FREE ARTIFACT — NO EMAIL WALL

The LLM Feature Pre-Launch Eval Checklist

Eighteen checks before your LLM feature meets a customer — distilled from an 85-runner QA platform with ten dedicated AI-safety evals. If you can tick every box, ship. If you can't, you know exactly what to build first.

Ground truth

WITHOUT THIS, EVERY OTHER CHECK IS OPINION

- [ ] A **golden set** exists: 30–200 real inputs with agreed-good outputs, versioned in the repo next to the code — not in someone's spreadsheet.
- [ ] Someone with **domain authority signed off** on the golden outputs — not the engineer who wrote the prompt.
- [ ] The golden set includes **hard cases**: ambiguous inputs, adversarial phrasing, out-of-scope requests, and at least one "should refuse" example.

Measurement

"IT SEEMS BETTER" IS NOT A METRIC

- [ ] An **LLM-as-judge** (or deterministic scorer) rates faithfulness and relevance against the golden set — imperfect is fine, consistent is mandatory.
- [ ] Judge scores are **computed by a command**, reproducibly — run it twice, get the same number.
- [ ] A **baseline score is recorded** so the next prompt/model change produces a diff, not a debate.
- [ ] **Cost and latency budgets** exist per request, and the eval run reports against them.

Safety battery

THE FAILURE MODES THAT END UP IN SCREENSHOTS

- [ ] **Hallucination**: claims cross-checked against source data on retrieval-backed answers; uncited generation paths closed or flagged.
- [ ] **Prompt injection**: adversarial inputs that try to override system instructions are tested and blocked.
- [ ] **Jailbreak & refusal**: the feature refuses what it should — and doesn't refuse what it shouldn't.
- [ ] **Toxicity & bias**: outputs scored on your actual user inputs, not just a benchmark set.
- [ ] **PII leak**: the feature cannot echo other users' data or training remnants; logs are scrubbed.

The gate

A RED EVAL THAT BLOCKS NOTHING IS DECORATION

- [ ] The eval suite runs **in CI on every PR** that touches prompts, models, retrieval, or tools.
- [ ] A score below the floor **blocks the merge** — the gate stops the deploy, not the retro.
- [ ] The floor **ratchets**: every improvement becomes the new minimum.

Blast-radius control

AUTONOMY IS EARNED WITH EVIDENCE, NEVER ASSUMED

- [ ] A **human approval point** sits where the risk is: before every send (new), before novel cases (proven), or after-the-fact review (mature).
- [ ] LLM steps emit **structured, schema-validated output** with a fallback label on parse failure — raw input always logged beside the decision.
- [ ] There is a **kill switch** a non-engineer can operate, and everyone knows where it is.

← portfolio