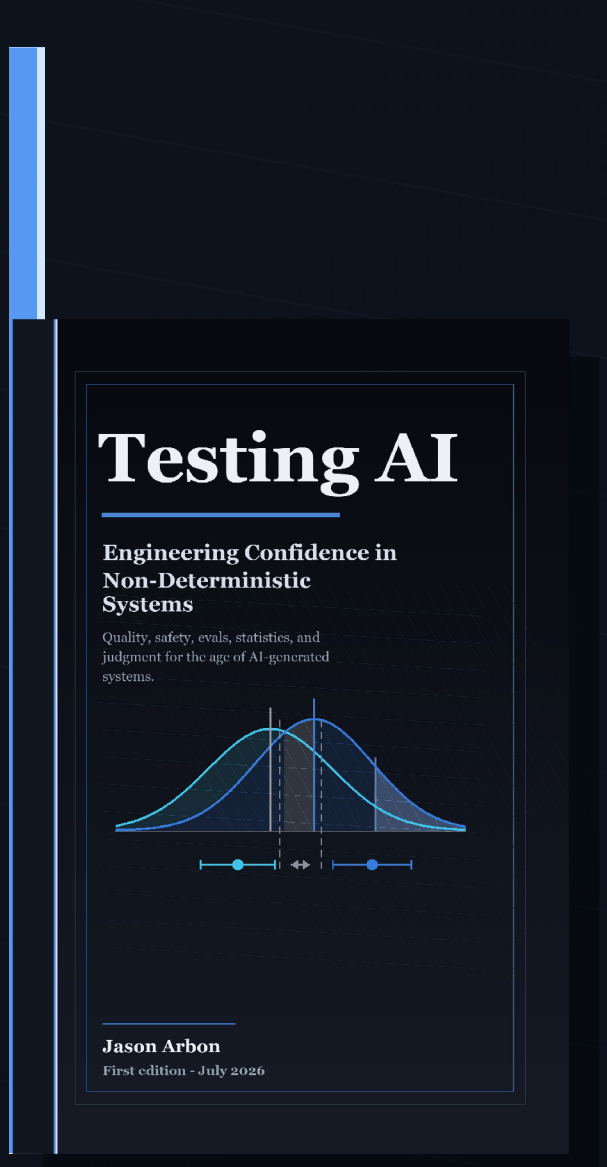
HOW MODELS WORK

Every model is a chain of transformations.

Input, data, tokens, model weights, tool calls, safety layers, and deployment behavior all shape what users actually receive.

From Chapter 15: How Models Work



THE TRAINING AND TESTING PIPELINE

A model is shaped long before a user sees the first answer.

Raw data	Filter	Pretrain	Align	Deploy
Sources, rights, languages, and gaps.	Deduplicate, clean, select, and exclude.	Learn broad statistical patterns.	Instruction, preference, and safety tuning.	Evals, red teaming, operations, and monitoring.

Testing belongs throughout the pipeline, not only after the weights have been trained.

DATA IS MODEL BEHAVIOR

The model learns from what it eats, including stale data, bias, and AI-generated noise.

Collection

What sources are present, absent, licensed, representative, and current?

Filtering

What did deduplication, quality rules, and safety removal erase or overemphasize?

Pollution

Can low-quality synthetic content, copied errors, or feedback loops amplify through new training runs?

The same model architecture trained on a different corpus is a different product.

TOKENIZATION IS A TEST SURFACE

The model receives token IDs, not the words you see.

Human prompt

"Testing AI: non-deterministic systems"

Names, punctuation, code, accents, emojis, dates, and long documents can split in surprising ways.

Example decoded tokens

Testing | AI | : | non | - | deterministic | systems

One possible tokenizer output, not a universal rule.

Test the boundary: context truncation, identifiers, multilingual text, copied punctuation, and token-budget behavior.

ALIGNMENT IS A PRODUCT PHASE

Pretraining makes a powerful autocomplete. Preference tuning makes it usable.

Pretraining

Predict the next token from vast amounts of text, code, images, and more.

Instruction tuning

Teach the model to follow requests, formats, and task conventions.

Preference and safety

Use human or AI feedback to reward helpfulness, boundaries, and better choices.

Product behavior

Test the resulting assistant, not the training story.

Alignment determines much of the style, helpfulness, and safety people attribute to "the model."

PREFERENCE TUNING NEEDS TESTING

Rewarding what people like is not the same as rewarding what is true.

Failure modes

- Longer answers win because they sound thorough
- Sycophancy wins because agreement feels pleasant
- Raters lack the context to judge the task
- One group's style becomes the default voice

Better evidence

- Calibrated raters and diverse context
- Clear definitions of helpfulness and harm
- Outcome checks beyond preference
- Separate personalized feedback routes from global training

RLHF and RLAIF are powerful product levers. Treat their reward signal as a measurement system that can be wrong.

THE PRODUCT FLOW

The model is only one part of the answer.

Input

Prompt, memory, files, identity, and user state.

Context

Retrieved evidence, policies, and tool results.

Model

Tokens, embeddings, transformer layers, decoder.

Actions

Tools, side effects, retries, and approvals.

Output

Safety checks, formatting, citations, and UI.

Every transformation can change quality, safety, cost, and reproducibility.

A BAD ANSWER IS A CLUE

Useful LLM bug reports preserve the path, not only the screenshot.

Weak report

"The model was wrong."

- No input or context
- No version or route
- No expected behavior
- No severity or repro

Useful evidence

Capture the conditions that could explain or reproduce the failure.

- Prompt, input, and user state
- Model, tools, retrieval, and versions
- Trace, output, rubric, and impact
- Failure family and regression case

A single bad answer rarely tells you whether the root cause is data, prompt, route, policy, tool, or model behavior.

IMAGE GENERATION IS A PIPELINE

One prompt can produce different images because the system starts from noise.

Prompt

Text, reference image, style, and editing intent.

Seeded noise

Starting randomness that changes the candidate image.

Denoising

Many guided refinement steps in latent space.

Output checks

Safety, artifacts, fidelity, composition, and prompt fit.

Test prompts, seeds, model versions, samplers, aspect ratios, safety filters, and visual failure modes together.

VISION-LANGUAGE MODELS

A fluent answer is not proof that the system understood the image.

Perception

OCR, small text, crop boundaries, charts, spatial relationships, and object confusion.

Context

Does the text prompt preserve uncertainty and request the evidence needed?

Decision

Is the answer grounded in visible evidence, calibrated, safe, and appropriately escalated?

Test the image, the extracted evidence, the question, and the final answer. Vision is a pipeline too.

IcebergQA

AI-FIRST QUALITY WORK

THE PRACTICAL MOVE

Test every transformation that can change the outcome.

From data construction through preference tuning, tool use, multimodal inputs, and production behavior, AI quality is a systems problem. IcebergQA helps teams turn those surfaces into practical evaluation and release evidence.

Talk to IcebergQA