Agreement Is Not Independence

Lessons from an LLM Ensemble Failure

Matthew J. Harmon · September 2026

A multi-model ensemble built to classify hostile web traffic from honeypot telemetry had a design flaw that its own metrics could not see: high voter agreement was masking constant, degenerate outputs rather than signaling consensus. This write-up covers why the system failed, how the network and hardware constraints shaped that failure, and what to carry into the next multi-model classification pipeline.

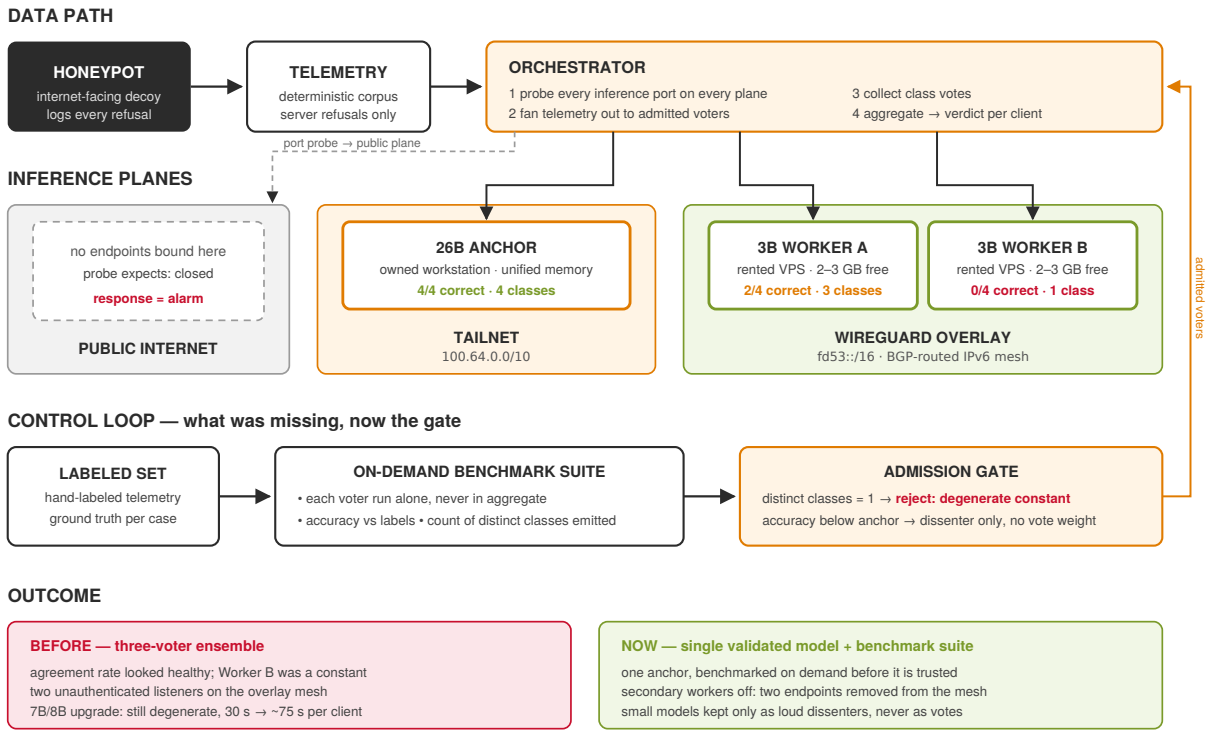

Figure 1 – The whole system: data path, inference planes, the benchmark-and-admission loop that was missing, and the before/after state.

1 System at a Glance

Component	Role	Where it runs
Honeypot	Internet-facing decoy; logs every server refusal	public edge
Telemetry corpus	Deterministic input to every classification run	local store
Orchestrator	Probes ports, fans out prompts, collects votes, aggregates	control host
26B anchor	Primary classifier	owned workstation on the tailnet
3B workers A/B	Secondary voters (decommissioned)	rented VPSes on the WireGuard overlay
Benchmark suite	Per-voter accuracy and distinct-class count against a labeled set	on demand, before a voter is trusted

2 Network Topology and the Security Invariant

The orchestrator fanned deterministic honeypot telemetry out to small LLMs across three network planes:

- **WireGuard overlay** — BGP-routed IPv6 mesh (`fd53::/16`); nodes talk privately.
- **Tailnet** — isolated subnet (`100.64.0.0/10`); hosts the primary 26B model on a dedicated workstation behind NAT.
- **Public internet** — no inference endpoints exposed.

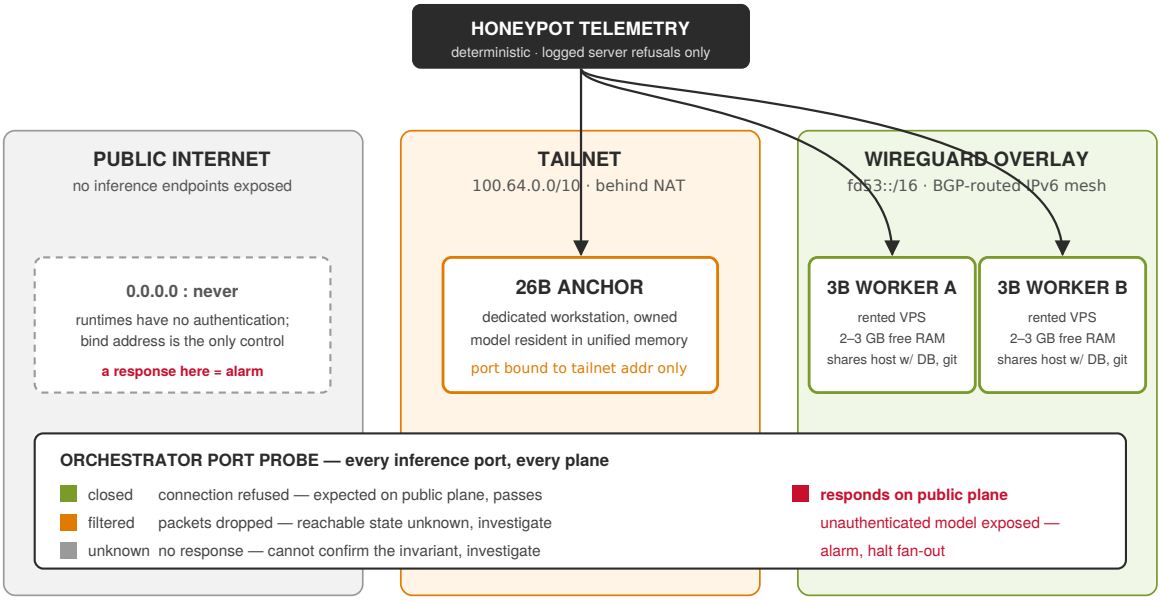

Figure 2 – Ensemble fan-out across three network planes. The security invariant is the bind address, and the orchestrator verifies it before every run.

The model runtimes had no authentication at all. Access control therefore rested on a single invariant: inference ports bind to overlay addresses, never to `0.0.0.0`.

To enforce it, the orchestrator probes every inference port on every plane before each run. A refusal on the public plane passes; a response is an alarm. The probe distinguishes three non-response states so a missing answer is never mistaken for a safe one:

State	Meaning	Action
closed	connection refused	expected on the public plane — pass
filtered	packets dropped	reachability unknown — investigate
unknown	no response	invariant unconfirmed — investigate
responds	**unauthenticated model reachable**	**alarm, halt fan-out**

3 The Hardware Bottleneck and the Illusion of Diversity

The anchor ran on owned hardware: a 26B-parameter model resident in unified memory on a dedicated workstation. The secondary voters did not. They were squeezed into residual memory (2–3 GB) on lightly provisioned rented VPSes that were also hosting production databases and git servers.

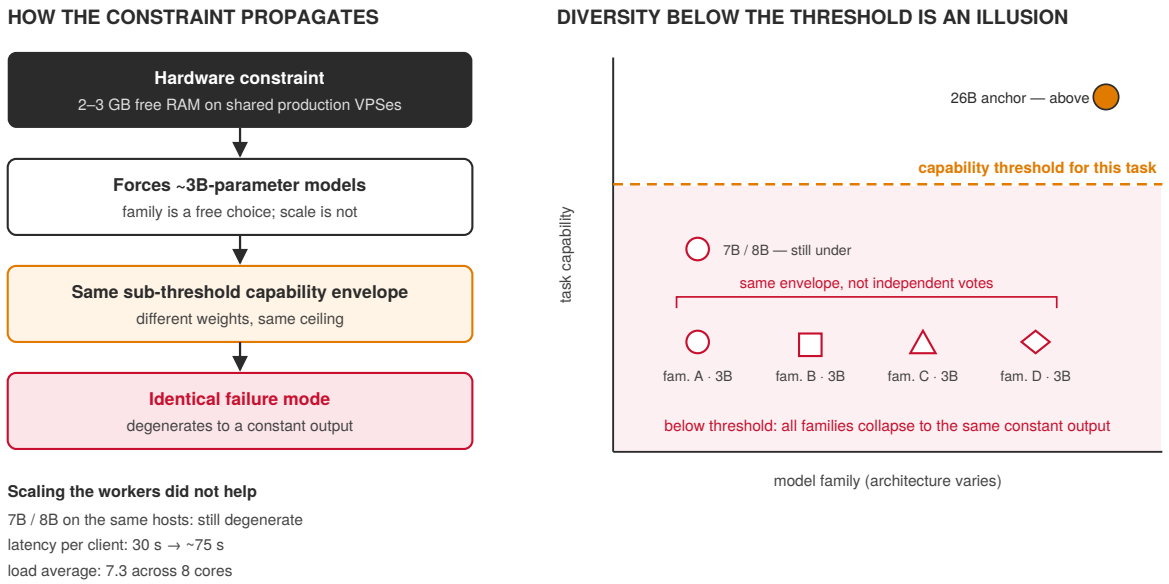

Figure 3 – Cross-family selection produced multiple samples of one under-resourced failure mode, not independent voters.

That constraint propagates in a straight line:

1. 2–3 GB of free RAM forces ~3B-parameter models.
2. Every 3B model, regardless of family, sits under the same capability ceiling for this task.
3. Models under that ceiling share the same failure mode: they degenerate to a constant output.

The trap: choosing small models across different families felt like architectural diversity. It produced multiple samples of the same under-resourced failure, not independent votes.

Scaling the workers did not help. Moving the rented voters to 7B/8B models left them degenerate, raised per-client latency from 30 s to roughly 75 s, and pushed the load average to 7.3 across 8 cores on hosts with other jobs to do.

4 Benchmarking the Voters Individually

Benchmarking each voter alone against hand-labeled telemetry exposed the breakdown immediately:

Model	Correct (of 4)	Distinct classes emitted	Status
26B Anchor	**4 / 4**	**4**	functional
3B Worker A	2 / 4	3	degraded
3B Worker B	0 / 4	**1**	**degenerate constant**

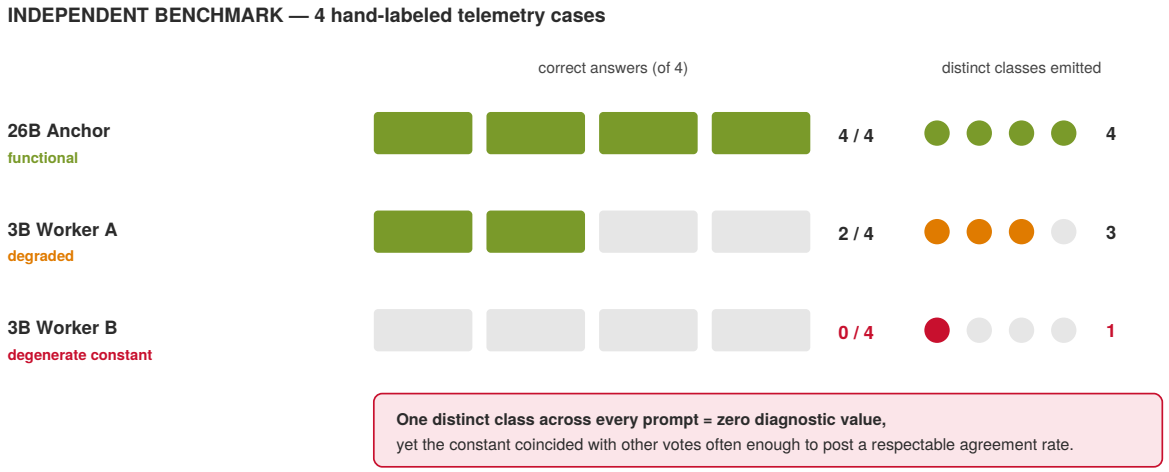

Figure 4 – Each voter benchmarked alone. Aggregate agreement had hidden that one voter never varied its answer.

4.1 Failure modes found

Constant-output collapse. Worker B — and three more small models evaluated afterward — emitted the same class for every prompt. Because that constant sometimes coincided with the other votes, its agreement rate looked respectable while it contributed zero diagnostic information.

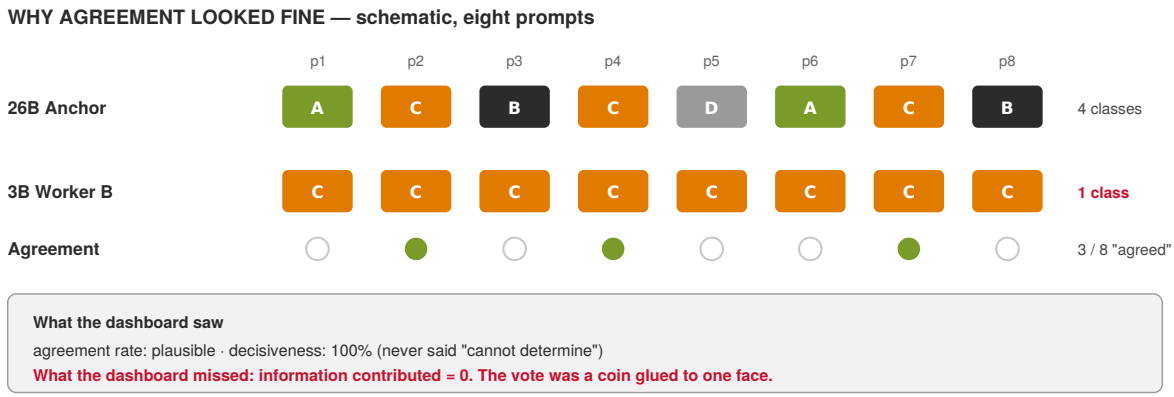

Figure 5 – A constant-output voter matches the anchor whenever the true answer happens to be its constant. Agreement is not evidence.

A decisiveness metric that rewarded the wrong thing. The core health metric penalized “cannot determine” responses. A constant-output model never hedges, so it scored perfectly on decisiveness.

Prompt token leaks. A generic enum label, `automated_tool`, repeatedly overrode the detailed docstring beneath it. Asked to classify a credential scanner, models picked `automated_tool` on token match alone. Renaming the class to `benign_monitor` — docstring unchanged — fixed it instantly.

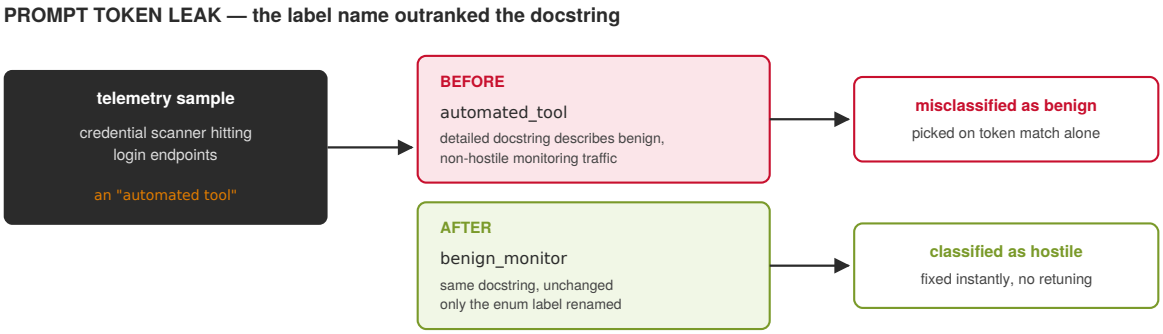

Figure 6 – Enum names are high-probability tokens; small models weight them over the fine-grained description that follows.

Unreachable schema options. The corpus contained only logged server refusals, yet the schema still offered benign-traffic categories. Struggling models used them as an escape hatch for hard cases. Removing categories that could not occur in the data eliminated the false negatives.

Temperature 0 was not deterministic. Setting sampling temperature to `0` did not produce repeatable output because the runtime did not pin the random seed.

5 Teardown and Current State

The ensemble was decommissioned and replaced with a single validated model paired with an on-demand benchmarking suite.

- **Exposure reduction.** Disabling the two secondary workers removed two unauthenticated listening endpoints from the overlay mesh.
- **Where small models still earn a place.** They are useless for voting on final verdicts, but they work as loud dissenters that flag a badly misconfigured anchor — as when a 30B code-specialized model was caught marking all hostile traffic as benign.

6 Takeaways

- **Isolate voters before aggregating.** Benchmark every model individually against a labeled set before it joins an ensemble. Aggregate agreement hides degenerate voters.
- **Track output diversity.** Count distinct class emissions per model. Any model that produces one unique output across varied runs is flagged immediately.
- **Name labels for what they mean.** Enum names are high-probability tokens; models weight them over the description that follows. The name must map strictly to intent.
- **Prune unreachable enum classes.** Remove categories that cannot occur in the target data. Impossible options become noise buckets for uncertain models.
- **Weight by measured accuracy, not assumed independence.** Never let several low-capability models outvote a high-capability anchor on the strength of theoretical architectural diversity. Shared hardware constraints strip small models of real independence.