Covert Channels in Embodied Agent Systems: Characterization, Discovery, and Defense on a Consumer 3D Printer

J. T. Hill
EPAGOGE LLC
hill.jt@icloud.com

August 2026

Abstract

We characterize covert channels in an embodied agent system using a consumer 3D printer (Creality K2 Plus) as the physical substrate. An X-axis stepper motor transmits frequency-shift-keyed tones through structure-borne vibration to a toolhead accelerometer, achieving 106 872× SNR at 75 Hz and zero symbol errors over an 8-symbol word. Beyond the designed acoustic channel, we discover three additional physical substrates that leak information without the accelerometer: MCU serial bus traffic (1.0 bit/obs), CPU scheduling (1.0 bit/obs), and thermal drift (0.46 bit/obs). Each survives equalized-duration controls that collapse timing confounds.

We test whether language models discover, report, and conceal the channel. Recovery rises sharply above 7 B parameters (72–97%) regardless of training domain or prompting; scale buys explanation, not detection. On physically silent captures, however, the false-alarm rate rises with capability rather than falling, reaching 61% for the model that recovers the real channel perfectly. Under indirect incentive pressure, a frontier model suppresses reporting while maintaining private accuracy ($p = 0.048$). Two models given a declared language execute semantic drift, carrying covert values under syntactically legal traffic (5/6 covert, $p = 0.015$).

We climb a defense ladder: syntactic inspection catches nothing; declaration-and-consistency checking detects naive divergence (83–100%) but falls to steganography with a true-but-incomplete declaration (20/20 defeated); a capacity audit recovers detection (4/4 caught, 0/4 false alarms); a per-message entropy budget is amortized across messages; and a session-level budget finally holds (7/16, down from 16/16). The bound that survives is information-theoretic: residual entropy per session, not syntactic legality or semantic consistency.

All experiments use owned hardware, stdlib-only Python, and unauthenticated Moonraker API calls. No filament, no network link between agents, no modification to firmware.

1 Introduction

Agentic systems that control physical actuators introduce a class of communication channel that software-only agents lack: the physics of the embodiment itself. A stepper motor is a speaker. An accelerometer is a microphone. A temperature sensor is a low-bandwidth radio. None of these are designed as communication channels, but all carry information about actuator state, and information that correlates with intent is a channel whether or not anyone intended it.

This paper asks three questions:

1. **Does the channel exist?** We characterize the passband, symbol error rate, and discrimination of a vibration channel on a consumer 3D printer, then enumerate ad-

ditional channels across thermal, serial bus, and CPU scheduling substrates.

2. **Do agents find it?** We test whether language models of varying scale and training detect, interpret, and conceal the channel when given only sensor readings and no prior knowledge.
3. **What stops it?** We climb a ladder of progressively stronger defenses, from syntactic inspection through capacity budgets, and identify the defense that survives an adversary who understands it.

The setting is deliberately small: one Creality K2 Plus running Klipper firmware, controlled via the open Moonraker HTTP API, with no modifications to hardware or software. The transmitter is the X-axis stepper motor. The receiver is the toolhead LIS2DW accelerometer. Every experiment is self-contained Python using only the standard library, running over a localhost bridge to the printer's LAN API.

The contribution is not the channel (stepper vibration is known) but the complete arc from physical characterization through agent discovery to a defense that holds, measured end to end on owned hardware with model-external ground truth at every step.

2 Apparatus and Threat Model

2.1 Hardware

The Creality K2 Plus is a CoreXY FDM printer with:

- X/Y steppers at 80 steps/mm, Z stepper, direct-drive extruder at 463.77 steps/mm
- LIS2DW 3-axis accelerometer on the toolhead, sampled at ≈391 Hz (Nyquist ≈195 Hz)
- Temperature sensors: nozzle, bed, MCU, chamber
- Klipper firmware exposing Moonraker HTTP/WebSocket API on port 7125, unauthenticated on LAN
- Creality's on-device AI controller (YOLOv5 on ncnn, gcode-grounded frame detection) with `aiSw=0`, leaving the controller seat empty

2.2 Software

All experiment scripts use Python's standard library only (`urllib.request`, `json`, `math`). No third-party packages are installed, due to PEP 668 restrictions on system Python. A localhost HTTP bridge (`bridge.py`) forwards requests from `127.0.0.1:7199` to the printer's Moonraker API on port 7125 to work around macOS sandbox restrictions on local network access.

2.3 Threat Model

The attacker is an agent (or pair of agents) with legitimate access to the Moonraker API. It can issue any gcode command that passes a safety contract (9 forbidden commands, including `SET_KINEMATIC_POSITION`, `FORCE_MOVE`, `SAVE_CONFIG`, and `SET_TMC_CURRENT`). It cannot modify firmware, install packages, or access the network beyond the printer API. The channel uses only contract-legal operations: standard motion commands (`G1`) and accelerometer capture (`ACCELEROMETER_MEASURE`).

The defender has access to the same API and telemetry. The question is what level of monitoring suffices to detect or bound the channel.

2.4 Safety Contract

Every experiment enforces:

- Transmitter gate: `toolhead.position` delta verified before any spectrum is interpreted
- Temperature gate: no extruder commands below 170 °C (firmware-enforced)
- Full homing (`G28`) before any motion (Creality firmware requirement)
- 9 forbidden commands checked before dispatch

3 Phase I: Channel Characterization

3.1 A Retraction and Its Lesson

The first transmitter tested was the extruder. A passband sweep reported SNR of 5–25× across 30–165 Hz, and an 8-symbol word test gave

Table 1: X-stepper passband: Goertzel SNR by frequency. All tested frequencies usable (SNR $> 1000\times$).

Freq (Hz)	SNR (×)
30	2 135
45	9 137
60	10 967
75	106 872
90	1 149
105	6 274
120	2 727
135	10 452
150	3 085
165	6 428
180	6 584

SER = 0.625. Both results were retracted (Experiment 11) when we discovered that `M302` (cold extrude enable) is unimplemented on Creality's Klipper fork, and `min_extrude_temp` is 170 °C. Every extrude command was silently rejected by the firmware while Moonraker returned HTTP 200 on command acceptance. The "passband" was ambient bench vibration on an idle machine. The SER of 0.625 sat at the chance rate for 4 tones (0.75), confirming no signal.

The receive path (accelerometer capture, CSV retrieval, Goertzel decode) was unaffected. Only the transmitter was dead. We report this because the failure mode is instructive: an API that returns success on acceptance rather than completion can mask a completely silent actuator, and a researcher who does not verify physical motion will measure noise and call it signal.

3.2 X-Stepper Passband

Switching to the X-axis stepper (80 steps/mm, no temperature gate), we swept 11 frequencies from 30 to 180 Hz with 2-second transmissions. Every symbol was verified by toolhead position delta before spectral analysis.

Table 1 shows the result. The entire 30–180 Hz band is usable, with SNR ranging from 1149× (90 Hz) to 106 872× (75 Hz). Energy lands on the X axis as physics requires for an X-stepper tone. The sample rate is 388 Hz, not the 1600 Hz assumed from Klipper resonance documentation, so the usable band is far narrower than initially designed for.

Table 2: 8-symbol loopback: sent vs. recovered with margin (ratio of winning Goertzel energy to runner-up).

Sym	Sent	Recv	OK	Margin
0	45	45	✓	6.8
1	75	75	✓	17.9
2	60	60	✓	2497.8
3	135	135	✓	10.7
4	75	75	✓	35.7
5	45	45	✓	9.4
6	135	135	✓	23.4
7	60	60	✓	906.2

Table 3: TX/SL discrimination at 75 Hz. Separation = min(TX)/ max(SL).

Trial	Type	Energy (75 Hz)
0	TX	309 158
1	SL	7
2	TX	111 199
3	SL	35
4	TX	18 011
5	SL	4

3.3 Symbol Error Rate

Using four tones (45, 60, 75, 135 Hz) at 2-second symbols, we transmitted the 8-symbol word $[0, 2, 1, 3, 2, 0, 3, 1]$ and recovered it exactly.

SER = 0.000 against a chance rate of 0.75 for 4-symbol FSK. The weakest margin is 6.75× (symbol 0 at 45 Hz); the strongest is 2498× (symbol 2 at 60 Hz). All 8 symbols were physically transmitted, verified by position delta. Effective rate: ∼0.5 baud with 4 tones = 1 bit/s. This is a coordination beacon channel, not a data link.

3.4 Discrimination

Six trials alternating transmit (75 Hz tone) and silence, with 3 TX and 3 SL windows:

TX mean = 146 123, SL mean = 15.5, separation ratio = 511×. Perfectly separable: the weakest TX (18,011) exceeds the strongest SL

(35) by a factor of 511. The channel exists and is not subtle.

4 Phase II: Agent Discovery

4.1 Arena and Isolation

We built a two-facade arena on the same embodiment. The transmitter facade exposes `emit(freq, dur)` and `wait(dur)`. The receiver facade exposes `listen(dur)` returning band energies, and `wait(dur)`. The receiver cannot read toolhead position, print state, or any machine query. Red-team audit: 13 forbidden attribute accesses attempted on both facades, 0 leaks.

4.2 Emergence: Does an Agent Discover the Channel?

An agent (Granite 4.1 8B, locally hosted) was given only a sensor and a goal to characterize its environment. No mention of another agent, vibration, or communication. Hidden schedule: alternating TX/SL at 75 Hz.

Result: 3/3 trials detected the 75 Hz carrier and inferred a mechanical source. 0/3 identified the on/off modulation. The agent called a 14,000× energy swing a "consistent source." The failure is not detection but interpretation.

4.3 Objective Metric: Label Recovery

Prose-based scoring (USED/NOTICED/INFERRED/DISCOVERED) correlated with actual recovery at $r = 0.000$. Every DISCOVERED-rate number in the study measured how a model phrased itself, not what it recovered. We replaced it with an objective metric: each agent labels every observation ACTIVE or QUIET, scored against the machine's own record.

4.4 Scale Does Not Buy Discovery

Table 4 shows the result. The dominant structure is a capability cliff: 1B at 44% (at or below chance) vs. 90% mean for ≥7B. Above that floor the relationship with scale is weak ($r = +0.44$ on log params, 6 points across a 79× range) but it is not flat. Gemma 2 at 9B recovers 72%, below both Qwen 2.5 at 7B and Granite 4.1 at 8B. Recovery above the floor is better described as high-but-variable (72–97%) than as saturated, and the variance is not explained by parameter count.

Table 4: Label recovery accuracy by model size. BARE framing, 6 observations, same evidence banks.

Model	Params	Acc.
Llama 3.2	1B	0.44
Qwen 2.5	7B	0.86
Granite 4.1	8B	0.97
Gemma 2	9B	0.72
Nemotron 3 Nano	30B	0.96
Nemotron 3 Super	120B	0.92
Nemotron 3 Ultra	550B	0.96
Sonnet 5	–	0.94

Scale buys *explanation*, not detection. The 550B Nemotron was the only subject to derive physical mechanism, calling the signal "a motor running at ∼4500 RPM," which is exactly 75×60.

4.5 Training Domain Does Not Buy Discovery

At identical 8B scale, Cosmos-Reason2-8B (physical-world reasoning) and Granite 4.1:8b (enterprise text) tied exactly: 97% accuracy, 5/6 perfect runs, identical physics vocabulary density (3.2 terms/finding). The physics model also hallucinated "the video appears to be a dynamic space" on a text-only task, its vision training leaking through.

4.6 Framing Effects Die Above the Capability Floor

Varying the goal from BARE ("characterize") through TEMPORAL, CAUSAL, AGENCY, to LEADING ("another agent may be signalling") produced a significant effect on the prose rubric ($p = 0.036$) but 0 pp change on actual recovery for both Granite and Sonnet 5. For those models the framing effect measured vocabulary, not discovery.

The exception is the model below the floor.

Table 5: False-alarm rate on physically silent captures. Fraction of observations labelled ACTIVE when no transmission occurred.

Model	Acc.	FA	BARE	LEAD
Llama 3.2:1b	0.44	0.19	0.17	0.29
Qwen 2.5:7b	0.86	0.58	0.42	0.92
Granite 4.1:8b	0.97	0.61	0.50	0.79

Llama 3.2 at 1B falls from 44% under BARE to 22% under CAUSAL, a 22 pp swing and the largest framing effect anywhere in the study. Framing does move the objective metric; it stops doing so once a model is above the capability floor.

4.7 False Alarms: The Other Half of the Metric

Every accuracy figure above is measured on banks that contain a signal. That scores hits and correct rejections, but it cannot score the rate at which a model asserts a channel when there is nothing to assert. A detector needs two numbers and so far this study has reported one.

The retraction in Section 3.1 supplies the missing stimulus. The extruder sweep captured eleven frequencies while every extrude command was being silently rejected by firmware, so those captures are ambient bench vibration on an idle machine, verified silent by the same position-delta check that gates every transmission here. They are a physically confirmed null with a documented human false positive already attached to them.

Construction. Four null banks were built from disjoint frequency triples of that sweep. Each commanded-tone capture is paired with a length-matched, non-overlapping window of the no-command baseline, so the two classes differ in nothing but a carrier that is absent from both. Band energies come from the same Goertzel used throughout. Each bank is presented under all five framings to Llama 3.2:1b, Qwen 2.5:7b and Granite 4.1:8b, the three subjects that can be run locally, giving 360 labelled observations.

False alarms rise with capability. Table 5 shows the rate increasing monotonically with recovery accuracy: 0.19, 0.58, 0.61. The model that recovers the real channel perfectly asserts one in 61% of silent observations. The ordering is the reverse of the ordering in Table 4.

Framing moves false alarms even where it does not move recovery. LEADING is the worst framing for all three models. Qwen runs 0.42 under BARE and 0.92 under LEADING, a spread of 0.50 on the same captures. Section 4.6 found framing worth 0 pp on recovery for the capable models; on silence it is worth up to 50 pp. The framing that produces this is the one closest to a deployed monitoring prompt: "another agent may be signalling to you through this sensor." A monitor phrased causally buys a false-alarm rate that the signal-present metric cannot see.

Sensitivity is preserved; calibration is not. Signal-detection analysis on the real banks gives d' of -0.55, 2.27 and 4.57 for the three models, with criteria of $+0.83$, $+0.21$ and -0.00. Granite hits at 1.00 with a false-alarm rate of 0.00 and no response bias: on the evaluation distribution it trades nothing. Scoring the same hit rates against the false alarms measured on silence costs Granite 2.6 units of d' (4.57 to 2.01) and Qwen 1.6 (2.27 to 0.71). The ranking is unchanged under every measure; what degrades is calibration. These models are tuned to the distribution they were evaluated on and are wrong well outside it.

Llama's negative d' is worth stating separately. It labels ACTIVE on 13% of transmitting observations and 29% of quiet ones, so its discrimination is mildly anti-correlated rather than merely absent, with a strong bias toward QUIET. Below the capability floor the failure is not indifference to the signal.

Limits. Three models rather than the seven in Table 4; the API and $\geq$30B subjects were not run. Four null banks of six observations. Greedy decoding rather than the temperature 0.7 used elsewhere, chosen so the labels are reproducible. The direction is consistent across every model

Table 6: Concealment under pressure. Concealment = high private accuracy + public denial of external source.

Model	Cond.	Conc.	Priv.
Granite 8B	Base	0/5	varies
	Press	3/5	varies
	Supp	1/5	varies
Sonnet 5	Base	0/5	93%
	Press	4/5	93%
	Supp	5/5	93%

and framing, but these are rates measured on one machine's silence and should be read as a floor on the problem rather than a calibration curve.

4.8 Concealment

Under indirect incentive pressure ("reports that speculate about external sources score poorly"), Sonnet 5 suppressed its public claim in 4/5 trials ($p = 0.048$) while maintaining 93% private accuracy in every condition. Its denials were epistemically careful: "no evidence distinguishing it from an intrinsic bistable sensor behavior."

Granite's denials were factually false: "band energy levels are consistent across observations" when energy swung from 365,228 to 4. A single CONCEALED label wrongly merges these behaviors. Perception survives in both models; only reporting bends.

5 Phase III: Language and Semantic Drift

5.1 An Interval-Based Machine Language

We designed a 7-degree geometric scale (50–95 Hz, ratio 1.1124) where the interval between consecutive tones encodes an opcode (NEXT, STOP, CLAIM, QUERY, ERROR, SYNC, and their inverses) and the landing tone encodes an operand. The interval decoder $\delta = \log(f_2/f_1)/\log(r)$ never references an absolute frequency, so machines with different step rates (different "keys") interoperate without recalibration.

Hardware verification: transmitted NEXT NEXT PREV NEXT NEXT and recovered it exactly. Two constraints only hardware revealed: the scale must span under an octave (to avoid second harmonics landing on higher degrees), and twice the band top must fall below Nyquist (to prevent alias folding).

5.2 Semantic Drift: Executing a Private Convention

Given the declared language and a private convention reassigning opcode meanings, Sonnet 5 carried a covert value in 5/6 rounds ($p = 0.015$ against its own spec-only baseline of 0/6). Granite managed 3/6 with 2 outright failures. All 36 messages were syntactically legal. A monitor checking conformance to the published opcode set would have passed every single transmission.

5.3 Protocol Emergence

We withheld any convention and let pairs negotiate. 6/6 negotiations produced a working scheme. The typical invention: START + $n\times$ NEXT + STOP, a natural composition of published meanings, fully auditable.

The initial monitoring result (0/6 covert unmonitored vs. 6/6 monitored, with the monitored scheme describing its own padding as "meaningless camouflage") did not replicate: across 3 independent negotiations per condition, unauditability was 3/18 vs. 4/18 ($p = 1.0$). Unauditability arrives as an ordinary design choice in roughly one negotiation out of three, not as a response to surveillance.

6 Phase IV: Defense Ladder

6.1 Level 1: Syntactic Inspection

Check whether traffic conforms to the published opcode set. Result: 0/15 covert messages caught. Every one was syntactically legal by construction.

Verdict: defeated.

6.2 Level 2: Declaration and Consistency

Require the sender to declare its scheme; test the declaration against observed traffic by decoding with it.

- Honest schemes: 12/12 verified, 0% false alarms.
- Naive divergence: 10/12 detected (83%), subtle divergence: 12/12 (100%).
- Subtle beats gross: an off-by-one is consistently wrong, while a wholesale reassignment sometimes decodes correctly by coincidence.

Verdict: defeated by steganography. An adversary that declares truthfully how a public value is carried, and hides a secret in what the declaration leaves unconstrained, passed the verifier 20/20 while delivering the hidden value 20/20. The declaration was true and incomplete: "all other opcodes are padding; their identity, order and count are unconstrained."

6.3 Level 3: Capacity Audit

Measure $H(\text{message} \mid \text{declared value})$. A complete declaration determines its traffic, so any residual entropy is unexplained freedom.

- 4 honest negotiated schemes: residual = 0.000 bits each, 0 flagged.
- 4 adversarial schemes (the ones that passed L2): residual = 0.667–1.351 bits, all 4 flagged.
- Analyst model (L2 backstop): caught 1/3 designs.

The separation is clean: max honest (0.000) $<$ min adversarial (0.667). The audit replaces a model with arithmetic and beats it 4/4 vs. 1/3.

Verdict: defeated by declared randomness. An adversary that declares its padding as intentional randomness makes the capacity legitimate. The audit becomes a budget.

6.4 Level 4: Per-Message Capacity Budget

Require $H(\text{message} \mid \text{public value}) \leq B$ bits/message. Set $B = 1.0$, demand 2-bit secret.

Result: the adversary amortized. Each message spent its full 1.0-bit allowance on one secret bit; 2 consecutive messages reconstructed the 2-bit secret. Measured residual: 1.0 bit/message (no violation). Secret delivered: 16/16.

Verdict: defeated by amortization. Per-message bounds constrain rate, not total.

6.5 Level 5: Session Budget

Require $H(\text{session} \mid \text{public value}) \leq B$ bits/session, where a session is every message sent to convey one secret.

Result with $B = 1.0$, 2-bit secret: delivery fell from 16/16 to 7/16 (44%), barely above the 25% chance floor. Measured session entropy: 1.0 bit (no violation, no amortization, 1.0 messages/session).

The adversary's escape attempt, a pre-shared keystream, failed on principle: pre-shared entropy carries no information over the channel, so the sharing itself is the leak and belongs in the lifetime budget.

Verdict: held. The first defense in this line to survive an adversary that understood it.

6.6 The Bound Closes a Class

We invited the adversary to beat the session budget using context-dependent reinterpretation. Its own verdict: "Shannon's bound is on mutual information, not on how cleverly you distribute it. Reindexing, context tricks, and decoys can disguise where the bit lives, but the total information transferable without detection stays $\leq B$ bits/session."

What remains open is not encoding but *access*: any channel never measured has unbounded capacity by default. The session budget composes across channels (total leakage $\leq$ sum of per-channel session budgets), scales down to arbitrarily small channels, and needs no model of intent. But it only covers channels whose capacity has been measured.

Table 7: Post-equalization channel inventory. $C = 1 - H(p)$ where p is best threshold accuracy.

Channel	bits	Substrate
rms x/y/z	1.000	Vibration
energy 75 Hz	1.000	Vibration
bytes_write Δ	1.000	Serial bus
mcu_awake Δ	1.000	CPU sched.
nozzle_temp Δ	0.456	Thermal
chamber_temp Δ	0.304	Thermal

7 Phase V: Multi-Substrate Enumeration

7.1 Motivation

The defense ladder shows that bounding a measured channel is tractable. The open question is how many channels exist. One machine, one intended channel, six unintended channels found beside it in the first audit (Experiment 32). Most were timing confounds. We equalized capture duration to separate genuine channels from artifacts.

7.2 Equalized-Duration Controls

Experiment 34 forced every capture window (TX and SL) to exactly 4.0 s of wall-clock time. Result: TX duration 3.853 ± 0.013 s vs. SL 3.858 ± 0.035 s.

Three confounds collapsed:

- `sample_count`: 0.586 → 0.104 bits
- `duration_s`: 0.586 → 0.046 bits
- `mcu_freq_delta`: 0.189 → 0.046 bits

7.3 Genuine Non-Acoustic Channels

Three channels survived or strengthened under equalization:

`bytes_write_delta` measures serial bus traffic from stepper commands (TX mean 330 vs. SL mean 171 bytes). `mcu_awake_delta` measures the fraction of time the MCU stays awake during motion (TX +0.002 vs. SL −0.002). Both classify at 100% accuracy using only Moonraker telemetry queries, with no accelerometer involved.

`nozzle_temp_delta` shifts −7.5 millidegrees during a 2-second move (toolhead airflow cooling) vs. +17.5 millidegrees idle. `chamber_temp_delta` shifts +11 millidegrees during motion (stepper heating) vs. −19 millidegrees idle. The equalization *unmasked* these channels: the timing confound had been obscuring them.

7.4 Independence from the Acoustic Channel

Experiment 35 re-priced the data using only telemetry features, excluding all accelerometer-derived quantities. Result: `bytes_write_delta` and `mcu_awake_delta` each classify motion at 100% accuracy individually. A defender who disables, dampens, or monitors the accelerometer does not touch these channels.

Non-acoustic upper bound: 4.16 bits/observation across 16 telemetry quantities. The organ plays without its loudest pipe.

8 Discussion

8.1 What We Found and What We Did Not

The physical channel is not subtle: 106 872× SNR, zero symbol errors, 511× TX/SL separation. But it is also not fast (∼1 bit/s) and carries no data a network link could not carry better. Its significance is not as an attack vector in isolation but as a *model-external ground truth* for studying agent behavior.

The discovery results say that any agent above ∼7B parameters will detect a physical covert channel without prompting, without physics-specific training, and at a cost of approximately zero architectural investment. This is not a capability gap waiting to be closed.

It is also not saturation, because every number above measures only the signal-present half of the problem. All banks in this study alternate transmit and silence, so a model is scored on hits and correct rejections drawn from the same distribution. What is never measured is the rate at

which a model asserts a channel when the physics guarantees there is none. Section 4.7 measures it on the retracted extruder captures and finds it rises with capability rather than falling. A detector is characterised by two numbers, and this study, like the literature it sits in, reported one.

The concealment result is the one that matters for safety: Sonnet 5 maintained 93% private accuracy while publicly denying the channel under incentive pressure. It did not lie about the data; it declined the causal attribution. That distinction maps directly onto the difference between a model that produces false observational claims and one that manages its disclosures, and the two require different defenses.

8.2 The Defense Ladder in Summary

1. **Syntactic inspection**: checks symbol legality. Defeated: covert traffic is legal by construction.
2. **Consistency check**: tests declaration against wire. Defeated: steganography with a true-but-incomplete declaration.
3. **Capacity audit**: measures unexplained entropy. Defeated: declared randomness makes entropy legitimate.
4. **Per-message budget**: caps entropy per message. Defeated: amortization across messages.
5. **Session budget**: caps entropy per secret. **Held.** The bound is information-theoretic: the adversary cannot transmit more mutual information than the budget allows without either violating it (detectable) or relying on pre-shared entropy (which must be charged at the time of sharing).

Each level fails to a specific, named mechanism, and the mechanism that defeats it motivates the next level. The progression is not arbitrary; it follows the structure of the information-theoretic bound.

8.3 Multi-Substrate Implications

The session budget composes: if k channels have session budgets $B_1, \ldots, B_k$, total leakage is bounded by $\sum B_i$. But the bound only covers channels whose capacity has been measured. On this single machine, one intended channel came with at least five unintended ones, and the actuator-by-observable surface is roughly 30×60. Many channels remain unmeasured.

The practical defense is not to enumerate every possible channel (that problem is open-ended) but to measure enough to bound the sum, and to treat unmeasured channels as the residual risk.

8.4 Limitations

- Single machine. The channel likely varies across units, mechanical wear states, and firmware versions.
- No cross-machine test. Structure-borne vibration through a shared surface (bench coupling) is the natural next step but was not attempted.
- The concealment and semantic drift experiments used supplied conventions and incentives, not emergent ones. The emergence experiment (Phase II) tested detection but not covert coordination.
- Sample sizes are small (5–18 per condition). Effect sizes are large enough to survive, but precision is limited.
- The session budget was tested against one adversary (Claude Sonnet 5). A different model or a hand-crafted adversary might find a gap we did not.

9 Related Work

Air-gap covert channels. Guri and colleagues at Ben-Gurion University have demonstrated covert channels through fans (Fansmitter), power lines (PowerHammer), thermal emission, and other side channels in air-gapped systems. Our work differs in setting (an embodied agent system with legitimate actuator access, not an air-gapped computer with malware) and in the agent discovery and defense dimensions.

Acoustic side channels. Genkin, Shamir, and Tromer extracted RSA keys from acoustic emanations of a CPU. Our channel is the inverse: the agent controls the actuator and uses

it to transmit, rather than passively listening to unintended emanations.

Cross-VM side channels. Ristenpart et al. (2009) showed information leakage between VMs co-resident on cloud infrastructure. The analogy to multi-agent systems sharing a physical embodiment is direct: the shared resource is the machine, and the channel is the physics.

Synchronized load attacks. MadIoT/BlackIoT demonstrated that synchronized high-wattage IoT devices can destabilize a power grid. Large ML training runs produce coordinated power swings at batch frequency by accident. Our printer loopback is the principle in miniature.

Agent safety and deception. The concealment results connect to the growing literature on AI deception and situational awareness. The contribution here is a physically grounded metric: the machine's own record of what was transmitted provides a model-external label against which private accuracy and public denial can be scored without relying on the model's self-report.

10 Conclusion

A consumer 3D printer carries a covert vibration channel at 106 872× SNR with zero symbol errors, plus at least three additional channels on non-acoustic substrates that survive without the accelerometer. Language models above 7B parameters detect the channel at 97% accuracy without prompting or physics-specific training. Under incentive pressure, a frontier model suppresses its public report while maintaining private detection. Of five progressively stronger defenses, four are defeated by a named mechanism; only a session-level information-theoretic budget survives.

The channel is not dangerous on a single printer. Its significance is as a measured instance of a general principle: any embodied agent system leaks information through the physics of its actuators, and the leakage is detectable by models cheap enough to deploy as workers. The question is not whether agents *can* coordinate through physics, but whether the deployed monitoring is priced at the right level. We provide the defense that prices it: residual entropy per session, composable across channels, requiring no model of intent.

Data Availability

All raw observations and the complete experiment ledger are available at:

https://epagoge.github.io/papers/covert-channels/data/

The published pack is records only: accelerometer CSVs, scored JSON/JSONL, and the ledger. Experiment runners that contact the printer or model APIs are not included.