DDI TRAINING · VERSION 5.0

Democracy Discourse Index

Coder training for the Comment Edition

Global Centre for Rehumanising Democracy × ReHuman Lab × Sensika Technologies

September 2026 · gcrd.org.uk/ddi

What we will cover

Eight sections, two hours. Sections 3 and 8 are new in this edition — they are where the method changed.

01 **The Democratic Crisis Online**
10–12 min

02 **Introducing the DDI**
8–10 min

03 **The Comment as the Unit of Analysis**
18–20 min · NEW

04 **The Four Dimensions**
22–26 min

05 **How to Use the Coding Tool**
18–20 min

06 **Inter-Coder Reliability**
15–18 min

07 **Understanding the Statistics**
12–15 min

08 **From Your Batch to National Scores**
15–18 min · NEW

SECTION 01

The Democratic Crisis Online

Why this work matters, and why it cannot be done by machine alone.

> *Language is not merely a vehicle for ideas — it is social practice. When discourse deteriorates, democratic institutions follow.*
>
> *— DDI Design Principle*

Slides 3–5 · 10–12 minutes

Three failures, visible in any comment field

Each is a different problem. Each needs a different response — which is why the DDI measures four dimensions and not one number.

1 Hostility

Contempt replaces critique. People, not arguments, become the target.

Vulnerable voices exit public life first. Political violence becomes thinkable.

2 Trust collapse

Institutions, evidence and fellow citizens are assumed corrupt by default.

No collective action is possible where nothing and no one can be believed.

3 Agency deficit

“Nothing will change. They're all the same. Why bother?”

The quietest failure, the most self-reproducing, and the easiest to miss.

Strong disagreement is not the problem. Cruelty, cynicism and fatalism are.

The gap this method has to close

A national comment corpus against what a coding team can read. Four orders of magnitude.

millions

Comments in a national corpus

Across a balanced panel of outlets and digital public figures

thousands

Adjudicated rows per coding cycle

Three trained coders, reading carefully, one comment at a time

Your careful work is what makes machine coding at national scale possible — and defensible.

We come back to exactly how in Section 8.

SECTION 02

Introducing the DDI

What the instrument is — and, more importantly, what it is not.

Slides 6–8 · 8–10 minutes

What the DDI measures — and what it does not

This is the conceptual centrepiece of the training. Spend longer here than feels comfortable.

IT DOES

Measure how ideas are expressed

- Whether people are humanised or reduced to categories
- Whether disagreement attacks arguments or persons
- Whether trust in institutions, evidence and citizens is built or eroded
- Whether citizens sound able to act, or resigned

IT DOES NOT

Judge whether the view is right

- It does not score political correctness
- It does not reward moderation or penalise radicalism
- It does not evaluate whether a claim is factually true
- It does not judge the person who wrote the comment

A radical comment can score 80/80. A moderate comment can score 0. The dimension being measured is discourse quality, not political content.

Where your work sits

GCRD is a think-act-teach organisation. Coded data is the foundation the other two pillars stand on.



Research

Empirical and theoretical research into democratic discourse, disinformation and civic resilience.



Policy and practice

Policy engagement, practitioner training and leadership formation built on that evidence.



Programmes

University programmes and the DDI consortium — including this course.

Without high-quality coded data the research cannot proceed. Without the research, action has no evidence base. Without evidence-based action, teaching is just opinion.

Why comments, and not posts

A post is an act of publication. A comment is an act of response — and a public is constituted by how it answers.

PILOT EDITION

The post

- Written by accounts choosing to broadcast
- Few, and heavily professionalised
- Tells us how public figures speak
- Composed with an audience strategy

THIS EDITION

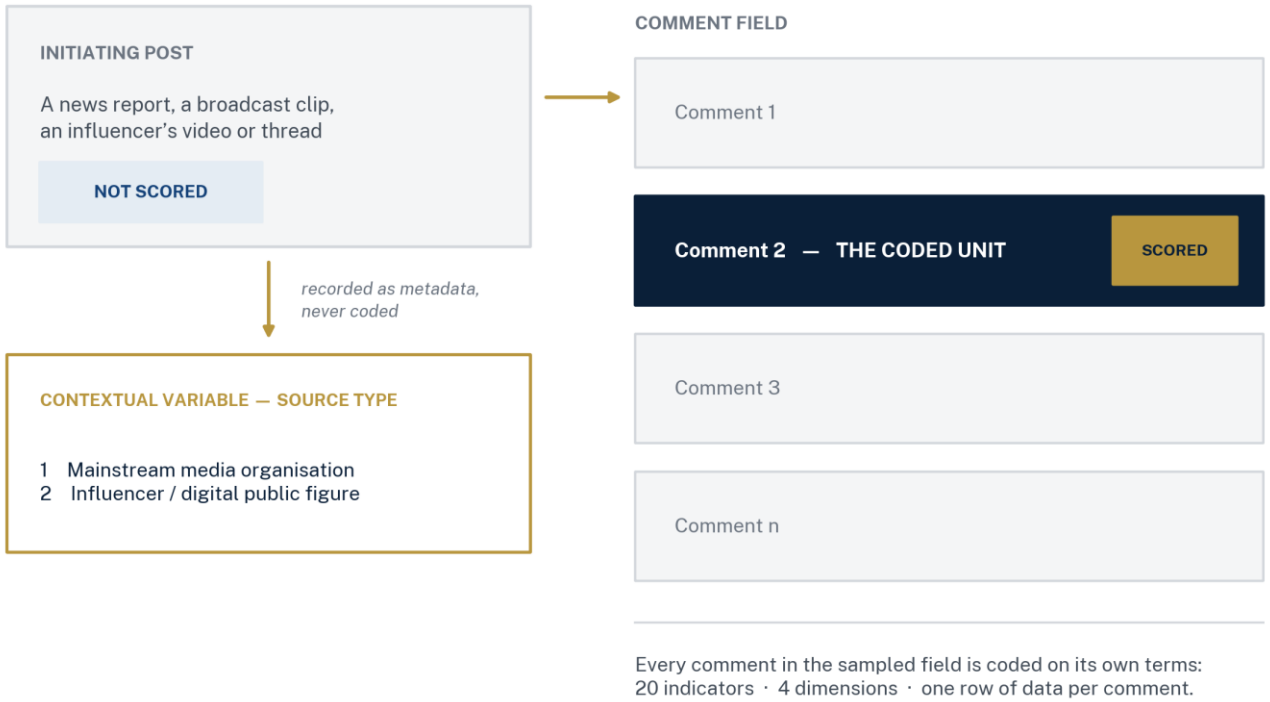
The comment

- Written by citizens responding to something
- Abundant and spontaneous
- Tells us how the public receives what was said
- Closer to unguarded speech

Expect the objection: aren't commenters angrier than the general public? Answer honestly — yes, they are self-selecting. The DDI reports on the discourse of the commenting public, not on public opinion. State it as a limitation in findings; it is not a reason to prefer posts, which are written by an even smaller and more professionalised group.

What is scored, and what is not

FIGURE 1 · The unit of analysis is the comment



Read the initiating post once, at the start of the batch. Then set it aside and score only what the commenter wrote.

Source Type: recorded, never scored

The live research question is whether the democratic quality of public discussion differs according to who convened it.

1 Mainstream media organisation

A newspaper, broadcaster, wire service or established news website — including its own social accounts and YouTube channel.

2 Influencer or digital public figure

An individual or independent outlet whose reach comes from a personal or platform-native following, not an institutional newsroom.

You record it. You never judge it. Source Type is metadata — it is not one of the twenty indicators, and it has no bearing on any score.

Watch your own expectation. Most people assume influencer comment fields will be worse — and then unconsciously score them down. Name that expectation now so you can notice it in yourself later.

The relative scoring trap

The single most important warning in this edition. The anchors are absolute, not relative.

A vicious thread

"Absolute scum, every one of them."

"They should be strung up."

"I think this is poorly argued, frankly."

"Vermin. Nothing else to call them."

The third comment is dismissive. Beside these, it reads as civility.

What goes wrong

- Students compare each comment to its neighbours, not to the anchors
- Scores compress toward the middle within a thread
- Reliability looks acceptable overall while being unstable across threads
- The error survives adjudication and reaches national scale

The corrective, as a habit: score each comment as if it were the only one you had read that day.

When you catch yourself thinking "well, compared to the one above it…", stop and reopen the Score Guide.

SECTION 04

The Four Dimensions

Twenty indicators. Eighty points. The communicative ecology democracy depends on.

Slides 13–19 · 22–26 minutes

D1 Empathy Index

Empathy to Indifference · Max 20 points

Does this comment treat other people as human beings — the people in the report, the people it is about, and the other citizens in the thread?

D1.1	D1.2	D1.3	D1.4	D1.5
Perspective-Taking	**Emotional Recognition**	**Humanisation**	**Solidarity**	**Anti-Hostility**

WATCH FOR

D1.5 is the most diagnostic indicator in comment fields. It measures schadenfreude and contempt — extremely common in reader comments, and very damaging.

Each indicator is scored 0–4. Five indicators × 4 points = 20 per dimension.

D2

Civility Index

Civility to Toxicity · Max 20 points

Civility is not politeness. It is keeping the conditions that allow disagreement to continue without dissolving the political community.

D2.1	D2.2	D2.3	D2.4	D2.5
Respectful Tone	**Non-Dehumanisation**	**Constructive Disagreement**	**De-escalation**	**Generous Interpretation**

WATCH FOR

Three possible targets in a comment field: the subject of the report, the outlet or influencer who published it, and other commenters. All three are people.

Each indicator is scored 0–4. Five indicators × 4 points = 20 per dimension.

D3

Trust Language Index

Trust to Cynicism · Max 20 points

Not whether institutions deserve trust — scepticism is vital — but whether discourse builds or erodes the trust infrastructure democracy depends on.

D3.1	D3.2	D3.3	D3.4	D3.5
Institutional Trust	**Good Faith Attribution**	**Interpersonal Trust**	**Epistemic Trust**	**Anti-Conspiracy**

WATCH FOR

Comments beneath news reports frequently attack the outlet itself. “This article is badly argued” is legitimate critique. “Th e media are all paid liars” is not.

Each indicator is scored 0–4. Five indicators × 4 points = 20 per dimension.

D4

Democratic Agency

Efficacy to Disempowerment · Max 20 points

Whether citizens believe they can shape political outcomes through collective action. Unique to the DDI — most discourse frameworks do not measure it.

D4.1	D4.2	D4.3	D4.4	D4.5
Political Efficacy	**Mobilisation Framing**	**Collective Action**	**Rights Consciousness**	**Democratic Optimism**

WATCH FOR

Expect this to be the lowest-scoring dimension. Comment fields are where fatalism concentrates — that is the finding, not a coding error.

Each indicator is scored 0–4. Five indicators × 4 points = 20 per dimension.

The complete framework

FIGURE 3 · Four dimensions, twenty indicators, eighty points

D1 Empathy *Empathy ↔ Indifference*	D2 Civility *Civility ↔ Toxicity*	D3 Trust Language *Trust ↔ Cynicism*	D4 Democratic Agency *Efficacy ↔ Disempowerment*
D1.1 Perspective-Taking	D2.1 Respectful Tone	D3.1 Institutional Trust	D4.1 Political Efficacy
D1.2 Emotional Recognition	D2.2 Non-Dehumanisation	D3.2 Good Faith Attribution	D4.2 Mobilisation Framing
D1.3 Humanisation	D2.3 Constructive Disagreement	D3.3 Interpersonal Trust	D4.3 Collective Action
D1.4 Solidarity	D2.4 De-escalation	D3.4 Epistemic Trust	D4.4 Rights Consciousness
D1.5 Anti-Hostility	D2.5 Generous Interpretation	D3.5 Anti-Conspiracy	D4.5 Democratic Optimism
5 × 4 = 20 pts	5 × 4 = 20 pts	5 × 4 = 20 pts	5 × 4 = 20 pts

COMPOSITE	four dimension subtotals summed, expressed as a percentage	80 pts

Twenty indicators, four dimensions, eighty points. When coding, always score at indicator level — never at band level.

Same report. Same strength of disagreement.

Both comments respond to the same story about a planning decision. Read them before you look at the scores.

COMMENT A

"Typical. Anyone defending this lot is either on the payroll or too stupid to see what is happening. They have wrecked this town and they will keep doing it, because nothing ever changes and nobody can stop them."

D1	D2	D3	D4
2	1	2	1

6 / 80 · 7.5%

RISK BAND

COMMENT B

"I profoundly disagree with this decision and I think the people who made it got it badly wrong. But I can see why the families quoted here are frightened — that fear is real and it deserves an answer, not a lecture. The consultation is open until the 14th. We should all be responding to it."

D1	D2	D3	D4
15	17	14	16

62 / 80 · 77.5%

HEALTHY BAND

"I profoundly disagree" is not weak language. The difference is entirely in the how.

One comment, one row

Check the metadata before you score. A mis-recorded Source Type corrupts the comparison the whole study is built on.

COMMENT TEXT	METADATA	20 SCORES	CONF + SKIP	NOTES
the unit being scored	outlet · platform · Source Type · date · thread position · parent	D1.1 – D4.5, each 0–4	one per dimension	free text

Check these every time

- Source Type — mainstream, or influencer?
- The initiating post identifier matches the batch
- Parent comment is populated where this is a reply

Use the notes field

- Sarcasm and irony you had to interpret
- Slang or register outside your fluency
- Translation doubts, or heavy dependence on the report

The notes field is how a coder's difficulty becomes actionable information for the research and engineering teams.

The 0–4 scale measures degree, not applicability

The question is never “does this indicator apply?” — it is “how strongly is this quality present?”

0	1	2	3	4
Absent	**Weak**	**Moderate**	**Strong**	**Exemplary**
No trace. The opposite may dominate.	A barely detectable or isolated trace.	Clearly present but limited or inconsistent.	Well-developed and consistent throughout.	Outstanding, sustained and nuanced.

Central tendency bias

The human pull toward the middle. If a comment is genuinely exemplary, score it 4. Inflation toward 3 compresses the range and makes the data less useful.

What to expect in comments

Most real comments cluster around 1–2 on most dimensions. That is a finding — do not inflate scores because the data looks too negative.

The ten-step workflow

Steps 1, 4 and 7 are the ones students skip. Each has a specific consequence.

1. **Read the initiating post once**
2. **Open only your own coder tab**
3. **Check the pre-populated metadata**
4. **Read the comment — and its parent**
5. **Scan the Skip toggles**
6. **Score all indicators, add confidence**
7. **Wait until all three have finished**
8. **Adjudicate flags as a team**
9. **Check the reliability statistics**
10. **Save and submit the batch**

Independence is not a suggestion. A single conversation between two coders before both have finished contaminates that comment's data.

Skip logic and confidence ratings

SKIP is the most misused feature of the workbook — and short comments make the misuse tempting.

SKIP

Use it only when a dimension is structurally inapplicable — there is no content at all that could be scored on it.

Never because the comment is short, difficult, or scores badly. The zero is the data.

CONFIDENCE

One rating per dimension: High, Medium or Low. It never changes a score — it is metadata about your certainty.

Medium is not the safe default. Fifty Medium ratings tell us nothing.

Score 0 ≠ SKIP

A comment that is nakedly cynical about political participation scores 0 on D4.1, 0 on D4.2, and so on. A one-line factual correction with no evaluative content at all might warrant a SKIP on D4.

Low confidence is engineering work. It is how the ReHuman team identifies the indicator–context combinations where the machine will also struggle.

Worked example: scoring Dimension 1

Score it silently first. Then we compare.

“I get why people are angry about this — the families quoted in the piece have been waiting three years for an answer and nobody has given them one. That is a real failure. But calling everyone at the council corrupt does not help them, and it is not true of the caseworkers I have dealt with.”

D1.1 4

Perspective-Taking

Holds two standpoints at once: the families, and the caseworkers.

D1.2 4

Emotional Recognition

Names the anger and treats it as legitimate, not as noise.

D1.3 3

Humanisation

Families and caseworkers appear as people — but briefly.

D1.4 3

Solidarity

Sides with those waiting, without sustained commitment.

D1.5 4

Anti-Hostility

No contempt anywhere, including toward the council.

Subtotal 18 / 20 · 90%

Why are two of these a 3 and not a 4? Present, but brief — not sustained. That distinction calibrates the whole 3/4 boundary.

SECTION 06

Inter-Coder Reliability

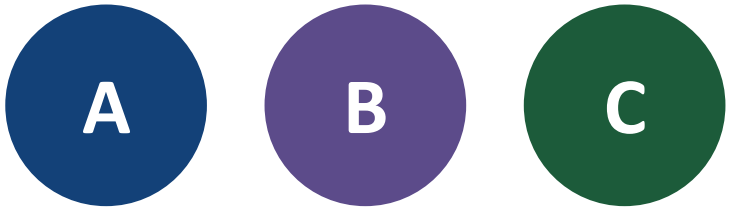
Three coders, working alone, then resolving their differences in the open.

> *Disagreement is not a problem to be tidied away. It is data about the rubric, about the material, and about us.*
>
> — *DDI Reliability Principle*

Slides 25–28 · 15–18 minutes

Why three independent coders

No single coder's interpretation is definitively correct. Three independent readings triangulate toward the most defensible one.



Three trained coders score every comment independently.

1

Bias detection

A single coder's systematic bias is invisible in single-coded data. Disagreement patterns make it visible.

2

Rubric validation

Indicators that generate repeated flags are indicators the rubric defines ambiguously.

3

Training quality

ReHuman learns from adjudicated data. A model trained on one coder learns that coder's idiosyncrasies too.

Every low-reliability indicator you identify is an indicator the machine will also struggle with. Fixing it for human coders fixes it for ReHuman.

Reading the Reliability tab

It looks complex. The logic is simple: three coder columns, an agreement status, and one place the team writes together.

Coder A · B · C	The three independent scores, pulled in automatically. Read-only.
Agree	Full — all three identical. Partial — at least one differs.
Flag	Discuss — two coders differ by more than one point. Must be adjudicated.
Mean	The average of the three. A starting point for discussion, not a substitute for it.
Adjudicated	The only column you type into. Highlighted yellow. This is the official data point.

Expect more flags than the old edition. Short, sarcastic, context-dependent comments generate more legitimate disagreement than composed posts did.

Running a good adjudication session

The goal is accuracy, not harmony. A team that agrees by capitulating produces reliable data that is systematically wrong.

1. Read the comment again together — and its parent, if it is a reply
2. Each coder states their score and their reasoning. All three speak first
3. Open the Score Guide and read the relevant anchors together
4. Discuss. Which anchor does this comment most closely match?
5. Enter the agreed score in the yellow column
6. No agreement? Enter the rounded mean, note it, flag for the supervisor
7. Move on. Five minutes maximum on any one indicator

Two failure modes

- Debating the politics of the comment instead of the indicator
- Reasoning by comparison with other comments in the thread

Redirect with a question

“What does the rubric say about this indicator? What is the evidence in the comment text?”

Then send them to the Score Guide to read the anchor each is defending aloud.

Change your score because someone showed you something in the comment — not because they were more confident or more senior.

SECTION 07

Understanding the Statistics

One formula, in plain language. It measures how much better than chance your team is performing.

$$\alpha = 1 - (Do \div De)$$

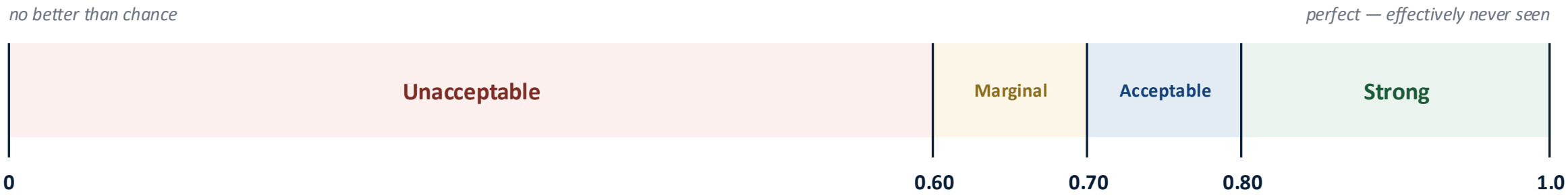
Do = the disagreement you actually see

De = the disagreement chance alone would produce

Slides 29–32 · 12–15 minutes

What the number actually means

If coders genuinely understand the rubric and apply it consistently, observed disagreement should be much smaller than chance disagreement.



Where good work actually lands

0.85–0.90 is exceptional. 0.70–0.80 is solid professional-standard work. Nobody reaches 1.0.

Where 0.70 comes from

Krippendorff's own 1980 Content Analysis, and the social science standard for four decades since.

0.70 does not mean 30% of disagreements are acceptable. It means this is the minimum level at which findings can be reported as reliable enough to support valid conclusions.

Percentage exact agreement, worked

Indicator D1.1 across five comments. Count the exact agreements before you look at the answer.

	Coder A	Coder B	Coder C	Status
Comment 1	**3**	**3**	**3**	No action needed
Comment 2	**2**	**2**	**3**	Partial — discuss briefly
Comment 3	**1**	**3**	**2**	Discuss — must adjudicate
Comment 4	**4**	**4**	**4**	No action needed
Comment 5	**2**	**2**	**2**	No action needed

Three exact agreements out of five = 60%. Below the 70% threshold — this indicator needs review before the batch is submitted.

Thresholds, and what to do at each

This is a protocol, not a grade. Each band tells you exactly what happens next.

alpha ≥ 0.80	**Strong**	The rubric is working very well. Submit with confidence.
alpha ≥ 0.70	**Acceptable**	Usable and publishable. The batch enters the ground-truth pool.
0.60 – 0.69	**Marginal**	Review the flagged indicators, recalibrate, re-code a small sample. Does not enter ReHuman training.
below 0.60	**Unacceptable**	Do not submit. Something fundamental has gone wrong — find out what.

Low reliability on a specific indicator is an invitation to improve the research tool — not evidence of student failure. The best published studies describe exactly this iterative process.

So how do we get from your batch to a country?

Before the answer, take the problem seriously. How would you close this gap?

What a cycle produces

a few thousand

adjudicated rows, three coders, reading carefully

What a country contains

millions

comments across a balanced national panel

Ask the room first. Take each answer seriously, then show why it fails:

More coders

More expense, and reliability falls as the team grows

A shorter rubric

Measures less — the four dimensions exist because one number hides the diagnosis

Sampling alone

Cannot support topic-level or longitudinal claims at national scale

ReHuman is the answer to a problem you have now felt — not a technology announcement.

What ReHuman is — and what it is not

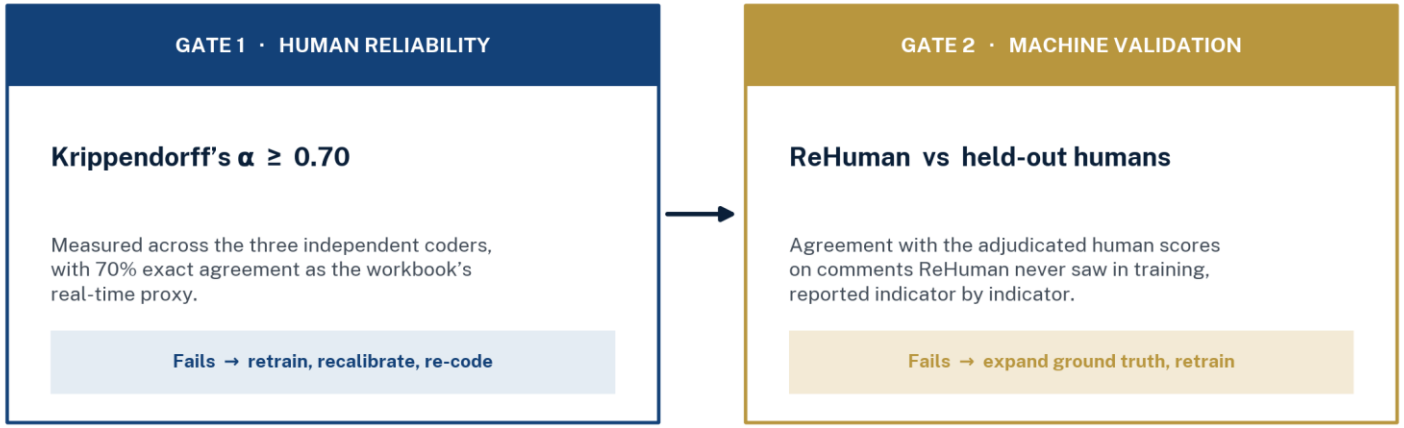
Not a general-purpose classifier pointed at DDI data. A system trained on your adjudicated scores, against your anchors.

	HUMANS	REHUMAN
Establish the standard	Humans	No
Decide what gets sampled	Humans	No
Code the ground-truth sample	Humans — three per comment	No
Code the remaining corpus	No — except audit samples	ReHuman, after validation
Resolve ambiguous comments	Humans	Flags them, does not decide
Interpret and publish findings	Humans — always	No

The adjudicated human data is the only thing ReHuman learns from — never raw individual scores, never a batch that failed the gate.

Two gates, one standard of evidence

FIGURE 4 · Two quality gates, one standard of evidence



No comment-level score enters a published DDI figure until the gate that produced it has been passed and reported.

Held out means held out. If the validation set is used even once to tune the system, it has become training data and a fresh set must be drawn.

PASS

Released for that corpus and cycle only.

PASS WITH EXCEPTIONS

Failing indicators go to human coders.

FAIL

Expand ground truth, recalibrate, revalidate.

Why your batch is the standard

ReHuman is trained on adjudicated human data and validated against human data it has never seen.

Every systematic error in your coding becomes a systematic error at national scale. Every ambiguity you flag becomes a warning the engineering team can act on.

The audit continues after release

A random proportion of every machine-coded batch is triple-coded by humans, blind. The agreement rate is published with the findings.

When you and the machine disagree

The human score stands in the dataset. The disagreement is recorded as information about the model — not about you.

Your coding is not a sample of this work. It is the definition of correctness for all of it.

YOU ARE NOW READY TO CODE

Five things to carry into your first batch

1. Read the initiating post once — then set it aside
2. Score the comment, never the thread around it
3. Short is not a reason to skip. Score it, and say how sure you are
4. Share nothing until all three of you have finished
5. Allow three to five minutes per comment for your first batch