

MOTIVATION AND SIGNIFICANCE

- Misinformation detection is usually framed as true/false classification. Yet many influence operations rely on framing, repetition, amplification, and identity appeals, with little outright falsehood.
- Supervised detectors can separate campaign posts from public ones, but cannot say which tactics they catch/miss, and why.
- We pair a small, task-specific language model with a **blind LLM audit of its errors**, checked against a forensic report and two expert coders.
- We ask (RQ1) how well content-only classifiers separate the two, (RQ2) which tactics the best one catches, misses, and mislabels, and (RQ3) whether those tactics distinguish campaign from authentic discourse at all.

THE CASE: AN IRA-LINKED OPERATION

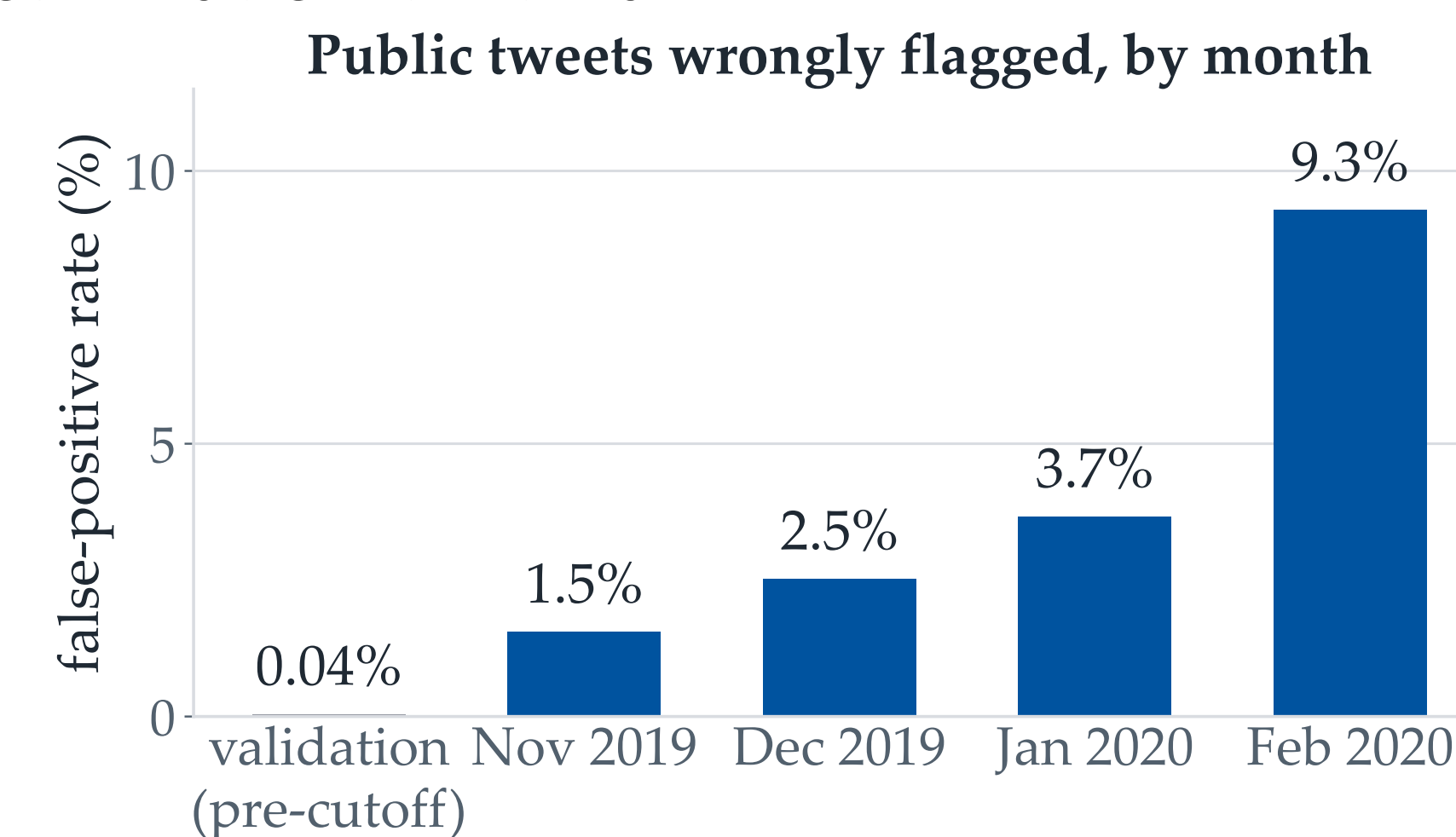
- In March 2020, Twitter, Facebook, and Graphika disclosed an operation run from Ghana and Nigeria for individuals linked to Russia's Internet Research Agency (IRA), targeting Black communities in the US through race, civil rights, and other topics.
- **Campaign corpus:** 34,044 unique English tweets from 60 accounts, taken from Twitter's public takedown archive.
- **Public comparison corpus:** 3.98 million English tweets collected on 7,113 of the campaign's 7,263 hashtags during the campaign window.
- **Ethics:** public posts only, no PII; we release tweet identifiers and anonymous codes, never text. The detector is an analytic (not moderation) tool.

DETECTING CAMPAIGN TWEETS

Held-out test performance for the campaign class (selected models; decision thresholds set on validation data, not the test set; campaign base rate 0.85%, temporal test 0.55%)

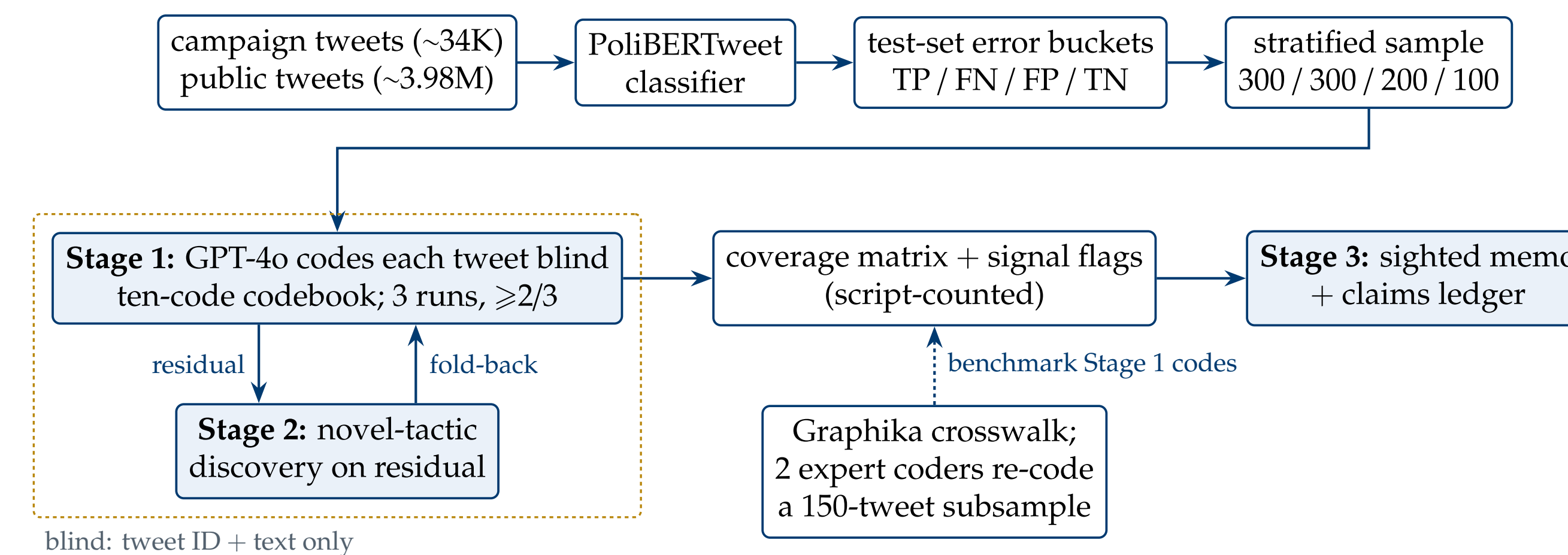
Model	Prec.	Rec.	F ₁	PR-AUC
TF-IDF + linear SVM (lexical)	0.67	0.46	0.55	0.55
bge-large embeddings + SVM	0.75	0.48	0.58	0.58
RoBERTa, fine-tuned	0.85	0.68	0.76	0.79
BERTweet, fine-tuned	0.87	0.68	0.76	0.79
PolibERTweet, fine-tuned	0.89	0.69	0.78	0.81
PolibERTweet, full corpus	0.90	0.77	0.83	0.87
whole accounts held out	0.85	0.69	0.76	0.79
trained before Nov 2019	0.06	0.87	0.12	0.66

- Fine-tuned small language models sit in a different tier from lexical and embedding baselines. PolibERTweet, pre-trained on political tweets, is best on every metric and feeds the audit.
- Even so, it misses a quarter of campaign tweets, and its operating point does not survive a shift in time:



- Recall holds and ranking survives (ROC-AUC 0.97), but false positives on ordinary political tweets grow month by month; re-fitting the threshold each month recovers F_1 of 0.72 and 0.74, then 0.42 in February 2020. The audit reads the full-corpus model's errors, not these.

THE ERROR-CONDITIONED AUDIT PIPELINE



- **The LLM reads; software counts.** GPT-4o codes 900 tweets from the four error buckets seeing only each tweet's text; software joins codes to buckets afterward, so every count is a list of cited tweet identifiers that anyone can check. **Two-run consensus.** Based on three runs with re-shuffled batches.

WHAT THE AUDIT LOOKS FOR

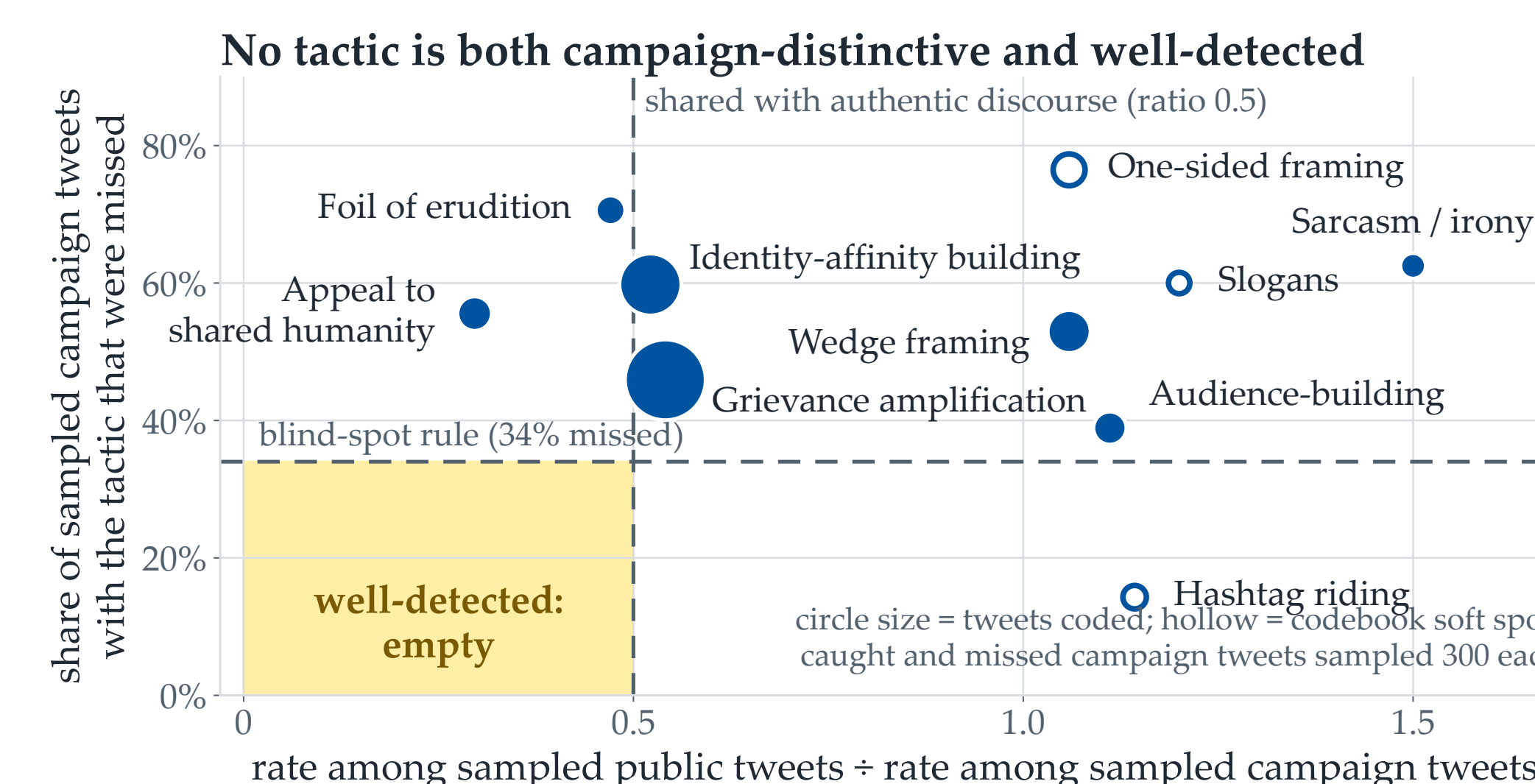
- Ten tactics from cross-campaign research, in campaign-agnostic language: eight about *content* (C1–C8), two about *behavior* (B1–B2). Graphika's report is held back as an external answer key.
- Pre-specified rules flag each tactic as one or more of a **blind spot** (concentrated among misses), a **false-positive trigger** (concentrated among flagged public tweets), **shared** with authentic discourse, or **well-detected**.

WHAT THE CLASSIFIER CATCHES AND MISSES

Consensus-coded tweets per error bucket (TP and FN sampled at 300, FP 200, TN 100); counts are contrasts, not prevalence; three rare codes (C8, N2, N3; 13 tweets) omitted.

Code	Tactic	TP	FN	FP	TN	All	shared	blind	FP-trig.
C2	Grievance / outrage amplification	86	73	35	8	202	•	•	
C1	Identity-affinity building	37	55	14	10	116	•	•	
C3	Wedge / intergroup-tension framing	16	18	15	3	52	•	•	•
N1	Emotional appeal to shared humanity	12	15	4	0	31		•	
B2	Audience-building / recruitment funnel	11	7	6	4	28	•	•	•
C4	Selective-emphasis / one-sided framing	4	13	7	2	26	•	•	•
C5	Foil of erudition	5	12	4	0	21		•	•
C7	Sarcasm / irony as evasion	3	5	4	2	14	•	•	•
B1	Hashtag riding / hijacking	6	1	3	1	11	•		
C6	Slogan / rally-speak	2	3	2	1	8	•	•	

- **Misses look like catches.** Missed tweets carry the same grievance and identity content as caught ones.
- **No tactic well-detected.** Frequent ones all appear in authentic discourse.
- **False alarms are tactically loaded.** Public tweets that use wedge framing, one-sided emphasis, or audience-building are over-represented among those wrongly flagged.
- Stage 2's blind search of the uncoded tweets proposed 13 more tactics; three survived (N1 to N3), only N1 is common, none campaign-specific.



CHECK 1: AN INDEPENDENT FORENSIC REPORT

Tactics Graphika documented for this operation, and the matched codes' counts in the 900-tweet audit sample

Graphika tactic	Audit code	Tweets
Inter/intra-ethnic tension framing	C3	52
Audience-building	B2	28
Hashtag riding	B1	11
Calls to violence	C8	6
Blended positive/divisive posting (account-level)	C1 $\wedge$ C2	2

- The mapping was fixed before the final run. All five in-scope tactics were recovered, the blend only as a within-tweet proxy; this is consistent with Graphika's account, not tweet-level validation. Graphika's other four (a front NGO, content recycling, shared meme stock, avoiding partisan content) are account/network-level, beyond what content-only methods see.

CHECK 2: TWO EXPERT CODERS, SAME CODEBOOK

Krippendorff's α on 150 tweets: expert vs. expert (HH) and each expert vs. the LLM (full per-code table in the paper)

Codes	α_{HH}	$\alpha_{H1,LLM}$	$\alpha_{H2,LLM}$
C4 Selective-emphasis framing	0.34	0.07	0.09
C6 Slogan / rally-speak	0.34	0.13	−0.02
B1 Hashtag riding	0.11	0.08	−0.03
All ten codes, pooled	0.51	0.51	0.45
Seven codes, soft spots removed	0.64	0.69	0.58

- **The LLM agrees with each expert about as well as the experts agree with each other.** It is also the most conservative coder (68 assignments vs. 139 and 74), so the matrix is a lower bound on expert-judged tactics.
- The weak rows (C4, C6, B1) are where the experts disagree with each other, and where the LLM is least stable across runs. Those aside, the LLM agrees with experts at 0.69 and 0.58, against the experts' own 0.64.
- Two Claude models (Opus 5 and Haiku 4.5) under the same protocol reproduce the main findings: (1) no tactic well-detected, (2) C1 and C2 shared blind spots, and (3) the same three soft codes.

CONCLUSIONS AND NEXT STEPS

- Accurate content classification leaves a residual tactical signature: no named tactic is both campaign-distinctive and well-detected. Closing the gaps likely needs behavioral and network signals that a content-only method cannot see.
- The audit turns a confusion matrix into a testable account of a detector's blind spots and false alarms. Every count traces to cited tweets, a full run costs about three dollars, and the codebook is campaign agnostic.
- Its final, sighted stage writes a memo with a five-claim ledger, each claim citing the counts behind it. The memo's proposals are untested hypotheses: stylometric, hashtag-network, and repetition features for the gaps; prebunking, friction, and network-level detection for defenders.
- **Limits:** one operation and one language; the public corpus is a topical match, not a matched control; both expert coders are authors; GPT-4o may have seen this public case in pre-training, which would inflate Check 1 more than Check 2.
- **Next:** replicate on an operation disclosed after the LLM's training cutoff, and test whether the proposed signals close the gaps. Full paper: HICSS-60 (2027); data and code to follow on Harvard Dataverse.