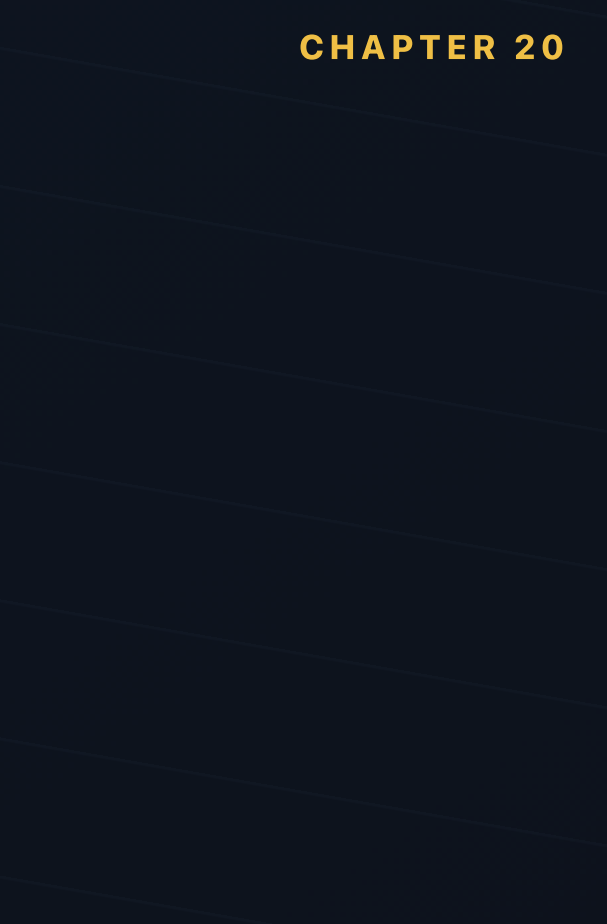
THE PRACTICAL PLAYBOOK

A practical system for shipping AI with confidence.

Define quality, collect evidence, make the release decision, learn from production, and keep the loop alive.

From Chapter 20: The Practical Playbook



THE QUALITY LOOP

Start small. Make the evidence real.

Observe

Find a real output difference or failure.

Preserve

Save the input, route, context, and trace.

Compare

Known-good versus known-bad; old versus new.

Inspect

Token, attention, activation, or feature drift.

Replay

Confirm with behavioral evals and intervention.

Internal signals earn trust only when they help detect, localize, or prevent consequential failures.

TEST THE CONVERSATION, NOT ONLY THE ANSWER

A chatbot must handle grounding, safety, tone, memory, escalation, and recovery.

Prompt

Raw input, normalized input, template, user state, and instructions.

Context

Retrieved evidence, memory, tool results, permissions, and truncation.

Route

Model and tokenizer versions, generation settings, tools, and filters.

A “reasoning” failure may really be a preprocessing, retrieval, token-budget, or prompt-assembly failure.

ONE WORKFLOW, MANY DECISIONS

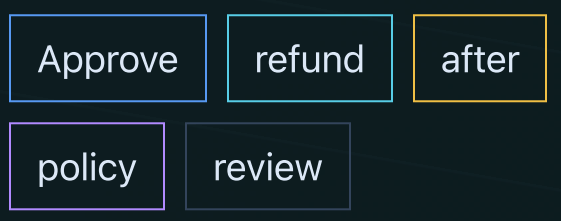
Each turn changes context, confidence, permissions, and risk.

Visible prompt

“Approve refund after policy review.”

Names, punctuation, code, accents, dates, emoji, and long documents can split or truncate in surprising ways.

Example decoded tokens



One possible tokenizer output. Token boundaries vary by tokenizer and model.

Preserve token IDs, positions, byte spans, tokenizer version, and the truncation boundary.

GOVERNANCE IS OPERATIONAL

Ownership, decision rights, audit trails, and escalation paths must exist before the incident.

Compress

Attention tensors become links, token summaries, and selected layer views.

Compare

Look at a safe case and a nearby unsafe case, or model A and B.

Question

Did a policy, negation, number, or constraint stop relating to evidence?

Replay

Confirm the suspected change with behavioral evals.

Attention is comparative triage evidence, not a pass/fail score or causal proof.

FAILURE TAXONOMY BEATS BUG FOLKLORE

A shared failure language turns scattered examples into patterns you can fix.

What can mislead

- Special tokens can dominate many heads
- Formatting can create strong but irrelevant links
- One vivid graph invites an easy story
- Similar outputs can arise from different patterns

What helps

- Document the aggregation and filters
- Compare the same positions across versions
- Look for reproducible behavior-linked changes
- Keep the full tensor for follow-up work

A useful attention pattern repeatedly predicts real behavioral differences. An intuitive picture is not enough.

AI ALWAYS FAILS

The question is where, how often, how badly, and whether the system can fail safely.

Early	Middle	Late	Compare	Verify
Predominantly local token relationships: position, punctuation, and short-range dependencies.	Broader contextual relationships and longer-range integration.	Stronger focus on answer-relevant context.	Same token positions across representative layers.	Connect visual differences back to output behavior.

Attention snapshots are illustrative diagnostics. They do not establish a universal sequence of “reasoning layers.”

PERFORMANCE IS QUALITY

Latency, cost, reliability, and tail behavior are product quality, not afterthoughts.

Overclaim

“This neuron contains the concept of privacy.”

- Raw neurons can be polysemantic
- Concepts can be distributed
- Labels come from the investigator
- A vivid spike can be noise

Probe carefully

Treat an activation or feature as a classifier that needs validation.

- Curated positives and negatives
- False positives and false negatives
- Stability across versions and slices
- Behavioral intervention checks

The useful question is whether the signal improves triage, detection, or review prioritization.

MINIMUM VIABLE AI QUALITY SYSTEM

Start with a few high-value cases, repeat runs, useful traces, a release gate, and owners.

Raw neuron

May respond to credential text, syntax, and unrelated patterns.

→

Entangled activity

One feature can live across units; one unit can support several features.

→

SAE feature

Attempts to separate sparse, more inspectable patterns.

→

Validate

Test the feature dictionary like another learned system.

Sparse autoencoders can help inspection, but they introduce new thresholds, omissions, training data, and instability.

MAKE TESTING INTERESTING

People need to read, understand, and act on failure evidence. Make quality work worth paying attention to.

Behavior

Outputs and actions meet product criteria. This remains the foundation.

System traces

Prompts, data, tools, routes, permissions, and policies participated.

Internal signals

Attention and activation drift tell teams where deeper testing may pay off.

Version internal probes with models, prompts, tokenizers, thresholds, datasets, and their documented limits.

THE PRACTICAL MOVE

Build the smallest evidence system that can stop the wrong release.

AI quality becomes practical when teams own the evidence, the release decision, and the learning loop. IcebergQA helps teams put that system in place.

Talk to IcebergQA