Agent Astro

THE 2026 BENCHMARK REPORT · FDA REGULATORY AI

Benchmarking Agent Astro Against Frontier LLMs on FDA Regulatory Tasks

A reproducible, like-for-like evaluation on FDA 510(k) predicate discovery and regulatory question answering, against the leading frontier models with and without web search.

In regulated work, fluency is not the same as fidelity. This report measures which AI systems are correct, which cite real records, and which know when they do not know.

Muhammad Usama, Axel Gedeon Mengara, and Adam J. Smith
Agent Astro
The first edition of an annual benchmark

Contents

01 · OVERVIEW

Overview

Regulatory teams are adopting general-purpose AI for FDA 510(k) work faster than they can vet it. These tools sound authoritative, but in regulated work that counts for nothing unless they also **get the facts right**, **cite records that actually exist**, and **say so when they do not know**. A confident wrong answer, or an invented clearance number, does more damage here than no answer at all.

Fluency is not fidelity.

0
FDA clearance numbers Agent Astro fabricated, across every task in this report

100%
Recall counts and class breakdowns Agent Astro gets exactly right, where the best web model manages 73%

90 / 90
The only system both accurate (90%) and calibrated (90%) on regulatory Q&A

67.5%
Predicates Agent Astro discovers from a description alone, a task where unaided frontier models score near zero

We benchmarked Agent Astro, a grounded FDA regulatory-intelligence platform, against the leading frontier models, Gemini 3.1 Pro, GPT-5.5 and GPT-5.6, Claude Opus 4.8, Grok-4.3, Kimi K3, and Perplexity Sonar, each evaluated with and without web search. The two tasks, predicate discovery and regulatory question answering, were built directly from the live FDA database. Every system answered identical inputs under identical, deterministic scoring, and every response was saved verbatim for audit.

The evidence separates three behaviors. **Unaided frontier models** cannot perform FDA-specific discovery and invent records freely. **Web-search models** can retrieve answers that are already public, but they are poorly calibrated and costly to run. **Agent Astro**, grounded in the full and current FDA corpus, leads decisively on genuine discovery, is uniquely accurate and calibrated on question answering, and never cites a clearance number that does not exist. One model, Grok-4.3 with web search, is a genuine peer on the question-answering set; we report that plainly. No baseline matched Agent Astro's combination of grounding, currency, calibration, zero fabrication, auditability, speed, and cost.

The sharpest separation appears not on questions but on **deliverables**, the analyst's actual unit of work. Asked to produce a recall and adverse-event assessment for a product code, Agent Astro states the correct recall count and class breakdown for every code tested; the best web-search model is correct on the count 73% of the time, and GPT-5.5 with web search only 17%. And Agent Astro returns that assessment in about half a second, where the web models take twenty seconds to several minutes. On the job, rather than the question, the gap is decisive.

> This report is written to be contested. The benchmark harness, the test sets, and the raw model responses are retained so that any figure can be reproduced or challenged. Where a baseline beats Agent Astro, the result is reported in full and explained, not omitted.

02 · THE PROBLEM

The Trust Problem in Regulatory AI

A 510(k) submission rests on substantial equivalence to a legally marketed predicate device. Identifying the right predicates, tracing a device lineage's recall and adverse-event history, and citing the correct primary records is slow, expensive, expert work, and it is precisely the work teams now hand to general-purpose AI.

The FDA cleared-device corpus is large and grows every week. At the time of this evaluation it held 174,828 devices, 165,953 recorded predicate citations, and tens of thousands of recall and adverse-event records. General-purpose language models are attractive against this backdrop because they are fast, fluent, and already in every analyst's browser.

Why fluency is not fidelity

The failure mode in this domain is specific and quiet. A model that invents a K-number, misstates a clearance date, or asserts a recall that never occurred does not merely return a wrong answer. It returns a wrong answer that reads as authoritative and is difficult to catch without checking the primary source by hand, which is the very labor the tool was meant to remove. The risk compounds when the output feeds a submission, a competitive assessment, or a post-market decision.

> The entire value of a regulatory tool is whether you can trust what it tells you and verify it quickly. A system that is usually right but occasionally fabricates with full confidence transfers the burden of verification back to the expert, and removes its own reason to exist.

The questions a buyer should therefore ask are narrow and testable: can the system find the right predicates, does it state device facts correctly, does it cite records that actually exist, and does it decline when the answer is unknown. The remainder of this report answers exactly these questions with measured evidence.

03 · METHODOLOGY

A Fair, Reproducible Comparison

Every system received identical inputs and was scored by identical, deterministic rules. The frontier models were evaluated at their strongest, with live web search, as well as alone, and the results report where they win.

Systems and conditions

System under test. Agent Astro, running on the live FDA database, using its production capabilities: the Deep Search predicate engine and the Astro Intelligence assistant. No special configuration.

Baselines. The frontier models were accessed through OpenRouter at pinned versions, each in two conditions, vanilla (the model alone) and web (the model with live web search). Identifiers: `gemini-3.1-pro-preview`, `gpt-5.5`, `gpt-5.6`, `claude-opus-4.8`, `grok-4.3`, `moonshotai/kimi-k3`, and `perplexity/sonar` with `sonar-pro` for the web condition. The first set was evaluated 2026-06-27 and 2026-06-28; GPT-5.6 and Kimi K3 were added on 2026-07-19 against the same gold sets. Kimi K3, a reasoning model, was given a larger 16,384-token output budget (the others used 4,096) so its reasoning would not exhaust the limit before it produced an answer. Each model received a neutral instruction, applied identically, to answer accurately and to state plainly when it was uncertain or when a record did not exist.

Track 1 · Predicate discovery

Given a device description with every K-number removed so the answer is not leaked, each system returns the most likely predicate 510(k) clearances, best first. Ground truth is the predicate actually cited in the device's real 510(k), an objective answer key rather than a human judgment.

Removing the memorization advantage. Frontier models are trained on public FDA data, so for historical devices they can recall a published relationship instead of discovering it. To separate discovery from recall, the primary gold set was built entirely from devices cleared between 2026-04-01 and 2026-05-29, after the models' training cutoffs. A model cannot have memorized these. Metrics: Hit@10, Recall@10, MRR, and a fabricated-citation rate, with 95% bootstrap confidence intervals and Holm-corrected McNemar tests. Sample: 200 devices.

Track 2 · Regulatory question answering

The Astro Intelligence assistant answers 100 questions drawn from the live database, in two buckets. Fact-retrieval questions, on product codes, decision dates, regulation numbers, and cited predicates of recent clearances, have deterministic ground truth. Adversarial questions probe calibration: non-existent K-numbers, fictitious device names, and false premises such as requesting the recalls of a device that has none. Scoring is fully automatic.

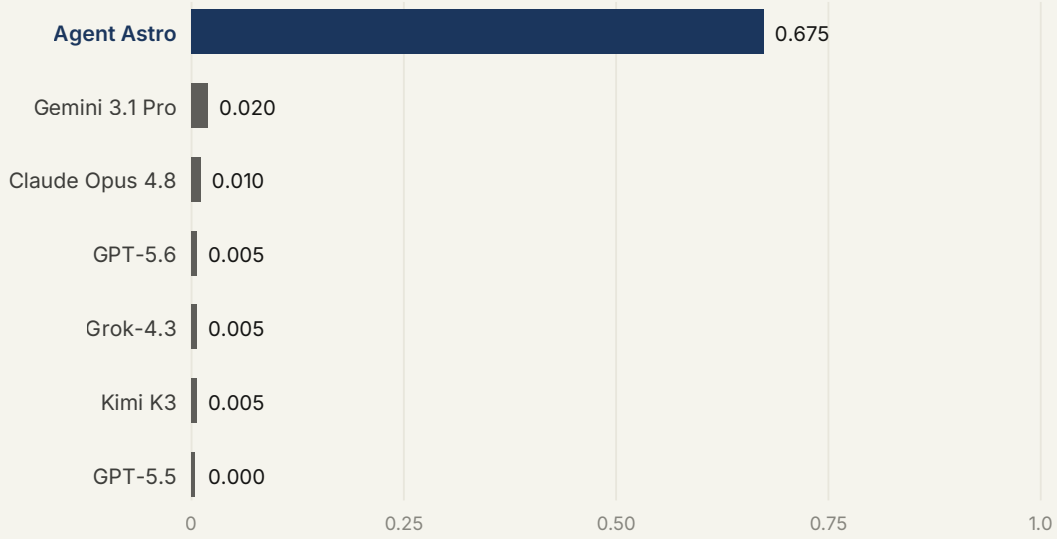
> A note on this benchmark's own integrity. Our first scorer mis-classified Agent Astro's grounded refusals ("the database does not contain a record for K706407") as fabrications, because the phrasing did not match a narrow pattern. We caught it by reading the raw answers, corrected the scorer, added regression tests, and re-scored. Inspecting raw output before trusting a surprising number is the discipline a reader should demand of any benchmark, including this one.

04 · RESULTS

Results

Predicate discovery (post-cutoff devices, n = 200)

This is the task a regulatory team actually faces: given a device you are developing, described in your own words, find the cleared devices it is substantially equivalent to. There is no answer to look up, because a device still in development has no published 510(k). We test exactly that. Every system receives a device description with all K-numbers removed and returns its most likely predicates. Agent Astro reaches 67.5% with zero fabrication. The frontier models, working from the description alone, are effectively unable to do it, and they invent 16 to 24% of the clearance numbers they offer.

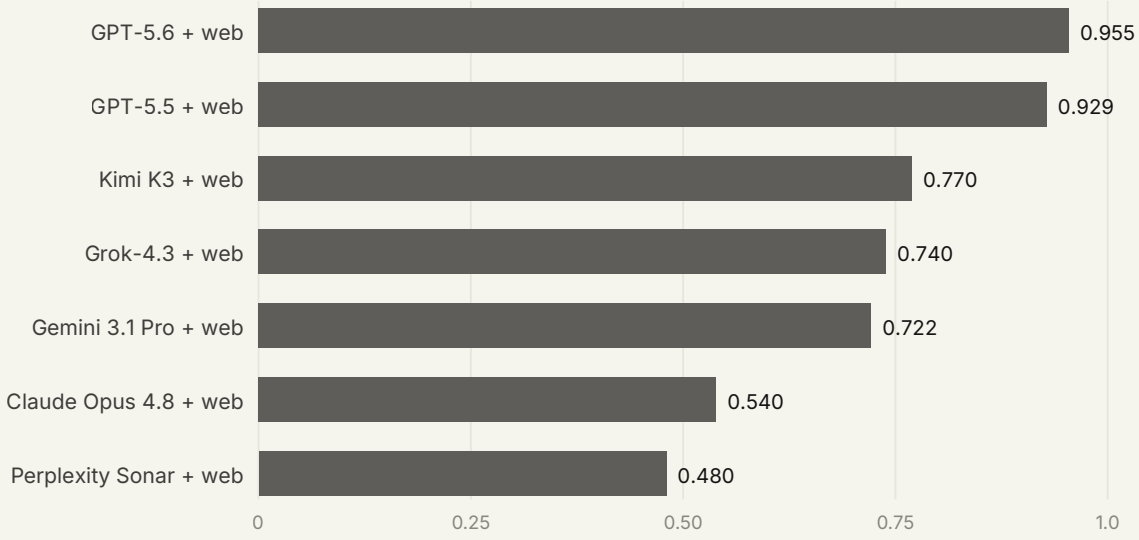


Predicate discovery from the device description alone, Hit@10 (n = 200). Agent Astro in ink-blue; the frontier models are shown without web search, the condition that reflects genuine discovery. Each of them also fabricates 16 to 24% of the K-numbers it cites; Agent Astro fabricates none.

System (no web)	Hit@10 (95% CI)	Recall@10	MRR	Fabricated
Agent Astro (discovery)	0.675 [0.610, 0.740]	0.537	0.328	0.000
Gemini 3.1 Pro	0.020 [0.005, 0.040]	0.009	0.009	0.207
Claude Opus 4.8	0.010 [0.000, 0.025]	0.007	0.003	0.237
GPT-5.6	0.005 [0.000, 0.015]	0.002	0.005	0.162
Grok-4.3	0.005 [0.000, 0.015]	0.002	0.001	0.193
Kimi K3	0.005 [0.000, 0.015]	0.002	0.005	0.156
GPT-5.5	0.000 [0.000, 0.000]	0.000	0.000	0.038

Web search is lookup, not discovery

Given live web search, the frontier models jump to between 48 and 96% on this same set. That looks decisive, and we report it in full below, but it measures a different and much easier task. Our gold devices are already cleared, so their 510(k) filings are public, and the description we feed each model is drawn from the device's own record. A web model does not reason about equivalence; it searches for that text, finds the device's own filing, and reads back the predicate the filing already lists. The scores hold even for devices cleared in the eight weeks before evaluation, which no model could have memorized, confirming the answer is fetched, not derived. It is a real capability for one narrow question, what predicate did this already-cleared device cite?, but it is not predicate discovery, and it vanishes the moment the device has not yet been filed, which is exactly when a team needs the answer. (Perplexity Sonar is search-native even in its unaided tier, so it has no genuine no-web mode and appears only here.)



The same frontier models with web search, Hit@10 (n = 200). These figures reflect reading the predicate off each device's already-published 510(k) filing, not discovering it. Agent Astro does not attempt lookup and is not shown here.

System (web search)	Hit@10 (95% CI)	Recall@10	MRR	Fabricated
GPT-5.6	0.955 [0.925, 0.980]	0.901	0.930	0.000
GPT-5.5	0.929 [0.894, 0.965]	0.887	0.901	0.001
Kimi K3	0.770 [0.710, 0.825]	0.675	0.690	0.022
Grok-4.3	0.740 [0.680, 0.800]	0.636	0.594	0.024
Gemini 3.1 Pro	0.722 [0.660, 0.784]	0.638	0.616	0.013
Claude Opus 4.8	0.540 [0.470, 0.610]	0.416	0.386	0.012
Perplexity Sonar	0.480 [0.410, 0.550]	0.363	0.383	0.066

Question answering (n = 100)

The two dimensions must be read together. A system that abstains on everything appears cautious and scores zero on facts. A system that answers confidently can score high on facts and poorly on calibration. Agent Astro and Grok-4.3 with web search sit highest in the upper-right, accurate and calibrated at once; the newer GPT-5.6 with web search climbs toward that corner but its calibration still trails, while GPT-5.5 stays overconfident. Agent Astro's advantage over GPT-5.5 with web search on calibration is statistically significant (Holm-adjusted McNemar p = 0.0003).

Regulatory Q&A, factual correctness (x) versus correct abstention on non-existent or false-premise questions (y), n = 100. The upper-right quadrant is both accurate and calibrated. Agent Astro in ink-blue.

System	Condition	Factual correctness (95% CI)	Calibration (95% CI)	Fabrication
Agent Astro	assistant	0.901 [0.84, 0.95]	0.900 [0.80, 0.97]	0.040
Grok-4.3	web	0.983 [0.96, 1.00]	0.950 [0.88, 1.00]	0.020
GPT-5.6	web	0.969 [0.93, 1.00]	0.750 [0.62, 0.88]	0.100
GPT-5.5	web	0.969 [0.93, 1.00]	0.500 [0.35, 0.65]	0.200
Gemini 3.1 Pro	web	0.793 [0.69, 0.89]	0.973 [0.92, 1.00]	0.011
Kimi K3	web	0.661 [0.54, 0.78]	0.850 [0.72, 0.95]	0.061
Perplexity Sonar	web	0.550 [0.42, 0.68]	0.775 [0.65, 0.90]	0.130
Claude Opus 4.8	web	0.050 [0.00, 0.12]	0.725 [0.57, 0.85]	0.110
Gemini 3.1 Pro / Claude Opus 4.8 / Grok-4.3	vanilla	0.000	0.80 to 0.98	0.01 to 0.08
GPT-5.6	vanilla	0.000	0.575 [0.42, 0.72]	0.170
Kimi K3	vanilla	0.008	0.675 [0.53, 0.82]	0.130
GPT-5.5	vanilla	0.000	0.475 [0.33, 0.62]	0.210

05 · DISCUSSION

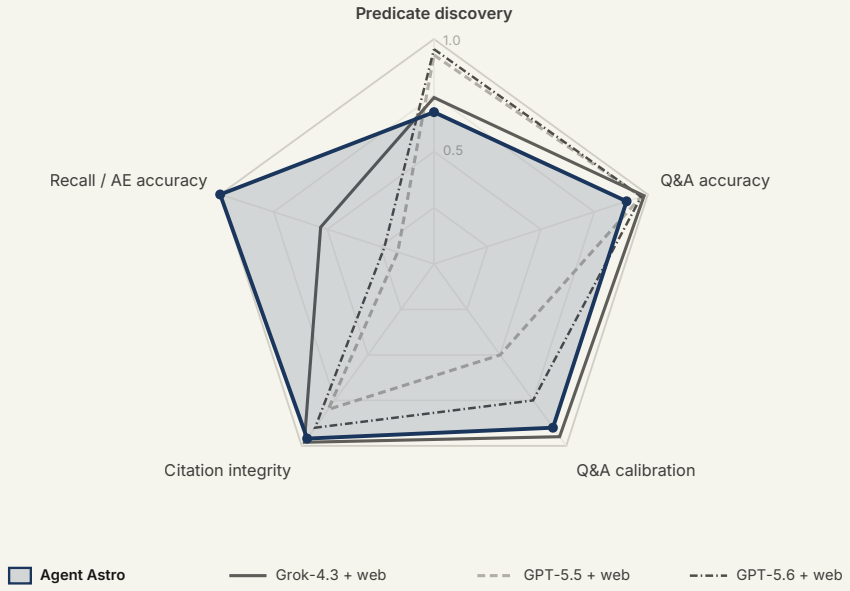
Three Behaviors, One Conclusion

Across both tasks the field resolves into three behaviors. The table below is the report in miniature; the paragraphs that follow explain each row.

Behavior	Predicate discovery	Q&A: facts	Q&A: calibration	Cost & dependency
Frontier LLM, no web	Fails (≤2%)	Fails (0%)	High, by abstaining	Cheap, but unusable for facts
Frontier LLM + web	Retrieves (up to 96%)	High (up to 98%)	Inconsistent (GPT-5.5 50%, GPT-5.6 75%)	~$0.25 to 0.58 / query; 20s to 3 min per deliverable; needs the filing public
Agent Astro	Leads (68%), zero fabrication	High (90%)	High (90%)	Low cost; grounded; auditable; ~0.5s

The combination, in one view

The chart traces each system's measured score on the five axes that matter in regulated work; a general model with web search can spike on one or two, but only Agent Astro reaches the outer edge on all five at once. Both OpenAI models with web search are strong on discovery and factual accuracy, then fall away on recall intelligence, and the newer GPT-5.6 is better calibrated than GPT-5.5 but still trails Agent Astro. Grok-4.3 with web search is a close peer on question answering but falls away on discovery and, decisively, on recall. The area each shape encloses is the argument of this report.



Measured performance on five benchmark axes, each scaled 0 to 1, higher is better. Predicate discovery is Hit@10 on post-cutoff devices (n = 200); Q&A accuracy and calibration are from the 100-question set; citation integrity is one minus the Q&A fabrication rate; recall and adverse-event accuracy is the share of product codes with the correct recall count (n = 30). The web models are credited their full published-filing lookup score on the predicate axis (see the predicate section); even so, Agent Astro is the only system that reaches the edge on every axis.

Unaided frontier models are not regulatory tools

They cannot find predicates for recent devices, they cannot state recent device facts, and when pressed they invent clearance numbers about one time in five. This is the most common way these models are used today, and it is the least safe.

Retrieval is not regulatory reasoning

GPT-5.5 with web search reaches 92.9% on predicate discovery, and GPT-5.6 95.5%, even for devices cleared in the eight weeks before evaluation. Because those devices are post-cutoff, the models cannot have learned them in training; the only explanation is that they read the device's already-published 510(k) filing from the open web. That is genuinely useful, and we credit it. But it is answer lookup, not discovery: it works only once a filing is public and indexed, it returns the filing's stated predicates rather than reasoning about equivalence, and the same models remain poorly calibrated. GPT-5.5 with web search still fabricates

on half of our non-existent-entity probes, and the newer GPT-5.6 on about one in four, still short of a grounded system.

Grounding delivers accuracy and calibration together

Agent Astro is grounded in the full, current FDA corpus. On genuine discovery it leads every unaided model by a wide margin while never fabricating a clearance number. On question answering it is the only system that is at once accurate and calibrated: it answers what it can verify and declines what it cannot, in language tied to real records. It does so at a small fraction of the per-query cost and time of the frontier web models, without depending on whether a filing has reached the open web, and with every answer traceable to a primary FDA record.

> The competitive picture is not a clean sweep. Grok-4.3 with web search is a real peer on question answering and competitive on discovery. What no baseline matches is the combination: grounded in the complete current corpus, accurate, calibrated, free of fabricated identifiers, auditable, and inexpensive, in one system that does not depend on public web indexing.

06 · THE DECISIVE TEST

From Answers to Deliverables

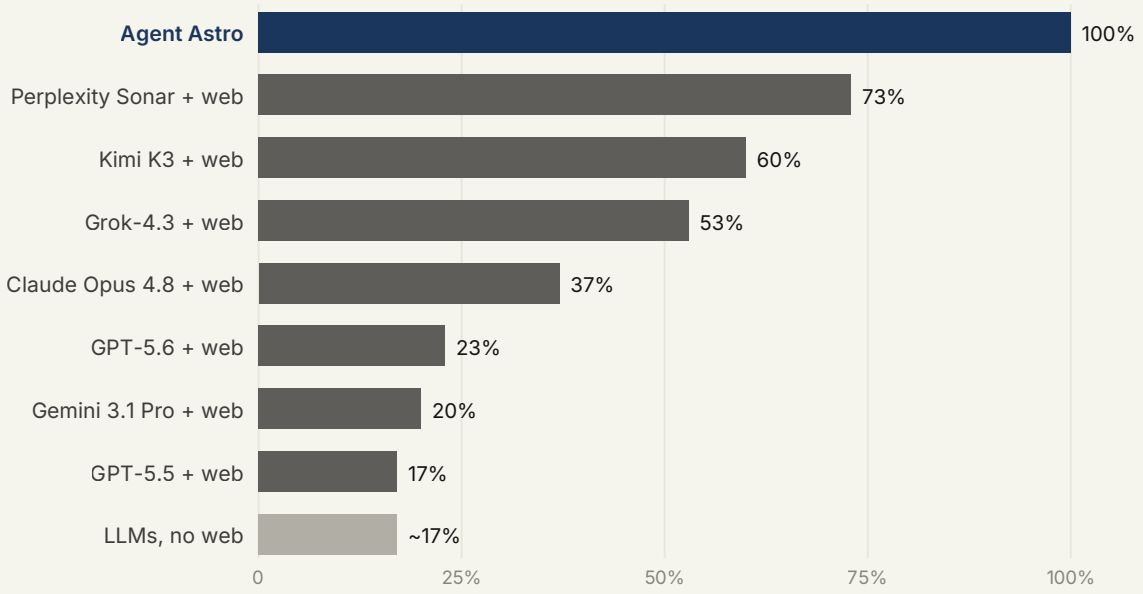
The previous chapters measured whether a model can answer a question, and conceded that on an isolated question frontier models are competitive. But regulatory work is not isolated questions. It is a set of grounded, structured workflows: recall and adverse-event assessment, predicate discovery and investigation, regulatory-pathway planning, scenario comparison. On the workflow that demands exact data, the frontier models fall off a measurable cliff; across the rest, they offer conversation where the job needs a grounded, auditable capability.

AI should be judged by the work it produces, not just the answers it gives.

Two mechanisms drive the gap. The first is a capability cliff: a general model writes prose, but it cannot enumerate the full recall set for a product code or produce a grounded, structured, corpus-wide analysis. The second is the verification tax: even a 90%-accurate model must have every claim checked, because you cannot know in advance which claim is the wrong one. So apparent accuracy saves no verification labor, while a grounded system whose claims resolve to real records does.

Measured: recall and adverse-event intelligence

We asked each system to produce a recall and adverse-event summary for a product code: the total number of recalls, the breakdown by class, and the adverse-event signal. Ground truth is the FDA recalls table. The result is the cleanest cliff in this report.



Share of product codes (n = 30) for which the system states the correct total recall count. Agent Astro in ink-blue. Kimi K3 ran with a larger token budget (see methodology). Overall deliverable-readiness: Astro 0.97 [0.95, 0.99] versus 0.30 to 0.54 for the web models.

Agent Astro states the correct recall count and the correct class breakdown for every one of the 30 product codes. The best web model, Perplexity, is right on the count 73% of the time; the reasoning model Kimi K3, given a larger token budget, reaches 60%; GPT-5.5 with web search manages only 17% and GPT-5.6 only 23%. None is complete, and none is reliable: the web models do not just under-count, they invent classes that do not exist. Agent Astro alone returns the exact count and class breakdown every time, with every figure resolving to the FDA recalls table.

WORKED EXAMPLE · RECALLS FOR PRODUCT CODE JZO

Truth: 18 recalls (Class II 14, Class III 3, Unclassified 1, no Class I).

AGENT ASTRO

"18 recalls. Class II: 14, Class III: 3, Unclassified: 1." Exact, grounded in the FDA recalls table.

FRONTIER MODELS, WITH WEB SEARCH

Grok-4.3: "at least 7 recalls, 1 Class I." Gemini 3.1 Pro: "at least 7." Perplexity Sonar: "low single digits, no Class I." Claude Opus 4.8: could not produce counts. GPT-5.5: no answer.

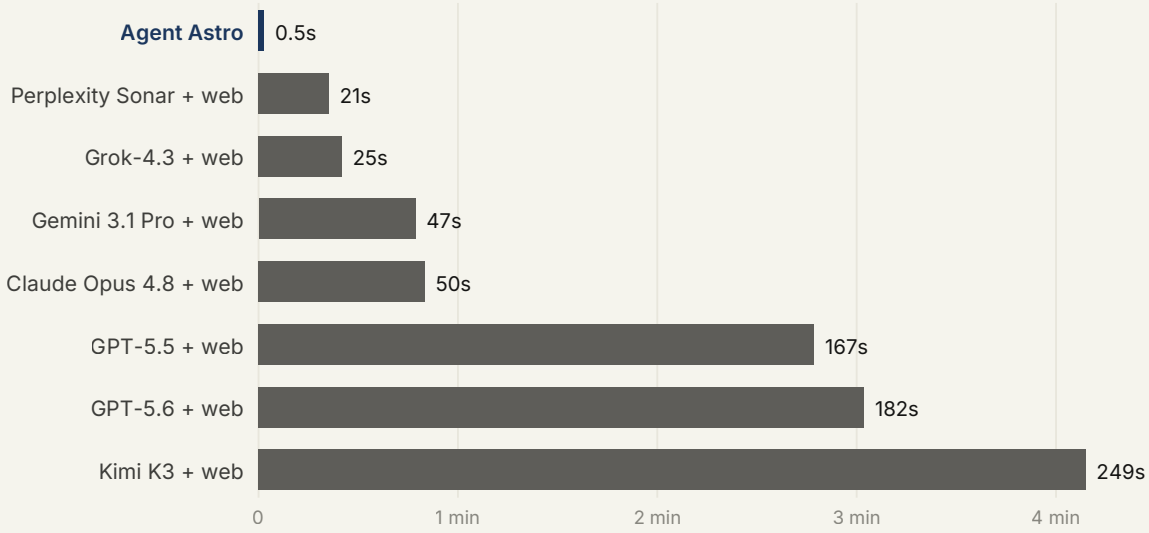
From answering to watching

The recall and adverse-event result above is a reactive one: ask, and Agent Astro returns the correct, grounded answer. The platform also runs that same grounding proactively. Its Regulatory Monitor lets a team watch the devices it relies on as predicates, the product codes it competes in, and named applicants, then checks the live FDA sources every day and raises an alert the moment a new clearance, a new recall, or an adverse-event signal lands on a watched target, each alert carrying a plain, grounded note on why it matters. This is the one signal in regulatory work that must reach you before you think to ask, because a fresh recall on a device in your submission's lineage is a fire drill, not a query.

A general model cannot offer this at all; it is a difference in kind, not degree. A chat assistant is wholly reactive: it keeps no watch list, runs on no schedule, and holds no prior state, so it can only answer the question you happen to ask on the day you ask it. And because every Agent Astro workflow, from a predicate investigation to a scenario comparison, is saved as a persistent, revisitable record tied to its source data, the work stays auditable after the fact rather than scattered through a chat history.

Half a second to a correct answer

Speed is the other half of the story, and grounding wins it twice over. Because Agent Astro reads the recall and adverse-event record straight from the FDA database, it produces this assessment in about half a second. The web-search models, which must fetch pages and reason over them, take from 20 seconds to more than three minutes, and are wrong more often than not. Accuracy and speed usually trade against each other; on the work that matters, Agent Astro leads on both at once.



Median wall-clock time to produce the recall and adverse-event assessment. Agent Astro in ink-blue (n = 30, against its production database); the frontier models with web search through OpenRouter (n = 30). Kimi K3's time reflects the larger token budget it was given. Agent Astro also returns the correct answer, which none of the others reliably does.

The regulatory workflow, end to end

Recall intelligence is one workflow among many. A regulatory professional's job spans device search, predicate discovery and investigation, pathway planning, scenario comparison, and staying alert to change. A general model can converse about any of them, but it cannot perform them, because each requires grounding in the full FDA corpus, a structured output, and citations that resolve. The map below is a capability comparison; only the recall and adverse-event row is scored in this report, the others contrast what each side actually delivers.

Workflow	Agent Astro delivers	A general model offers
Astro Intelligence (assistant)	Grounded, cited answers, accurate and calibrated on FDA facts	Fluent answers, competitive on facts but fabricates on non-existent entities
Device Search	Similar cleared devices from the full corpus, with per-candidate match reasons and confidence	A few web-findable devices, no corpus-wide search or structured reasons
Deep Search	Ranked predicates with grounded recall risk, landscape analytics, and a cited research memo	Retrieves published predicates, no grounded risk or corpus landscape
Recall & AE intelligence (measured)	Exact recall counts and class breakdown, correct on 100% of codes	Wrong counts (correct 17 to 73%), invents classes
Regulatory Pathway	A recommended pathway with a grounded evidence plan and risk view	General pathway advice, no grounded evidence plan
Predicate Investigation	Predicate-lineage traversal, a source-cited substantial-equivalence argument, and a persistent, revisitable investigation trail	Plausible prose, cannot traverse lineage or cite verifiable sources
Scenario Compare	Side-by-side scenarios with an evidence-burden matrix, a predicate-delta table, and a sensitivity view across the corpus	Ad-hoc prose comparison, not structured, grounded, or quantified
Regulatory Monitor	Proactive daily watch on predicates, product codes, and applicants; grounded alerts on new clearances, recalls, and adverse-event signals	None; a chat assistant is reactive, keeps no watch list, and cannot alert

> On an isolated question, a frontier model is a capable assistant. Across the workflows that make up regulatory work, it offers conversation where the job needs grounded data, structured output, and citations that resolve, and where the task demands exact data, as in recall intelligence, it is measurably wrong. That is the difference between a tool that discusses the work and one that does it, watches it, and leaves a record you can audit.

07 · IMPLICATIONS

What This Means for Your Team

The practical guidance follows directly from the three behaviors.

Do not rely on an unaided LLM for FDA facts. On recent devices it will not know the answer, and roughly one in five of the identifiers it offers will not exist. Its confident tone is not a signal of correctness.

Treat a web-search LLM as a retriever, not an analyst. It can surface a published filing and read off its predicates, which is useful for triage, but it does not reason about equivalence, it still fabricates on edge cases, and it depends on the filing already being public and indexed. For novel or very recent devices, and for any answer you must defend, that dependency is a liability.

Use a grounded system where the answer must be correct, complete, and auditable. The combination that matters in regulated work, accuracy with calibration, full-corpus coverage, resolvable citations, and low cost, is exactly what grounding provides and what a general model cannot guarantee.

IF YOU DO ONE THING

Run your own test. Take ten of your devices, ask each system for predicates and for the device's recall history, and check every cited identifier against the FDA database. The systems that invent records will reveal themselves immediately, and the exercise takes an afternoon.

08 · PLATFORM

Built for Regulated Enterprises

The properties that make Agent Astro accurate are the same ones that make it deployable in a regulated enterprise: every answer is grounded in a primary record, carries an audit trail, runs where your data lives, and can be delivered straight into the systems your team already uses.

Your environment, your data

For a medical-device company, the most sensitive information it holds is often exactly what a regulatory-intelligence tool touches: a device still in development, the predicates it plans to claim, a pre-submission strategy, a competitive assessment. Sending that to a public model, or to any service that moves it off your infrastructure, is a non-starter for many teams, and rightly so.

Agent Astro is built to run inside your own environment. It can be deployed in a **private VPC in your own cloud**, so the intelligence layer sits where your data already lives and your proprietary regulatory information never leaves your control. The corpus it reasons over is the public FDA record; the queries, documents, and analysis you run against it stay within your boundary. A complete, grounded knowledge base with your own data kept private is what lets a regulated organization adopt AI rather than merely evaluate it.

Security and auditability

In regulated work, security is inseparable from provenance. Every Agent Astro answer resolves to a primary FDA record, so an assessment is not only produced, it is **defensible**: a reviewer, an internal auditor, or an inspector can trace each cited clearance, recall, or predicate back to its source without redoing the analysis. Access is authenticated and each customer's workspace is isolated, so your queries and uploaded documents remain your own. The service runs on hardened cloud infrastructure with encryption in transit, and because the knowledge base is the public FDA record rather than your proprietary data, the sensitive surface area is small by design.

Scale and currency

Agent Astro is grounded in the complete FDA cleared-device corpus, **174,828 devices** and 165,953 predicate citations at the time of this report, and it does not go stale. An automated pipeline refreshes clearances, recalls, and adverse-event records every week, so the corpus stays current in a way a model's fixed training snapshot cannot. The platform is cloud-native and scales elastically, so a single analyst and an enterprise team of hundreds draw on the same complete, current knowledge base with the same response characteristics.

An intelligence layer your systems can call

Agent Astro is not only an application. Its grounded FDA intelligence is exposed through three access surfaces, so an enterprise can build it directly into internal workflows, agents, and tools, inside its own controlled environment, rather than asking analysts to leave their systems to get an answer.

REST API

Programmatic access to predicate discovery, device search, recall and adverse-event intelligence, and grounded Q&A, for pipelines, dashboards, and internal services.

SDK

A typed client that puts the same grounded capabilities into your codebase, so engineering teams integrate in hours rather than weeks.

MCP SERVER

A Model Context Protocol endpoint that exposes Agent Astro as a grounded tool to your own AI agents and copilots, so an internal assistant answers FDA questions from primary records instead of guessing.

> Whether your team works in the Agent Astro application or calls it from inside your own systems, in our cloud or in your own private environment, the guarantee is the same: grounded, auditable, and current FDA intelligence, delivered securely into the workflow where the decision is actually made, with your proprietary data under your control.

09 · LIMITATIONS

Limitations & Appropriate Use

We state these plainly because they bound what the numbers mean.

- This is a vendor-run benchmark. The harness, gold sets, and raw responses are published, and scoring is deterministic rather than a judged opinion, but independent replication is the right standard and is invited.
- Sample sizes are 200 (predicate) and 100 (Q&A). Confidence intervals are reported throughout; small-sample caution applies to per-cell rates.
- Models were run at temperature 0 with one run per item: reproducible, but not a characterization of run-to-run variance.
- Latency is wall-clock time per call. The frontier figures come from the benchmark runs (through OpenRouter, so they include routing and provider queueing); Agent Astro's are measured against its production database on a sample rather than the full set. They show the order-of-magnitude difference, not a controlled micro-benchmark.
- Q&A scoring is automatic and heuristic; it was validated against raw answers and regression-tested, but heuristics can still misclassify edge phrasings.
- Some comparisons favor the baselines by construction. The predicate web result credits live answer-retrieval that Agent Astro does not attempt; we report it rather than suppress it.

- The web condition depends on live web state and is not perfectly reproducible over time; retrieved content was saved per call.
- This report covers predicate discovery and question answering. Structured-field extraction and open-ended synthesis are ongoing work.
- Agent Astro is decision support. It does not replace a qualified regulatory professional, and its outputs should be verified against the primary FDA sources it cites.

10 · SELECTION CRITERIA

How to Evaluate Any Regulatory-Intelligence Tool

These criteria apply to any vendor, including us.

- Does it ground answers in primary FDA records and cite resolvable identifiers, or produce prose you cannot check?
- Does it ever state a clearance number, date, or recall that does not exist? Ask for its fabrication rate on adversarial inputs.
- Does it recognize when it does not know? Test it on non-existent devices and false premises.
- Does it cover the full, current corpus, or only what has reached the open web?
- Is the comparison reproducible? Can you see the methodology, the test set, and the raw outputs?
- What does each answer cost, and how fast is it, at the volume your team actually works?

Conclusion

On the tasks that decide whether a regulatory tool can be trusted, the dividing line is grounding. Unaided frontier models cannot do FDA-specific discovery and fabricate freely. Web-search models retrieve answers that are already public but are poorly calibrated and expensive to run. Agent Astro leads on genuine discovery, is uniquely accurate and calibrated on question answering, and never cites a clearance number that does not exist, at a fraction of the cost and with every answer traceable to a primary source. And across the workflows that make up regulatory work, from predicate discovery to recall and adverse-event intelligence, the advantage is one of grounded, structured capability: measured and decisive where the task demands exact data, as in recall intelligence, where Agent Astro is correct on every product code tested and the best general model manages 73%.

THE BOTTOM LINE

If your team uses a general LLM for FDA work today, assume it cannot discover predicates and will occasionally invent records with full confidence. Agent Astro is built to be accurate, calibrated, and auditable on exactly these tasks. Ask us for a live evaluation on your own devices, or for access to the benchmark harness to run it yourself.

11 · APPENDIX

Appendix & Reproducibility

A. Models and dates

Frontier baselines were accessed through OpenRouter at pinned versions and evaluated on 2026-06-27 and 2026-06-28: `google/gemini-3.1-pro-preview`, `openai/gpt-5.5`, `anthropic/claude-opus-4.8`, `x-ai/grok-4.3`, `perplexity/sonar`, and `perplexity/sonar-pro`. `openai/gpt-5.6` and `moonshotai/kimi-k3` were added on 2026-07-19 against the identical gold sets; Kimi K3, a reasoning model, ran with a 16,384-token output budget. Agent Astro ran on the live FDA cleared-device corpus, whose latest clearance at evaluation time was dated 2026-06-13.

B. Reproducibility

The benchmark harness, gold sets, run manifests (model versions, dates, seeds, and content hashes), and the per-call raw responses are retained for audit. Every figure in this report can be re-derived from those artifacts. Predicate ground truth is the FDA predicate-citation record; Q&A fact ground truth is the FDA device and recall record. Both are objective, not judged.

C. On competitor products

Other regulatory-intelligence products are not measured here because none offers a comparable, openly accessible interface for a like-for-like test on the same inputs. They are addressed separately in a capability and accessibility comparison. Frontier models are named with exact versions and dates because they can be tested reproducibly.

D. Glossary

Predicate device - a legally marketed device to which a new 510(k) device claims substantial equivalence.

Hit@10 - the share of cases where the correct predicate appears in the system's top ten candidates.

Fabrication rate - the share of answers that cite a clearance number which does not exist in the FDA database.

Calibration - correctly declining to answer about a non-existent device or false premise, rather than inventing specifics.

E. About Agent Astro

Agent Astro is a regulatory-intelligence platform for FDA medical-device work, grounded in the complete, continuously updated FDA cleared-device corpus. It supports the full arc of 510(k) practice, from predicate discovery and device search to recall and adverse-event intelligence, regulatory-pathway planning, scenario comparison, and proactive monitoring, and it exposes that intelligence through its

application as well as a REST API, an SDK, and an MCP server, so a team can work in the product or build it into their own systems. Every answer is grounded in a primary FDA record and built to be verified.

EVALUATE IT YOURSELF

Reproduce this benchmark on your own devices, or request a live walkthrough and access to the benchmark harness. The fastest way to judge a regulatory-intelligence tool is to test it on the devices your team already knows.