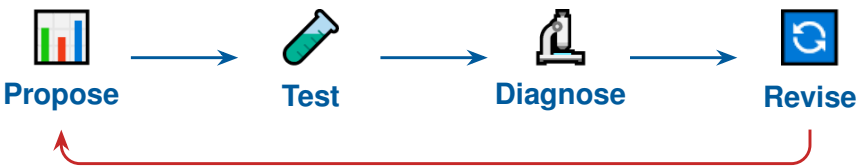
# AI and Social Science Research

Failing Faster, Learning Better

Valentina González-Rostani
University of Southern California

# Research Is an Iterative Process

AI agents organize tasks, use tools, and revise their work as new information becomes available.




**Explore alternatives and identify weaknesses before scaling up the study.**

# Two Risks in Research

### Too Little Exploration

Discarding promising ideas before testing them.

Example: abandoning a project because data collection is too costly.

**More exploration calls for better tests.**

### Too Little Criticism

Overinvesting in weak designs or measures.

Example: applying a measure that conflates concepts to more cases.

# Criticism Should Seek Out Flaws

| | |
|---|---|
| **Propose** | Design alternatives and make assumptions explicit. |
| **Challenge** | Look for errors, counterexamples, and conditions under which an approach fails. |
| **Validate** | Check against data, sources, and human review. |

**Agreement among agents does not replace independent evidence.**

# Creativity and Criticism in Research

| Application | Creativity Failure<br>Missed opportunities | Criticism Failure<br>Overinvestment in weak ideas |
|---|---|---|
| **Organizational Measurement** | Data-intensive projects are ruled out by collection costs. | Flawed measures are adopted and affect downstream analyses. |
| **Experiments and Audits** | Weak or underpowered designs can obscure real effects. | Contaminated or poorly validated evidence appears credible. |
| **Role of Agentic AI** | Expands research by reducing the costs of exploration and iteration. | Improves evaluation through iterative validation and stress testing. |

# AI as a Research Tool

Six uses: validate for the research goal

# Six Uses of AI in Social Science

Large language models (LLMs)

**1. Annotation and Measurement**

**2. Experiments**

**3. Synthetic Surveys**

**4. Simulations**

**5. Data Collection**

**6. AI Auditing**

# The Cost Revolution

Illustrative 2026 example: classifying one million tweets.

**Resource-Intensive Setup**

GPT-5.2.
High reasoning effort.
1 tweet per request.
No execution optimization.

**US$149,000**

**Optimized Setup**

GPT-5-nano.
No reasoning.
100 tweets per request.
Grouped inputs and batch API.

**US$2.90**

**Estimated cost reduction: ≈50,000-fold. Quality still needs validation.**

# Six Uses for Failing Faster, Learning Better

Applying the framework from Failing Faster, Learning Better.

| | |
|---|---|
| **1. Annotation and Measurement** | Test measures and correct errors early. |
| **2. Experiments** | Pilot treatments before scaling up the study. |
| **3. Synthetic Surveys** | Identify problems in questions before pilot studies with people. |
| **4. Simulations** | Test mechanisms and assumptions. |
| **5. Data Collection** | Make previously unaffordable projects feasible. |
| **6. AI Auditing** | Look for flaws in model responses. |

Simulations and AI auditing extend the chapter's general framework.

# 1. Annotation and Measurement

**Turn text, images, or videos into categories or measures.**

### Define the Measure

Coding criteria, examples, and review of disagreements.

Evaluate errors by group and language using holdout data.

### Hypothetical Example

A category appears in 0.5% of cases.

**99.5%**

accuracy from always answering "no," missing every positive case.

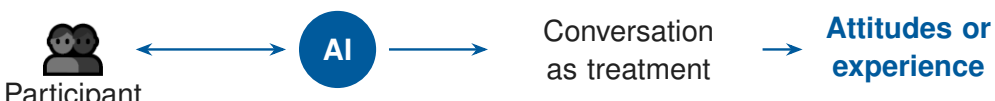
**Overall accuracy can conceal errors in rare categories.**

# 2. Experiments

**The stimulus is what a participant receives or experiences.**

### AI Creates the Material

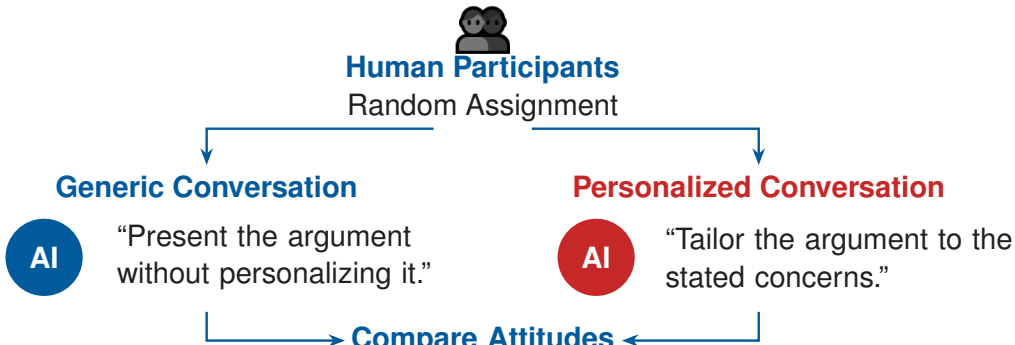

### AI Is the Conversation Partner




**Define what varies, save generated content, and run a pilot with participants.**

# 2. Experiments: Talking to AI

**Does personalizing a conversation change attitudes?**




Illustrative example. Keep the topic, length, and information comparable.

# 3. Synthetic Surveys

**Constructed Profile**

Age 35, Lima
Informal Employment

LLM

Language Model

**Generated Response**

**6/10**

Illustrative Example

“Using this profile, rate your trust in the local government from 0 to 10.”

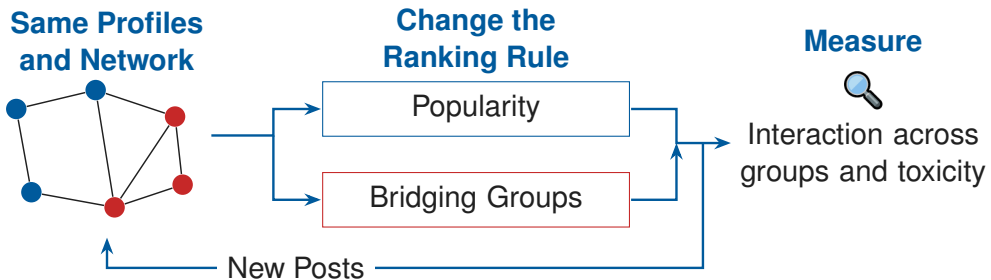
**Validate for the Intended Use**

| | |
|---|---|
| **Individuals** | Errors by individual and subgroup. |
| **Populations** | Means, dispersion, and distributions. |
| **Relationships or Effects** | Compare with independent evidence. |

**Matching averages does not guarantee matching relationships or effects.**

# 4. Simulations

**How do different recommendations change the conversation?**




Example: “Use your profile and memory to decide what to post in response.”

**Verify**

Test prompts, memory, and ordering.

**Validate**

Compare patterns with external data.

# 5. Data Collection

| | |
|---|---|
| **Locate** | Define the document population and inclusion criteria. |
| **Extract** | Save each value, its source, and the access date. |
| **Verify** | Check coverage, missing data, and extraction errors. |

**A useful dataset keeps each value traceable and protects sensitive data.**

# 6. AI Auditing

**Audit AI performance and study the consequences of its use.**

### What to Study

Capabilities and biases in specific tasks.

Effects on the decisions and behavior of people or organizations.

### Auditing Responses

Control context, conversation history, and option order.

Specify repetitions and scope: languages, versions, and dates.

**Model responses alone do not establish effects on people.**

# Learning from Failure

Identifying a failure helps us revise the design or narrow the conclusion.

| Failure Identified | Next Step |
| --- | --- |
|  The measure conflates economic and cultural frames. | → Distinguish the criteria and recode. |
|  Personalization also changes the tone. | → Revise the stimuli and repeat the pilot. |
|  The pattern disappears with minor changes to the prompts. | → Revisit assumptions. Validate with external data. |

**Detecting failures early allows revision before investing more.**

# Evaluation and Rigor

Look for failures before trusting results

# Challenges in Research with AI

### Reproducibility

Models change or become unavailable.

### Validation Benchmarks

Human judgments also need scrutiny.

### Model Bias

Errors can vary across groups and tasks.

### Equity and Access

Research resources differ across institutions.

### Opacity

Closed systems limit what we can inspect.

### Rapid Obsolescence

Model updates require results to be reassessed.

# Best Practices for LLM Research

1. Compare cost and quality before scaling up.

2. Separate generating alternatives from critiquing them.

3. Validate for the question, group, and context.

4. Record the version, date, prompts, and parameters.

5. Revalidate when the model or data change.

6. Finalize the protocol and set aside data for confirmatory analysis.

7. Save inputs and outputs while protecting sensitive data.

8. State the intended inferences and their limits.

**Records enable audits. Repeating the process may produce different results.**

# Key Takeaways for Research with AI


Explore more alternatives and detect failures earlier.


Compare cost and quality before scaling up.


Validate for the question, context, and population.


Document prompts, versions, and decisions.


Check results against independent evidence.


The research team is responsible for inferences, biases, and privacy.

**Thank you**

gonzalez.rostani@usc.edu gonzalez-rostani.com

# References


**Crabtree, González-Rostani, and Kim (2026)**

Charles Crabtree, Valentina González-Rostani, and Jae Yeon Kim.
*Failing Faster, Learning Better: Agentic AI and Empirical Social Science Research*.
In Rohan Alexander (ed.), *Quantitative Social Science in the Age of AI*, Chapter 8.
Read the Chapter

**Barrie et al. (2026)**

Christopher Barrie, Lisa Argyle, James Bisbee, Michael Heseltine, Christopher Lucas, Jon Mellon, Alexis Palmer, Margaret Roberts, and Arthur Spirling.
*AI and Research Methods*. APSA Preprints.
doi:10.33774/apsa-2026-h59kk