ACCEPTANCE CODE MSR27

ISPOR ASIA PACIFIC SUMMIT 2026 · BANGKOK · 6–8 SEPTEMBER 2026 · METHODOLOGY

Validating LLM-Assisted Classification Of Patient Voice From Social Media: A Dual-Artifact Workflow With Held-Out Reproducibility, Demonstrated In Fertility Care

Rushabh Mehta[1], Varsha Patial[2], Vaidehi Ketkar[2], Rahul Srivastava[3] [1] SIROai · [2] Aksigen IVF · [3] SIRO Clinpharm Pvt. Ltd. · Mumbai, India

1 OBJECTIVE

CURRENT APPROACH

Posts → LLM → Labels → **Unknown accuracy**

A model returns a confident label for every item whether or not the instrument behind it is sound, and nothing in the output distinguishes the two.

VALIDATED APPROACH

Posts → Codebook → Prompt → Independent human validation → κ agreement → **Trusted labels**

OBJECTIVE

To validate a workflow that separates the human-facing codebook from the LLM-facing prompt, routes each disagreement to its source, and confirms reproducibility on a held-out sample.

4 domains
25 sub-themes
6 classification dimensions
2 independent human raters

2 METHODS · THE WORKFLOW

Nine stages, one pass, instruments locked before the held-out draw

1. **109,889 posts** — Reddit & YouTube patient voice
2. **Cleaning** — Deduplicate, then filter on content
3. **Codebook** — Human-facing · v1.0 → v1.5
4. **Prompt** — LLM-facing · v1.1 → v1.3
5. **Human rating** — 2 raters, blind to AI output
6. **LLM classification** — Same items, same codebook
7. **Agreement analysis** — Cohen's κ vs pre-set κ ≥ 0.60
8. **Disagreement routing** — To codebook, prompt, or both
9. **Held-out → production** — n=~200 unseen, then 22,619 items

DEVELOPMENT
n=137 dispatched → n=111 analyzable

ROUTING
R10, R11 → codebook v1.4 · R12 → codebook v1.5 + prompt v1.3

LOCK

HELD-OUT
n=80 drawn after lock (62 Reddit / 18 YouTube)

PRODUCTION
22,619 items · codebook v1.5 · prompt v1.3

Development ran on codebook v1.3.1 · prompt v1.2. The held-out sample was drawn only after both instruments were locked, and was unseen during codebook iteration v1.3.1 → v1.5 and prompt iteration v1.2 → v1.3.

2b METHODS · TWO INSTRUMENTS, ONE LOGIC

Codebook
Human-facing · v1.0 → v1.5

- Conceptual definitions and boundaries
- 12 numbered decision rules
- Worked examples for rater training
- Distributable to human raters

SHARED LOGIC

Prompt
LLM-facing · v1.1 → v1.3

- Mechanical operationalisation
- Anti-keyword tests, removal procedures
- Output schema and error handling
- Executed by the model via API

Every disagreement is routed to the surface that caused it

Disagreement

- **Human vs human** → codebook ambiguity → **REVISE CODEBOOK**
- **AI vs human** → three types, routed separately
  - Literal-vs-meaning gap — PROMPT
  - Codebook gap the AI's literalism surfaced — BOTH
  - Rule requiring both surfaces — BOTH

3 RESULTS · DEEP DIVE, D4 TREATMENT NAVIGATION

BEFORE

"… we cannot afford any more IVF cycles. I'm terrified."

↓

AI flags D4 Navigation on cost- and insurance-adjacent keywords

↓

0.452
development κ, the worst-performing dimension in the study; both candidate models below threshold

> *"AI classification should be done on the meaning of the posts, not just picking the words suggesting specific domain."*

REMEDIATION · RULE R12

Conceptual statement
→ codebook v1.5

Mechanical procedure
→ prompt v1.3 only

R12 — Domain 4 Active Navigation Test

Domain 4 requires an active, expressed navigation need. Not background context or emotional colouring. Flag D4 only if the navigation barrier is the explicit subject of the patient's question or expressed need.

The Removal Test. If the navigation-adjacent sentence can be removed and the primary question remains fully expressed, D4 does not qualify. The navigation element must be load-bearing.

Do not flag D4 if: financial hardship explains distress (D3); a clinic or specialist is named as the source of an information gap (D1); "what are my options" sits in a post primarily about emotional distress (D3).

D4 = N

"Tomorrow is transfer day. This is our last embryo and we cannot afford any more IVF cycles. I'm terrified. Everything is riding on this."

Remove the affordability sentence and the post still stands as an emotional support request.

D4 = Y

"My insurance just denied coverage for IVF, saying it's 'elective.' My clinic says I need to appeal within 30 days but I don't know how to start. What documents did you need?"

Remove the insurance content and nothing remains.

Five over-flags examined item by item (#24, #34, #57, #58, #73); both raters confirmed on structured review that these were AI errors, not codebook ambiguity. Worked examples from study codebook v1.5 — not verbatim patient posts.

AFTER

KEYWORD MATCHING	MEANING UNDERSTANDING
"insurance"	patient intent
↓	↓
D4 Navigation	D4 Navigation

0.769
re-validation on the same development sample

0.97
HELD-OUT — DATA THE FIX HAD NEVER SEEN

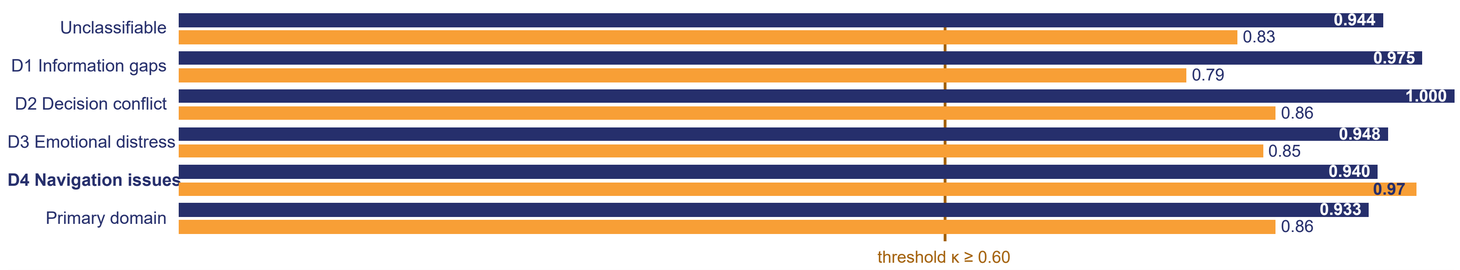
AI agreement with the two raters was 1.000 and 0.940.

3b RESULTS · VALIDATION

All six dimensions cleared the pre-specified threshold on unseen data

22,619 posts classified
80 held-out sample, unseen
2 independent human raters
97.3% valid classifications in production
6 / 6 dimensions above κ ≥ 0.60 on held-out

HELD-OUT REPRODUCIBILITY · N=80



AI–human κ was at or above its development-stage value on every dimension — refuting the interpretation that iteration had overfitted the development sample. AI–human κ is the mean of AI-versus-each-rater agreement.

MODEL SELECTION · FIXED BEFORE ANY RESULT

DEVELOPMENT K	OPUS 4.7	SONNET 4
Unclassifiable	**0.820**	0.750
D1 Information gaps	**0.750**	0.755
D2 Decision conflict	**0.601**	**0.544**
D3 Emotional distress	**0.607**	**0.511**
D4 Navigation issues	**0.452**	**0.482**
Primary domain	**0.670**	0.632
Cleared κ ≥ 0.60	**5 / 6**	**3 / 6**

Rule applied mechanically: Opus ≥ 0.60 and Sonnet < 0.60 on ≥1 dimension → Opus selected for production.

Human–human agreement at the development stage ranged 0.78–0.96. κ 95% CI half-width is approximately ±0.11 to ±0.13 at these sample sizes; the design targets threshold-based decision power rather than precise κ estimation. Results are conditional on a pinned model version (claude-opus-4-7) and are not directly portable across model versions or providers.

Production: 22,006 valid classifications; 613 errors retained in audit files (157 Reddit / 456 YouTube). No error was silently recoded as Unclassifiable. One mid-run failure recovered from checkpoint without item loss. Cost approximately 5× the rejected Sonnet path. Compared against roughly 1,500 double-rater hours for manual coding of the same corpus.

4 CONCLUSIONS

Separate codebook and prompt
Separating the instruments is what makes disagreement diagnosable rather than something to tune away.

Route every disagreement
AI literalism is a diagnostic probe. R10 and R11 came from human disagreement; R12 came from the AI's failure. A codebook ambiguity two experienced raters had been resolving silently, without ever noticing they were resolving it.

Validate on held-out data
The held-out partition is what distinguishes validation from fitting.

Deploy at scale
Locked instruments classified 22,619 posts at 97.3% validity, with every error retained for audit.

κ 0.45 DEVELOPMENT → **κ 0.97** HELD-OUT SAMPLE

HEADLINE FINDING

Prompt-only remediation improved agreement on completely unseen data. Moving the study's most error-prone dimension, D4 Treatment Navigation, from below threshold to near-perfect.

Funding · Aksigen IVF. No external funding.

AI-use statement · Classification performed by a large language model (claude-opus-4-7) under codebook v1.5 and prompt v1.3; validated against human coding on a held-out sample (n=80).

Disclosures · Authors are employees of the listed organisations.

RUSHABH@SIROAI.COM
Rushabh Mehta