When AI Collapses Fact and Assumption
v1.4

Blended inference is the baseline response mode of LLMs. Smooth prose is the goal. In software, that smoothness can hide the boundary between grounded analysis and inferred assumptions.

The generation process does not distinguish between a token the model can support and one it filled in. Everything comes out at the same confidence level.

I ran a small experiment on a Python caching service by asking:

> *We're seeing latency spikes on our report generation API. What should we look at?*

The baseline response correctly identified concrete areas to improve in the file: no lock or request coalescing on cache miss, a cleanup job that scans all of Redis, a stale flag that never gets checked, and synchronized TTL expiry.

In the same answer, at the same confidence level, it also said things like:

- "If this runs periodically on the same Redis used by the API, it is a strong candidate for periodic spikes."
- "If many hot reports are created around the same time—after deploy, after nightly prefetch, after business-hour traffic ramps up—they can expire around the same time too."
- "Correlate p95/p99 latency with cache hit rate for /reports/generate."

None of those lines are absurd. Some may even be useful.

The model did not know my Redis topology. It did not know my traffic shape. It did not know whether I had that telemetry. It did not verify the correlation it recommended. It moved from what it could support from the file to assumptions about the surrounding system and wrote both in the same voice.

Instead, the burden of sifting grounded analysis out of a flood of smooth prose falls on me.

That changed what review required from me. I could not just ask whether a sentence was wrong. I had to decompose the answer: what came directly from the file, what followed from reasoning over the file, and what entered because the model filled in missing context.

I then reran the same prompt and the same code using VDG protocol.

The concrete analysis stayed. But the response could no longer glide past what it did not know.

Instead of silently leaning on unknowns, the response had to put those unknowns in the Gap section:

- “No request metrics were provided, so it is unknown whether spikes are dominated by aggregate_transactions runtime, Redis latency, or concurrent duplicate work.”
- “No Redis topology was provided. It is unknown whether this cache is dedicated or shared, how many total keys live in db=0, and whether Redis CPU or memory pressure is present.”
- “No traffic-shape data was provided. It is unknown whether a small set of hot report keys dominates traffic or whether demand is evenly distributed.”
- “No client retry behavior was provided. It is unknown whether callers retry generate aggressively on slow responses, which would magnify stampedes.”

I could see what the file supported, what the model inferred, and what remained open.

That is the value of VDG protocol. Not just a consistent response shape, but a way to force the model to take on the burden of separating grounded analysis from inferred assumptions.

Get VDG Protocol | PDF