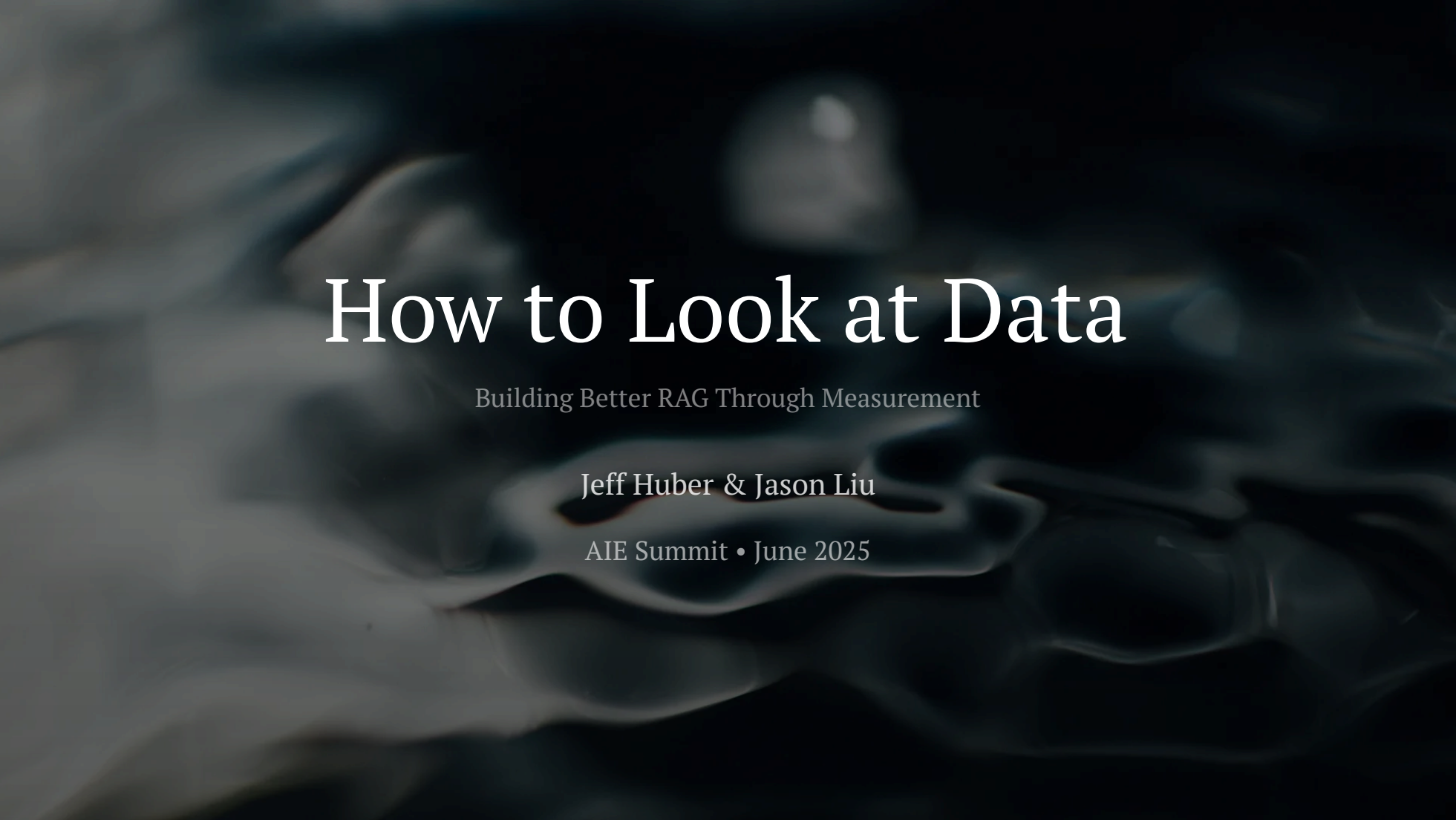


How to Look at Data

Building Better RAG Through Measurement

Jeff Huber & Jason Liu

AIE Summit • June 2025

The Two-Part Journey

Part 1: Measure Your Inputs

Jeff Huber

- Stop guessing, start testing
- Build fast evals on YOUR data
- Make empirical decisions about retrieval

Part 2: Measure Your Outputs

Jason Liu

- Turn chat logs into insights
- Cluster conversations to find patterns
- Build feedback loops that improve products

You can't manage what you can't measure

The Problem

You can’t manage what you can’t measure

- **Intervention Bias**: "hey guys would this work, would that work?"
- **Why guess when you can test?**
- **Stop flying blind** - without measurement, you’ll just thrash and crash

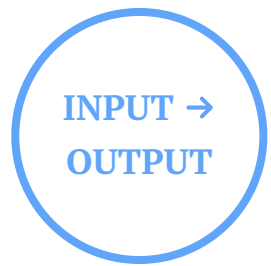
Word cloud of common problems:

Issues • Questions • Guesswork • Uncertainty

The Solution

Look at your data!

- **Great measurement makes systematic improvement easy**
- **Look at your inputs AND outputs**
  - **Inputs**: Build & Test - look at your documents
  - **Outputs**: Deploy & Monitor - look at your logs



Inputs: Look at Your Documents

There are many decisions in setting up your retrieval system

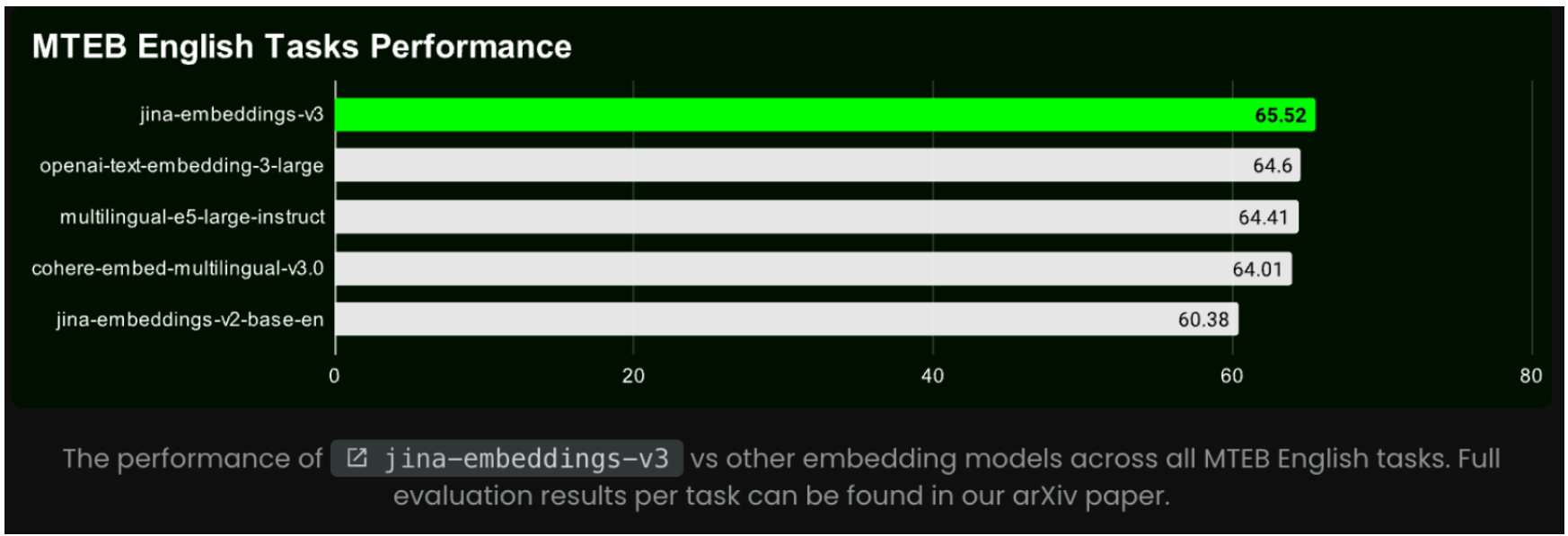
- **Which chunking strategy?**
- **Which embedding model?**
- **Query expansion?**
- **Reranking?**
- **And many more…**

⚠️ Critical Point

Public benchmarks are NOT representative of your data

Why Public Benchmarks Fail

- **Generic** - not specific to your domain
- **Overly clean** - real data is messy
- **Used for training** - LLMs and embedding models have seen this data



Fast Evals: The Solution

What is a fast eval?

- **"If this is queried, this document should be returned"**
- **A set of these = golden dataset**
- **Focus on "can you find it"**
- **Not "LLM as a judge over all your chunks"**

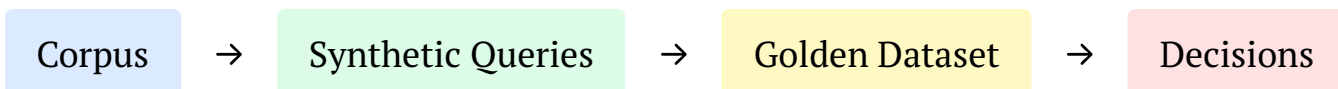
Key Insight:

Fast evals make decisions easy by focusing on retrieval accuracy

How Fast Evals Work

Use LLMs to write queries for your documents

1. **Take your corpus**
2. **Align an LLM to write representative synthetic queries**
3. **Get a golden dataset**
4. **Run your evals**
5. **Make empirical decisions**



Case Study: Weight & Biases Chatbot

Generative benchmarking revealed surprising insights

- **Their original embedding model performed worst** out of 4 models tested
- **Contradiction with MTEB model rankings**
- **36% of document chunks were judged irrelevant**

From Chat Logs to Actionable Insights

Turning thousands of support conversations into clear product direction.

The Pain Today

- User conversations have too much detail
- You’re not always the expert who knows where to look
- There’s too much volume to go through manually
- Outputs are noisy and hard to scan

The Feedback Is Already There

> "We built feedback widgets… but users already told us everything in the chat."

- Chats hold raw, unfiltered pain points
- Manually reading works until you go viral
- We need a systematic approach, not another form

Learning from Marketing Analytics

- Marketers slice by **age, gender, region**
- We slice by **topic, intent, capability**
- Goal: know where to **double-down** vs where to **improve**

The Data to Decision Loop

1. Define your success metrics (KPIs)
2. Group conversations into clusters
3. Compare clusters on your KPIs
4. Choose **build** / **fix** / **ignore** actions

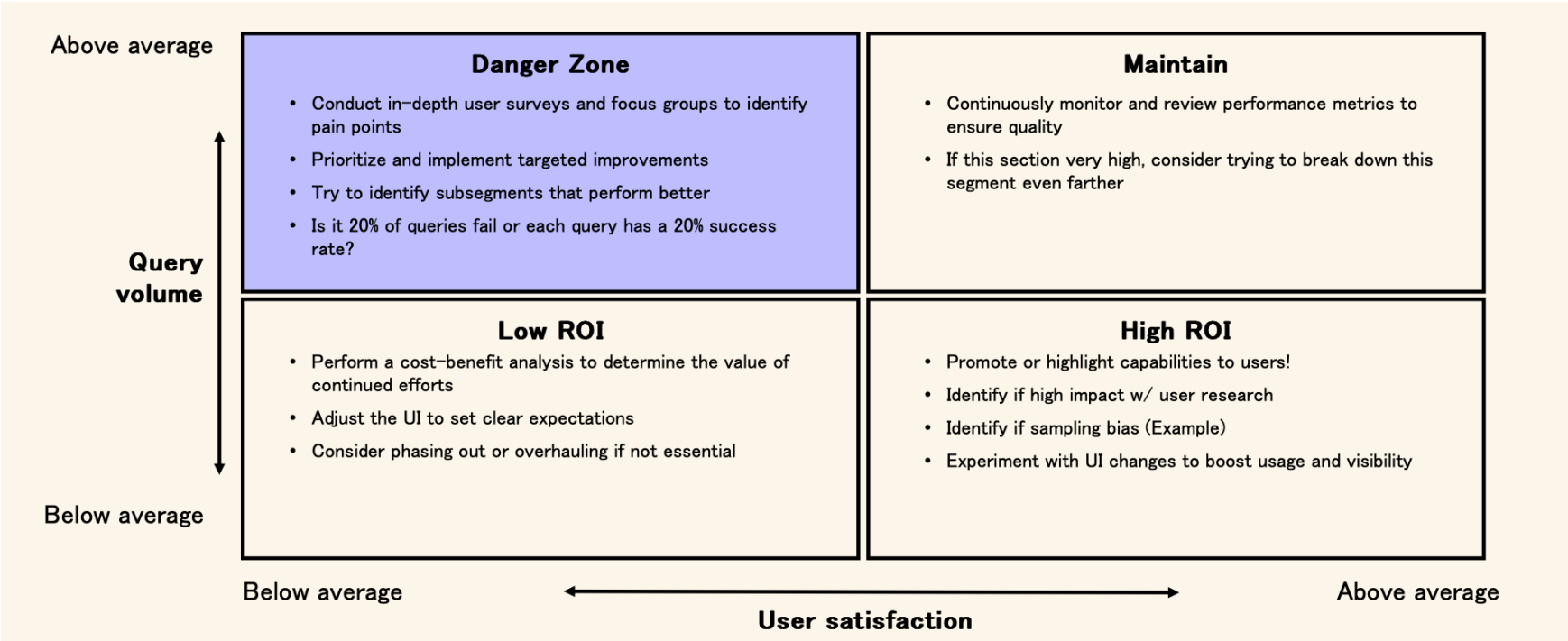
When Manual Reading Stops Working

- **10-100 chats**: Read everything manually
- **1,000+ chats**: Use language models for clustering
- Surface the most representative threads per cluster

What Clustering Reveals

- **Missing filters** users keep asking for
- **New data sources** to add to your index
- **Under-performing segments** (low satisfaction, long chats)

Quadrant Analysis Framework



- **High Traffic + High Satisfaction** = Double down
- **High Traffic + Low Satisfaction** = Fix immediately

Building the Feedback Loop

From Clusters to Better Products

Once you identify clusters, you can:

- **Build classifiers** to tag conversations in real-time
- **Create dashboards** to track trends over time
- **Set up alerts** for concerning patterns
- **Make data-driven roadmap decisions**

Getting Started

- **Example notebooks**: github.com/567-labs/how-to-look-at-data
- **Deep-dive blog**: improvingrag.com

Drop in your chat export and start clustering today.

Key Takeaways

Measure Inputs

- Public benchmarks ≠ Your data
- Fast evals make decisions easy
- Focus on retrieval accuracy
- Test on YOUR corpus

Measure Outputs

- Chat logs are gold mines
- Cluster to find patterns
- Build classifiers for tracking
- Close the feedback loop

Stop Guessing. Start Measuring.

Your data has the answers – you just need to look

Start Building Today

Resources

- **Example notebooks**: github.com/567-labs/how-to-look-at-data
- **Fast evals research**: research.trychroma.com/generative-benchmarking
- **Deep-dive blog**: improvingrag.com

Scan for resources

Questions?