Streaming Experimentation Readiness Checklist — One Page

Fora Soft Learn

An A/B test proves a change helped instead of trusting an opinion. Get five things right: choose the OEC (watch time and retention, not clicks), set quality-of-experience guardrails that can veto a launch, size the test for power, decide the stopping rule before you start (never peek), and pick the right method. Run visible-change tests long enough for novelty to fade. Netflix's published figures move - confirm before planning against them.

1 · CHOOSE THE OEC (what success means)

- ☐ **Optimize watch time and retention** — month-to-month retention and streaming hours, not clicks or plays started.
- ☐ **Clicks lie** — a tap that leads to a 30-second bounce is a failure dressed as a win; clickbait erodes trust.
- ☐ **Use fast proxies, validate slow** — move on engagement/streaming-hours proxies, confirm the big calls against retention.
- ☐ **Write the OEC down first** — one agreed metric before the test starts, not a dashboard of twenty read after the fact.

2 · SET GUARDRAILS (what must not get worse)

- ☐ **Quality-of-experience guardrails** — video start time, rebuffering ratio, crash rate, app errors.
- ☐ **Business guardrails** — unsubscribes, payment failures, support contacts; an OEC win can still hurt elsewhere.
- ☐ **A guardrail regression vetoes the launch** — engagement bought with a worse-quality stream is not a real win.
- ☐ **Check for sample-ratio mismatch** — if the groups didn't split as planned, the experiment is broken; stop and fix.

3 · SIZE IT & DON'T PEEK (the discipline)

- ☐ **Sample size ≈ 16 × variance / (effect size)²** — smaller effects cost quadratically more viewers and time.
- ☐ **Decide the stopping rule before you start** — fixed horizon and look once, OR a sequential test built to monitor.
- ☐ **Never stop a fixed test early** because it looks significant — peeking inflates false positives and fakes wins.
- ☐ **Use sequential testing to monitor live** — always-valid results let you kill a harmful rollout within hours.

4 · PICK THE METHOD & AVOID TRAPS

- ☐ **A/B test** — the decisive test of any change (UI, pricing, player, sign-up) against retention; needs large samples.
- ☐ **Interleaving** — compare two recommendation rankings with 100×+ fewer viewers; prune fast, then A/B-test survivors.
- ☐ **Quasi-experiment** — region/time before-after when you can't randomize (CDN swap, pricing); weaker, the fallback.
- ☐ **Beat the novelty effect** — run visible-change tests long enough for the spike to fade; CUPED can halve viewers needed.

THE ORDER OF OPERATIONS — DECIDE THE METRIC, SET GUARDRAILS, SIZE IT, FIX THE STOPPING RULE, THEN RUN

At streaming scale a one-percent improvement is millions of hours, so you cannot afford to guess which changes help - and a controlled experiment is the only reliable way to tell a real improvement from noise. Do it in order. First, choose the Overall Evaluation Criterion (OEC): the single outcome that defines success, anchored on watch time and month-to-month retention, never clicks or plays started, because a change tuned for clicks wins on clicks and loses on the watch time that pays the bills. Second, set guardrail metrics that must not get worse - video start time, rebuffering ratio, crash rate, unsubscribes - and let any guardrail regression veto the launch even on an OEC win, because engagement bought with a worse-quality stream is not a real win. Third, size the test: the viewers you need scale roughly as 16 times the metric's variance divided by the square of the effect you want to detect, so a smaller target effect costs quadratically more sample and time, and variance reduction such as CUPED can roughly halve it. Fourth, decide the stopping rule before the test starts - either fix the duration and look once at the planned end, or adopt a sequential test designed for continuous monitoring - and never stop a fixed-horizon test early just because it crossed the significance line, because that peeking manufactures false wins; sequential testing is what lets you watch a rollout live and kill a harmful change within hours. Finally, pick the method to match the change: a classic A/B test for the decisive measurement against retention, interleaving to compare recommendation rankings with over 100 times fewer viewers as a fast first-stage filter, and a quasi-experiment as the fallback when you cannot randomize at the individual viewer. Run visible-change experiments long enough for the novelty effect to fade. Netflix's published figures and platforms change - re-verify before you plan against them.