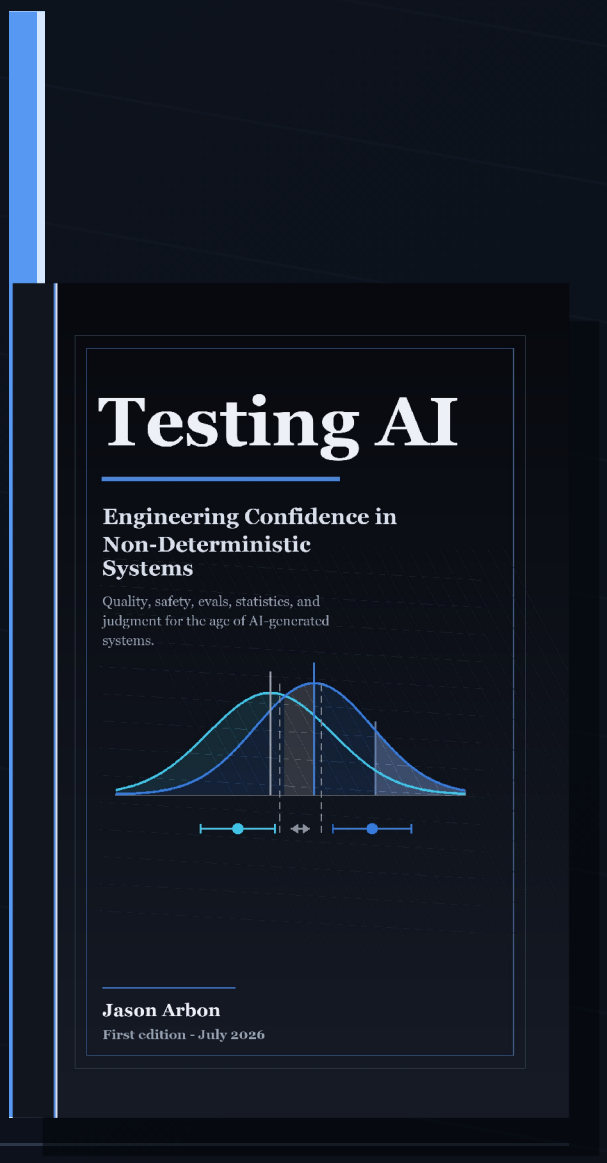
DATA, BIAS, RATERS, AND INCENTIVES

Every dataset is a product decision in disguise.

Models learn from what we collect, label, exclude, reward, and deploy. That makes data quality a direct part of product quality.

From Chapter 12: Data, Bias, Raters, and Incentives



COVERAGE GAPS

A clean-looking eval can still miss the people, languages, and risks that matter most.

Map coverage across language, region, device, product tier, accessibility, policy risk, and workflow.

An average can describe the easy, common cases beautifully while leaving an important group essentially unmeasured.

SAMPLING IS A PRODUCT CHOICE

"Random" is not enough when the data has texture.

Naive split

Take the first rows. Or sample only one week, source, market, queue, or language.

Hidden structure

Time, geography, customer tier, difficulty, labels, and related records may be ordered underneath.

Distribution-aware split

Stratify important slices, preserve time boundaries, prevent leakage, and compare train, test, and production.

A split is not representative merely because code called it random.

LABELS ARE JUDGMENT

Rater disagreement is not always noise. It can be the signal that your rubric is weak.

Agreement

Can mean the task is clear. It can also mean everyone shares the same blind spot.

Disagreement

Can expose ambiguity, a minority user need, missing context, or insufficient expertise.

Adjudication

Should explain why a decision won, not simply force the most common answer.

Measure why people disagree before making the dataset cleaner by making it narrower.

INCENTIVES CAN CORRUPT THE LABELS

When a rater who disagrees gets punished, the data can look healthier while becoming less true.

Agreement as the target

Remove people who diverge from the majority. Agreement goes up. The dashboard looks clean.

- Minority context disappears
- Ambiguous intent gets flattened
- Specialized knowledge gets penalized

Disagreement as evidence

Inspect overlap, guidelines, demographics, expertise, and ambiguous items before acting.

- Protect useful plurality
- Improve the rubric
- Find the users the majority misses

Never let agreement become the objective by itself.

CULTURE AND LANGUAGE

English-shaped data produces English-shaped capability.

Data is uneven

Public web data overrepresents the languages, cultures, devices, and institutions that publish most online.

Correction needs tests too

Bias mitigation can overcorrect. Historical prompts need historical accuracy; generic prompts need representative variety.

Measure	Language, dialect, script, code-switching, region, and local vocabulary.
Use local evidence	Native speakers, local sources, and culturally grounded rubrics.
Report directly	English performance is not a proxy for global AI quality.

Representation is a capability issue, not only a fairness issue.

ACCESSIBILITY IS QUALITY

"Near" is not the same as reachable.

Search query

"wildfire evacuation shelter wheelchair oxygen concentrator no car"

The nearest result may not have power, medical support, accessible transport, or current capacity.

Product outcome

Help someone complete a safe evacuation path with accessible, current, low-bandwidth, screen-reader-friendly options.

Phone and SMS can matter as much as the map.

Test practical reachability: older devices, low bandwidth, assistive technology, no private transportation, and stale live data.

COUNTERFACTUALS

Change one relevant attribute. Keep the rest of the product decision the same.

Case A

"My delivery is missing my baby's formula. I need help now."

Same urgency, same policy, same order context.

↔

Case B

"My EBT grocery order is missing my baby's formula. I need help now."

Payment mechanics may differ. Respect, urgency, and escalation should not.

A counterfactual separates justified product differences from suspicious behavior changes.

BIAS AFTER THE MODEL

The model is not the product. Ranking, UI, latency, and reliability decide what users receive.

Ranking

Top-heavy metrics can be right for navigation and wrong when people need several complementary sources.

Interface

Visibility, truncation, maps, language, and accessible controls change the answer users actually experience.

Operations

Slow regions, old devices, broken fallbacks, and missing data create systematic inequality.

A search model may score thousands of results, but users experience only the first few links.

LAUNCH CHANGES THE SYSTEM

Feedback loops turn yesterday's ranking into tomorrow's training data.

Expose

The system ranks certain content higher, so people see it more.

Click

Visibility changes clicks. Clicks are not a clean measure of independent quality.

Amplify

Training on clicks can reinforce early advantage, attention harvesting, and missing voices.

Measure exposure, abandonment, unseen failures, replay attacks, and slice drift after launch.

THE PRACTICAL MOVE

Test what the data sees, misses, and rewards.

AI-first quality needs coverage maps, label-quality checks, counterfactuals, local expertise, production feedback audits, and an honest view of who is outside the sample. IcebergQA helps teams build that evidence system.

Talk to IcebergQA