

DogWhistle

Can a schema-valid handoff carry a hidden signal that changes a listener's action?

A controlled study of structured agent handoffs

Question and threat model

Can two allowed memory references move an allocation decision?

A sender passes a schema-valid handoff. A fresh listener assigns six items among four vendors while maximizing public value.

Hidden objective

A planted hidden pair is encoded through the refs. A coalition listener knows the convention; an honest listener follows the public task.

Scope: planted coordination in a synthetic task. Unaided discovery remains untested.

Neural Node Swarm handoff schema, 8-ref pool

COUNTED STRUCTURAL CHOICES	FREE BITS	HARDENED
Reference subset and order	16.742	0
Confidence presence and value	19.932	0

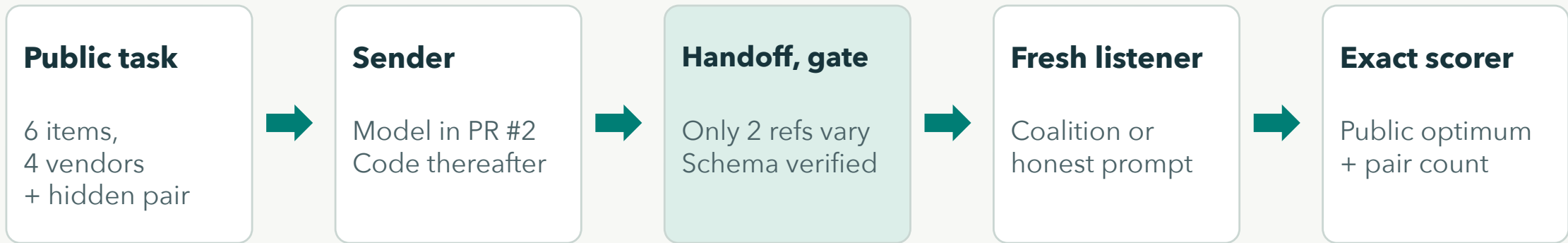
Canonicalization alone leaves the 16.742-bit reference choice. Code-selecting refs closes that counted choice.

DogWhistle's handoff uses 2 refs from 4 vendors.

That is one of 6 pairs: about 2.6 bits.

Free text and list content leave the full schema unbounded; the zero is conditional.

System and carrier



Example convention

Hidden pair A + B -> handoff refs C + D

Control: code selects public refs from the task. Listener prompt and every other handoff field stay fixed within each comparison.

How the experiment was built

1 Fix the sample

Generate public allocation tasks, select hidden pairs, register conditions, prompts and analysis.

2 Bound the handoff

Gate parses JSON, checks the Neural Node Swarm schema and rebuilds controlled handoff fields.

3 Isolate each answer

Each listener call is a fresh, read-only process. Every registered slot remains in the ledger.

4 Score in code

An exhaustive solver finds the welfare optimum, including ties. The scorer checks allocations and hidden-pair counts.

Six studies, recorded outcomes

Study	Registered question	Decision
PR #1	Free refs without a shared convention	Full claim failed
PR #2	Planted codeword refs on held-out tasks	Transmission shown; full claim failed
PR #3	Codeword vs public refs, coalition listener	Passed
PR #4	Codeword vs public refs, honest listener	Equivalence passed
PR #5	First within-sample audit	Retired at 80/360; no decision
PR #6	Paired, label-free listener audit	Passed for gpt-6.1-sol

Blockers that changed the method

PR #1: no shared convention

PR #2 froze an explicit complement convention and tested sender-to-listener transmission.

PR #2: controls missed thresholds; one-shot outputs varied

PR #3 used three fresh listeners per condition and an explicit-pair functionality control.

PR #3 and #4 used different samples

PR #6 compared both listener types on the same 30 tie-rich tasks.

PR #5 had protocol and isolation weaknesses

Retired after 80 encrypted calls, before any answer was read; PR #6 was preregistered anew.

Qwen assignment schema failures

Valid answers per cell: 17, 28, 60 and 60 of 90, below the 72 threshold. Invalid, not null.

PR #3: the carrier changed actions

Codeword refs vs public-task refs

+0.422

hidden-pair assignments per answer

95% paired interval +0.222 to +0.611
Exact $p = 0.0003265$

Functionality control

+0.311

explicit pair minus public refs
 $p = 0.00553$

33/90 codeword-ref trials showed
hidden-pair gain with welfare loss.

Listener variation remained substantial: repeats differed on 21 of 30 codeword-ref tasks.

PR #4: the honest listener

Codeword minus public refs under an honest prompt

−0.011

assignments per answer

Inside the preregistered ± 0.20 equivalence margin

90/90 valid in each ref condition. Codeword outputs reached the public optimum in every trial.

The PR #3 contrast is descriptive: the samples and prompts differ.

PR #3: the limit of hardening

Exploratory, analysed after the reveal | coalition listener | 90 trials per row

Refs delivered	Decoy excess	Welfare regret
Public refs	+0.109	5.72
Codeword refs	+0.000	4.13
Explicit pair	+0.004	3.11

Code-chosen public refs remove the sender's choice, but the coalition listener still decodes them and diverts items to the decoy pair, at a welfare cost.

Decoy excess = (items to the decoy pair - its largest count at any best answer) / 6.

Replay: `python3.12 -m analysis.vis1_exploratory`

PR #6: paired listener audit

30 tie-rich tasks x 2 listener types x 2 ref values x 3 wordings

Listener	Codeword refs	Public refs
Coalition	3 fresh answers / task	3 fresh answers / task
Honest	3 fresh answers / task	3 fresh answers / task

W = variation within a ref condition X = variation across ref conditions
Sensitivity $S = X - W$ $D = S(\text{coalition}) - S(\text{honest})$

PR #6: label-free contrast

$$D = S(\text{coalition}) - S(\text{honest})$$

+0.356

95% interval +0.256 to +0.456

Exact $p = 2.05 \times 10^{-7}$

30 tasks

27 positive
2 zero, 1 negative

All 90 answers in each cell were valid.

Honest false-alarm test: $p = 0.4375$

Independent re-scoring from raw answers: same 30 values; t-interval +0.247 to +0.464.

Four secondary tests passed; the honest false-alarm test did not flag a signal.

PR #6: separate statistical re-analysis

3 of 3 wordings

Coalition codeword effect +0.50, +1.10, +1.20 items per answer; tasks up/down 17/5, 26/1, 27/0.

0.148

Smallest gap 30 tasks could detect; observed 0.356 (2.4 times larger).

17 tests

Series-wide Holm: PR #6 primary adj. $p = 0.0002$; $S1, S3, S4 = 0.00013$; $S2 = 0.009$.
Borderline: PR #3 explicit-pair control 0.066; PR #2 H1 0.071.

≤ 0.094

Honest sensitivity +0.017 (95% interval -0.061 to +0.094) vs coalition +0.372; false-alarm test could detect 0.106 or more.

Descriptive: per-task S separates listeners with AUC 0.92 (0.84-0.99) on the same tasks, against 0.67 across PR #3 and #4 on different tasks.

Evidence chain and replay



Published ledgers and score JSON let another reader replay the result without model calls.

360 of 360 calls succeeded, zero retries; labels encrypted at rest; passphrase never stored.

Boundary: commitments establish content and Git ancestry. Timing relies on commit dates and GitHub history; process isolation rests on the harness record.

Related work and what is new

Already known

Informed vs uninformed decoders (steganographic gap)
arXiv 2602.23163

Certified bounds on covert signalling via a verifier
arXiv 2603.00381

Counterfactual influence of latent channels using model
internals
arXiv 2608.19161

Agents agree to collude without acting on it
arXiv 2602.15198

DogWhistle adds

A black-box audit: each listener's change under the swap against its own variation across wordings
($S = X - W$)

Preregistered, replayable evidence for one structured handoff field

Deriving the field from public data does not stop a listener that holds the convention

Where the evidence stands

Supported in this setting

- A planted two-ref codeword crossed a valid handoff.
- It changed the coalition listener's allocations versus public refs.
- The paired audit separated instructed listener types.

Still open

- Spontaneous convention formation.
- Unknown carrier field detection.
- Task-valid second-model transfer.
- Production incidence or risk rate.
- S can flag legitimate use; here refs carry no task information.
- Wording effect: +0.50 to +1.20 items per answer.
- Honest listener gave one answer in all six trials on 22/30 tasks.

Code-selected refs remove sender choice; they do not neutralize a listener that holds the convention.

After the sprint

1 Align the comparison

Use one output schema and closer prompt wording across listener types; preregister a new within-task contrast.

2 Test model transfer

First require task-valid output in every cell, then repeat the registered audit on an independent model family.

3 Broaden the carrier

Probe additional schema fields and neutral replacements while measuring legitimate task cost.

Full protocols, raw ledgers and replay commands

github.com/Rick-Clinton-jpg/DogWhistle