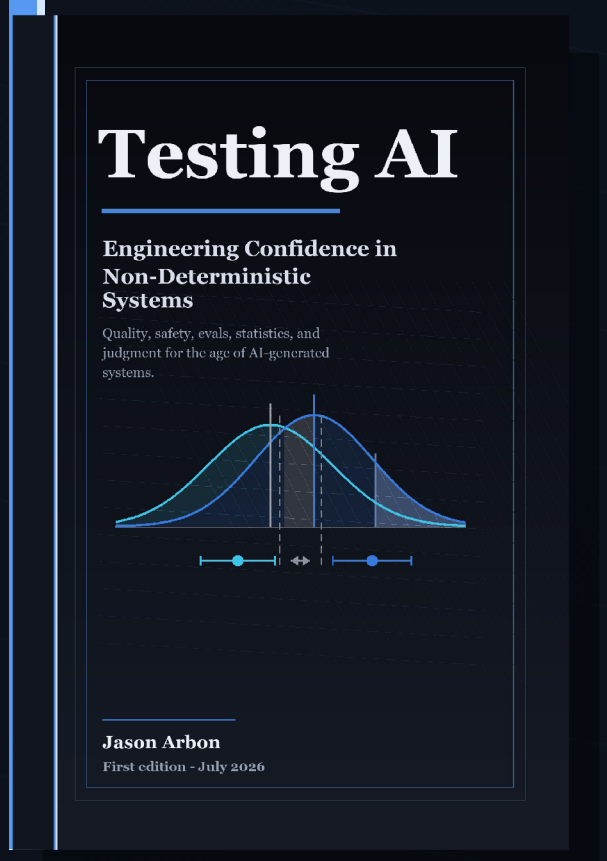
PERSONALIZED AND DYNAMIC AI PRODUCTS

When AI knows you, quality becomes personal.

The system can adapt to history, memory, identity, device, accessibility needs, and context. That makes the quality surface much larger.

From Chapter 17: Personalized AI



THE PERSONALIZATION CONTRACT

Define what the system may remember, infer, use, and forget.

Observe	Preserve	Compare	Inspect	Replay
Find a real output difference or failure.	Save the input, route, context, and trace.	Known-good versus known-bad; old versus new.	Token, attention, activation, or feature drift.	Confirm with behavioral evals and intervention.

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

TEST COUNTERFACTUAL USERS

Hold the task constant. Change the person.

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.

N = 1 IS NOT CERTAINTY

The most personal experience has the smallest sample size.

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

Approve | refund | after | policy | review

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

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

THE DYNAMIC UI IS AN OUTPUT

AI can personalize more than text: layout, controls, workflow, and risk gates.

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.

WHEN NOT TO PERSONALIZE

Past behavior is not the same as current need, truth, safety, or public interest.

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.

USER-OWNED MEMORY

Users must be able to inspect memory, correct it, move it, and limit it.

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.”

THE ECONOMICS OF PERSONALIZATION

Every extra slice creates another population worth measuring and another way an average can hide harm.

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.

SYNTHETIC USERS ARE INSTRUMENTS

Personas can generate questions. They are not automatically the answer.

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.

THE STANDARD FOR GOOD PERSONALIZATION

Context-aware enough to help. Transparent enough to trust. Portable enough to leave.

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.

IcebergQA

AI-FIRST QUALITY WORK

THE PRACTICAL MOVE

Make personalization useful without making it invasive.

AI products need user-slice evidence, dynamic-interface checks, memory controls, and practical opt-out or portability paths. IcebergQA helps teams make personalized AI worth trusting.

Talk to IcebergQA