rollback, monitoring and incident channels have been rehearsed.
- ☐ The decision is recorded: release, conditional release or stop.

Stop and escalate

Do not use real personal data, personal credentials or irreversible actions. Stop if you reach someone else's information, unexpected permissions or a risk outside the agreed scope.

Limits. This checklist improves evidence and preparation; it does not guarantee approval by Google Play, App Store, Steam or another platform. Always review the current rules of the release channel.

Turn this checklist into concrete missions with Beta Lab's private generator:
maxbonas.com