

What is a protest anyway?

Codebook conceptualization is still a concern in LLM-era classification

Andrew Halterman · Michigan State University

Katherine A. Keith · Williams College / Cohere

“Upset about the election results, a group of angry youth smashed the windows of businesses.”

Is this a **protest**?

“Upset about the election results, a group of angry youth smashed the windows of businesses.”

Is this a **protest**?

In daily use, probably. But four common social science *codebooks* (≈annotation guidelines) make different decisions:

ACE	ACLED	CAMEO	CCC
YES	NO	SUBCLASS	YES
“riots” are explicitly part of DEMONSTRATE	violence by protesters is explicitly excluded	coded as “protest violently, riot”	included; no violence exclusion

These are justifiable disagreements between domain experts with different substantive questions.

LLMs in a broader measurement pipeline

Social scientists want downstream estimates, using well defined concepts

CONCEPTUALIZATION

SCORING

INFERENCE

**Background
Concept**

*"a broad
constellation of
meanings"*

"Protest"

**Systematized
Concept**

Written in a
codebook

C

*"a gathering of 3+
people..."*

**Predictions/
Labels**

*LLMs/humans
label docs using
codebook C.*

$$\hat{Y}_i = f(X_i, C)$$

Estimate

*prevalence,
regression
coefficients, etc.*

θ

Our work

Most LLM work

Our work

SCOPE CONDITION

A "gold standard" label exists

1. A gold standard (oracle) label exists *given a codebook*

Given a fully specified codebook $\mathbf{C}$, each document $\mathbf{X}_i$ has a true label $\mathbf{Y}_i$.

(Excludes purely subjective tasks.)

2. An "expert" annotator always recovers the oracle

An expert annotator applies $\mathbf{C}$ perfectly:
 $\mathbf{Y} = \mathbf{f}(\mathbf{X}_i, \mathbf{C})$.

(A standard assumption in bias correction methods like DSL and PPI.)

The gold label is *conditional on the codebook*: $\mathbf{Y}_i \mid \mathbf{X}_i, \mathbf{C} \neq \mathbf{Y}_i \mid \mathbf{X}_i, \mathbf{C}'$

CLAIM 1

Annotation errors can be decomposed into two types

Scoring Errors

The model/annotator *misapplies* instructions.

$$\hat{Y}_i \neq Y_i$$

E.g., the definition excludes riots, but the annotator labels one as protest anyway.

Conceptualization Errors

The annotator is provided a codebook that does not reflect the researcher's concept.

$$C' \neq C$$

E.g. the research intends to measure peaceful protests, but the codebook fails to exclude riots.

Adcock and Collier (2001), Fariss, Kenwick, and Reuning (2020), Jacobs and Wallach (2021), Wallach et al. (2025).

CLAIM 2

LLMs can "fail silently"

A tale of two annotations

88% of 103 reviewed CSS papers recommend LLM annotation (Baumann et al. 2025); major benchmark studies prompt with bare label names, no definitions (Ziems et al. 2024)

CLAIM 2

LLMs can "fail silently"

A tale of two
annotations

Annotation in 2015

Researcher: *“Classify the following document as containing a protest.”*

Research Assistant: *“What do you mean by protest?
What about labor strikes?
How many people...?”*

88% of 103 reviewed CSS papers recommend LLM annotation (Baumann et al. 2025); major benchmark studies prompt with bare label names, no definitions (Ziems et al. 2024)

CLAIM 2

LLMs can "fail silently"

A tale of two
annotations

Human annotators
serve a useful
forcing function.

Annotation in 2015

Researcher: *"Classify the following document as containing a protest."*

Research Assistant: *"What do you mean by protest?
What about labor strikes?
How many people...?"*

Annotation in 2025

Researcher: *"Classify the following document as containing a protest."*

LLM: *"Certainly! I'd be happy to classify this document..."*

88% of 103 reviewed CSS papers recommend LLM annotation (Baumann et al. 2025); major benchmark studies prompt with bare label names, no definitions (Ziems et al. 2024)

CONCEPTUALIZATION

Three levels of definition (general → localized/bespoke)

TYPE I

Surface form

“Is this a protest?”

Meaning assumed self-evident.

What the LLM optimist uses.

TYPE II

Dictionary entry

“a public (often organized) manifestation of dissent”

Generic; built for every use-case

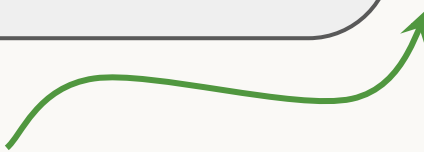
TYPE III

Stipulative Definition

“an in-person public demonstration of three or more participants in which the participants do not engage in violence and...”

Purpose-built, bespoke, with explicit inclusion *and* exclusion criteria. (ACLED)

Applied work needs Type III definitions for valid social science.



CLAIM 3

Conceptualization failure cannot be fixed with post-hoc correction or better LLMs

1. Use post-hoc bias corrections (e.g., PPI, DSL)?

Methods to correct bias from LLM annotations (DSL, PPI, etc.) rely on **expert labels** to learn a correction for noisy LLM labels.

By our assumption, even the oracle can produce incorrect labels when given an incomplete* codebook.

2. Use more accurate LLMs?

As annotator accuracy approaches oracle performance (i.e., LLM performance improves), errors in the flawed oracle persist.

CLAIM 3

Conceptualization failure cannot be fixed with post-hoc correction or better LLMs

Bias correction methods (DSL, PPI, etc.) rely on **expert labels** to learn a correction for noisy LLM labels.

By our assumption, even the oracle can produce incorrect labels when given an incomplete* codebook.

As annotator accuracy approaches oracle performance (e.g., LLM performance improves), this error persists.

Complete
codebook C

Incomplete
codebook
 $\tilde{C}$

Incomplete
codebook
 $\tilde{C}$

Oracle
labeler
 $Y \mid X_i, C$

Oracle
labeler
 $Y \mid X_i$
 $\tilde{C}$

LLM
labeler
 $\hat{Y} \mid X_i, C$

Y

$\neq$

$\tilde{Y}$

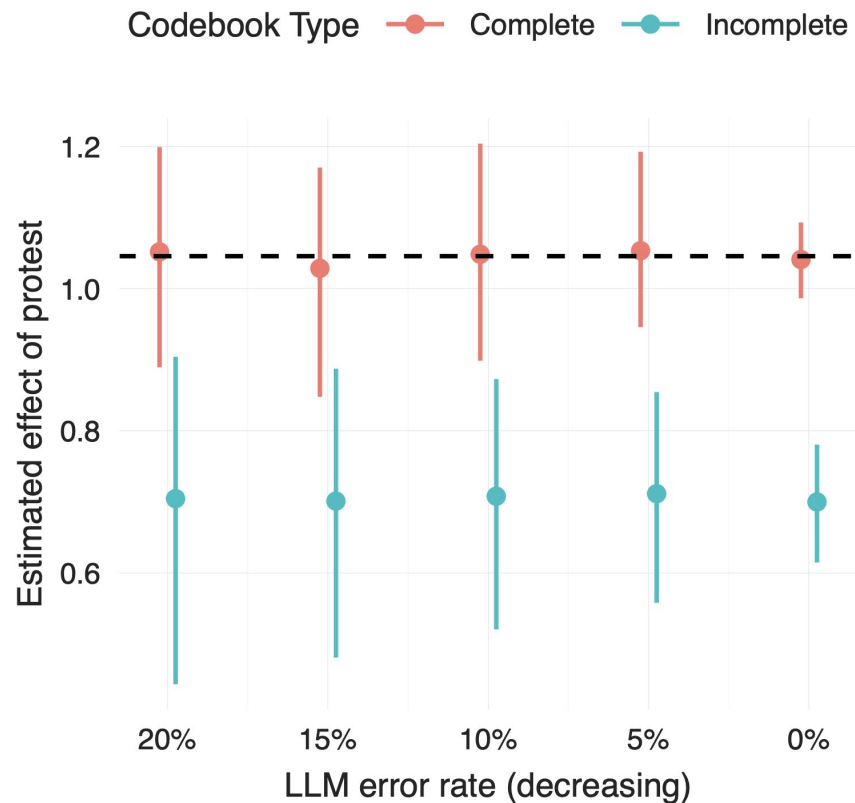
$\neq$

$\hat{\tilde{Y}}$

TESTING CLAIM 3 WITH SIMULATIONS

Simulation study:

- The researcher intends to measure the effect of (peaceful) protests on attitudes.
- Write an **incomplete** codebook *omitting* the violence exclusion.
- Bias correction (DSL, using incomplete annotations) does not eliminate bias.
- Increased LLM accuracy (with incomplete codebook) shrinks variance but not bias.



Our recommendations

The LLM Over-Optimist

- Use simple surface form definitions
- Throw it in the LLM
- No bias correction
- Very cheap, but *inevitably* biased

The Pragmatist

- Develop codebook
- Throw it in the LLM
- Use expert labels for bias correction
- Medium cost, unbiased

Thank you!

What is a protest anyway?

Codebook conceptualization is still a concern in LLM-era classification

Andrew Halterman · Michigan State University

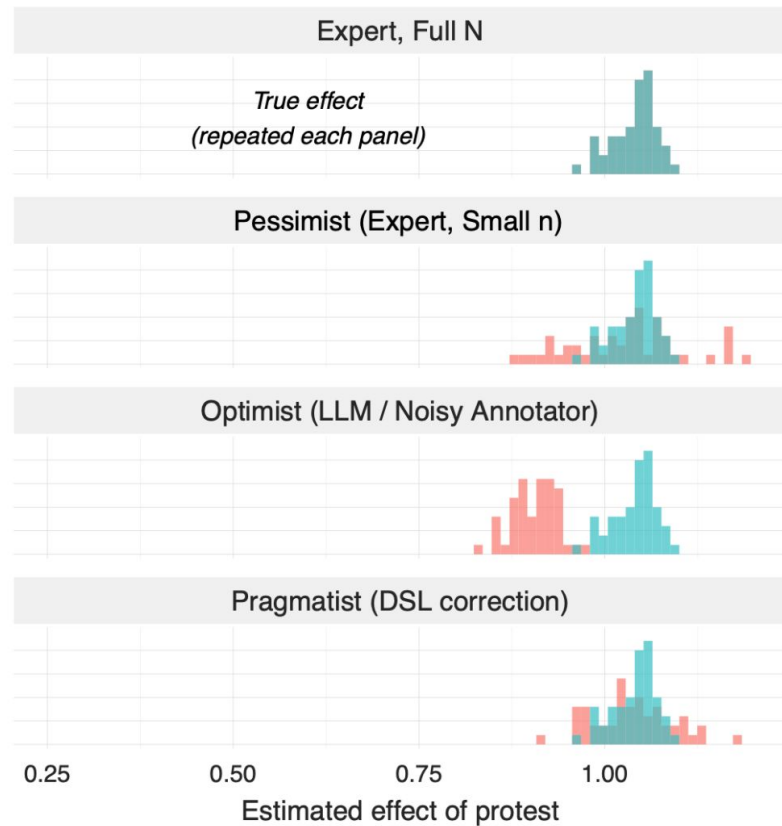
Katherine A. Keith · Williams College

BACKUP SLIDES

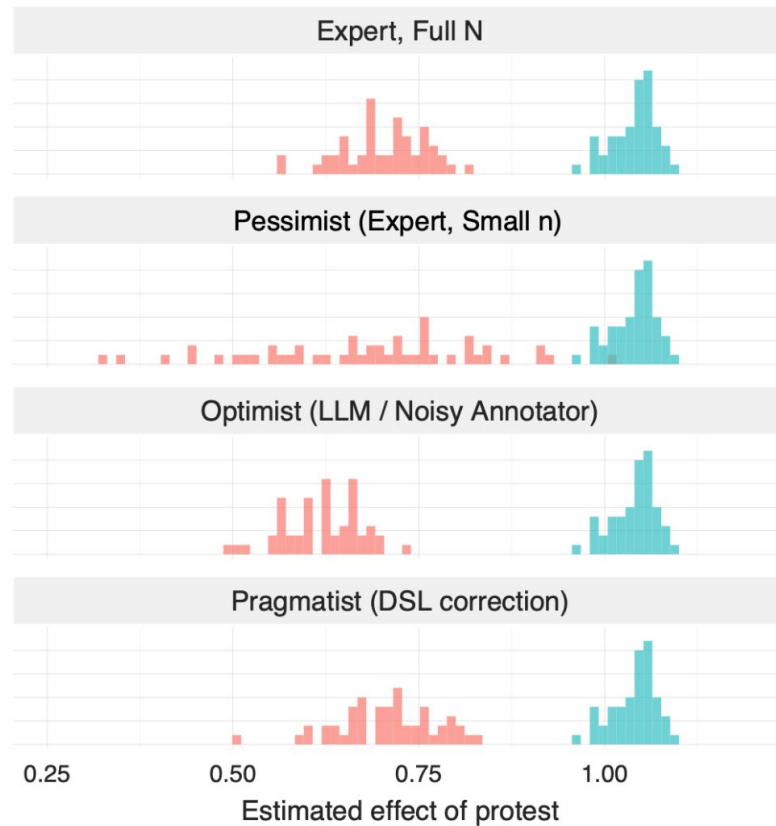
	ACE	ACLED	CAMEO	CCC
Single person protest	"a large number" required	3+ required	Implicitly included	No size limit
Protest with violence <i>by</i> protestors	"Riots" are included	Explicitly excluded	"Riot" subclass	Included
Protest with violence <i>against</i> protestors	No mention—likely included	Included	Included	Included
Protest in favor of a policy	No mention	Excluded	Coded as "diplomatic cooperation"	Coded as "rally"
Protest against business	No mention—likely included	Included	Included	Included
Online protests	Likely excluded ("public place")	Explicitly excluded	No mention	Included
Labor strikes	Included	Explicitly excluded	Subclass within protest	Not coded (resource constraints)
Hunger strike	No mention	Explicitly excluded	Subclass within protest	No mention
Civil disobedience	No mention—likely included	No mention—likely included	Coded as "defy norms"	Subclass within protest

Included
Subclass
Excluded
Separate class
Unclear

Complete Codebook



Incomplete Codebook



How do you know if a codebook is "complete"?

This is an open question. Two proposed standards:

1. If **two expert annotators are consistent** in annotating documents, the codebook is likely complete (no unresolved ambiguity)
2. If a **community of domain experts agree** that the codebook addresses all relevant aspects of the class (*although the specific decision can be different*), then the codebook is complete.

Can LLMs (help) auto-generate codebooks?

This is an exciting area of ongoing research (Xiong et al.'s Co-DETECT, HiCode, etc).

However, experimentally validating codebook variants requires expert re-annotation under *each* alternative codebook specification.

LLM label	Gold label	Inference	Description
–	$Y = \text{Expert}(X, C)$	Unbiased (high var.)	Pessimist. High variance from only human labels ($n \ll N$)
$\hat{\hat{Y}} = \text{LLM}(X, C)$	–	Biased	Optimist. No correction for annotation/scoring errors
$\hat{Y} = \text{LLM}(X, C)$	$Y = \text{Expert}(X, C)$	Unbiased	Pragmatist. A complete codebook permits post-hoc correction.
$\hat{Y} = \text{LLM}(X, C)$	$\tilde{Y} = \text{Expert}(X, \tilde{C})$	Biased	Procedural error. Gives incomplete $\tilde{C}$ to expert
$\hat{\hat{Y}} = \text{LLM}(X, \tilde{C})$	$Y = \text{Expert}(X, \cdot)$	Unbiased	Reliability error. Assumes an expert can label without C , see §3.5
$\hat{\hat{Y}} = \text{LLM}(X, \tilde{C})$	$\tilde{Y} = \text{Expert}(X, \tilde{C})$	Biased	Conceptualization error. Incomplete codebook, see §5

Table A1: **Comparing approaches to downstream inference**, many of which combine LLM zero-shot predictions with human expert gold-standard labels. Here, C is a complete codebook and $\tilde{C}$ is an incomplete codebook. We highlight the **pragmatist** row as the recommended approach, and also highlight the conceptualization error row which corresponds to Claim 3.