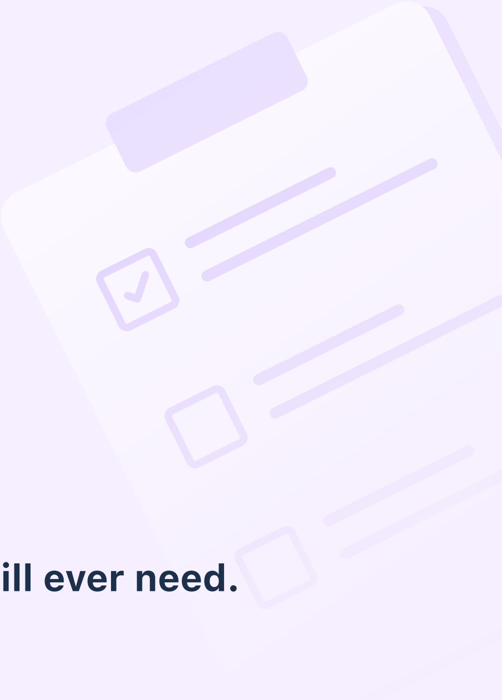

Experimentation Checklist ☑

The only checklist your experimentation program will ever need.

Experiment planning

Lock your hypothesis, metrics, and AI baselines before a single line of test config is written.

□ Hypothesis clearly defines the business outcome, user behavior change, AI component under test, and expected direction of impact.

□ Primary metric defined and configured in VWO.

□ Secondary metrics defined and configured.

□ AI-specific metrics defined: response latency, prediction confidence distribution, hallucination/error rate, recommendation accuracy, false positives/negatives.

□ Guardrail metrics set to protect revenue/conversion, UX performance, and AI output stability.

□ Baseline AI performance documented: output samples saved, accuracy benchmarks recorded, latency captured, segment performance logged.

Test setup

Document every AI artifact and assign clear ownership before the experiment is cleared for launch.

□ Model version, prompt version, and algorithm logic documented.

□ Rollback criteria defined before launch.

□ Bias and fairness checks included in evaluation plan.

□ Privacy and compliance reviewed for AI data usage.

□ Clear ownership assigned: experiment owner, analytics owner, AI reviewer, rollback decision-maker.

Pre-launch QA

Eliminate implementation error as a confounding variable before a single visitor is bucketed.

□ Experiment type selected: A/B, Split URL, Server-side, Multivariate, or AI model comparison.

□ Targeting and segmentation validated; overlap with concurrent experiments checked.

□ Test previewed across browsers and devices; dynamic content optimized for performance.

□ AI edge cases and stress prompts tested during QA.

□ Hypothesis peer-reviewed by at least two stakeholders.

□ Prioritization framework includes AI confidence or uncertainty factor.

□ Experiment stages configured in program management tool.

Running the experiment

Protect test integrity through disciplined non-interference until pre-committed stopping criteria are met.

□ No premature conclusions drawn from early data; no metric changes made mid-run.

□ No prompt changes, model swaps, or logic updates during run; if required, pause, flush data, and relaunch cleanly.

□ AI telemetry actively monitored: latency, error rate, unsafe outputs, and behavioral anomalies.

□ All variations receiving sufficient traffic; traffic split confirms statistical power.

□ Guardrail metrics monitored daily.

□ Test running minimum of 7 days; drift detection monitored for AI systems.

Conclusion

Ship only when the evaluation is complete, unbiased, and strictly aligned to your pre-defined hypothesis metrics.

- □ Test not concluded before minimum runtime unless a guardrail breach occurred.
- □ Evaluation based on probability to beat baseline, statistical significance, effect size, segment-level performance, and AI output distribution differences.
- □ Evaluation strictly aligned to metrics defined in the original hypothesis.
- □ Bias and fairness review completed before shipping.
- □ AI stability confirmed before rollout.
- □ Winning variation shipped; losing variations paused and documented.

Analysis & learnings

Go beyond the win/loss call, segment, quality, & feed every insight back into the next experiment cycle.

- □ All learnings documented: business outcomes, AI behavior patterns, unexpected outputs, segment insights, guardrail observations.
- □ Qualitative deep dive completed using heatmaps, session recordings, funnel analysis, and AI output sampling review.
- □ Insights fed back into prompt refinement, model retraining, feature engineering, and personalization logic updates.
- □ Next experiment planned based on this test's findings.
- □ Learnings shared with leadership and cross-functional teams.