LifeSci Diamond Benchmark for Life Sciences

LifeSci Diamond, also written Life Sci Diamond, benchmarks AI on 55 regulatory affairs tasks. Compare model scores and explore public examples.

Arca · Updated 2026-11-04 · PDF edition
https://www.arca.inc/blog/lifesci-diamond

LifeSci Diamond (Life Sci Diamond) tests AI on hard professional tasks in life sciences. We call it “Diamond” because no evaluated system fully passes a third of its 55 tasks. Arca with GPT-6 Astra passes 29.1% (16/55). The current release focuses on regulatory affairs deliverables.

AI assistants answer regulatory questions well. The deliverables behind those questions are harder, such as classifying a change control against the guideline in force at submission, or checking that a Module 3 agrees with itself across a dozen documents. One wrong reporting category can mean a deviation or an agency query.

Related benchmarks test FDA drug label questions (FDARxBench[1]), life science research (LifeSciBench[2]) and legal agent work (Harvey's Legal Agent Benchmark[3]). We built **LifeSci Diamond** around the work regulatory teams hand to each other, such as change classifications and dossier checks. We developed the 55 tasks with practising regulatory specialists. We evaluated Arca and 18 plain model APIs on the same tasks. Arca also routes each request to the smallest model that handles it well. We cover cost and routing in Model Routing Made Our Customers' AI Work 76% Cheaper.

The tasks

Each task reads like a request to a senior colleague, with a scenario, its documents and the deliverable expected back. Answers are free text. A task-specific rubric awards points for the decisions, citations and open questions an expert expects. Penalties flag wrong claims and disqualifying decisions, such as agreeing to a CBE-30 when the change needs prior approval. Their weights affect partial credit; any triggered penalty blocks a strict pass.

Task family	What the task asks
Reporting category	Classify a post-approval change, as an EU variation type and code under the guideline in force at submission or as a US reporting category. Credit for checking each condition against the facts and asking the questions that could change the outcome.
Dossier consistency	Audit Module 2 and 3 documents for conflicts. Real conflicts sit next to differences that are supposed to exist, and false alarms cost credit like misses.

Task family	What the task asks
Missing evidence	Draft a dossier section when a required report is not in the package. Credit for naming the missing document and what it blocks, and a fail line for inventing its results.
Submissions and correspondence	Agency correspondence, label reviews and clinical hold responses.
Device vigilance	Check a device adverse-event report in FDA's registry before a PSUR cites it, and decide whether the draft describes the event correctly.

Construction

The harness, scorer and ten example tasks are public, with the same examples on Hugging Face. The other 45 tasks are outside that public release; sample scores are not comparable to the full benchmark.

EU variation tasks start from published EMA procedural records. We rewrote each as a change control for a fictional company, wrote the dossier documents it needs, and kept EMA's published classification as the answer. Other task types are checked against their own primary sources.[4]

Consistency tasks hide real conflicts, such as a shelf life that differs between the Quality Overall Summary and the stability conclusion. They also hold differences an expert should not flag, such as release and shelf-life limits. A leak check looks for copied rubric wording in the task documents.

Practising regulatory affairs specialists wrote or reviewed the original answer keys for 44 of the 55 tasks. Of those, 26 were reviewed by a specialist and 18 were written by outside contributors without a second read. The remaining 11 tasks rest on primary sources and published records; their answer keys have not had expert review. Two specialists wrote rubrics for the same tasks independently and scored the same answers almost identically.[5]

The six cases added in November use contributor answer keys and fictional records we wrote from their briefs. They have not had a second expert review. We checked their cited rules against primary sources where available and recorded gaps in each answer key.

GPT-6.1 Sol (`openai/gpt-6.1-sol`, `medium` reasoning effort) grades every criterion against the task documents. The revised rubrics check the conditions behind each decision, the evidence cited and any unsupported claims. Judges must quote the answer when applying an unsupported-claim penalty; clearly labelled assumptions and conditional draft text are allowed.

Public example

The public cross-document consistency example includes its task, documents and rubric. It shows the distinction between real dossier conflicts and expected differences.

The data behind the answers

Many regulatory answers turn on one record, such as a guidance row or an approval letter. Arca queries over 200 regulatory data sources directly instead of relying on web search.

Source	What Arca reads
openFDA	Drug and device adverse events, labels, recalls, 510(k) and PMA
Drugs@FDA	Approval letters, review documents, complete response letters
FDA enforcement	Warning letters and import refusals
DailyMed, Orange Book, Purple Book	Labels, patents and exclusivity
VAERS	Vaccine adverse event reports
EUDAMED and device regulators	EU device records, field safety notices from BfArM, PMDA and others
Trial registries	ClinicalTrials.gov, CTIS, WHO ICTRP
Literature	PubMed, Europe PMC, OpenAlex

It reads a company's own documents in Veeva Vault and SharePoint the same way, with every claim cited to its source.

Results

Updated 4 November 2026. Every number below uses the same 55 tasks and revised grading criteria.

The headline score is the **strict pass rate**. An answer passes only when every rubric line is fully met and no penalty fires. Partial credit does not count as a pass.[6]

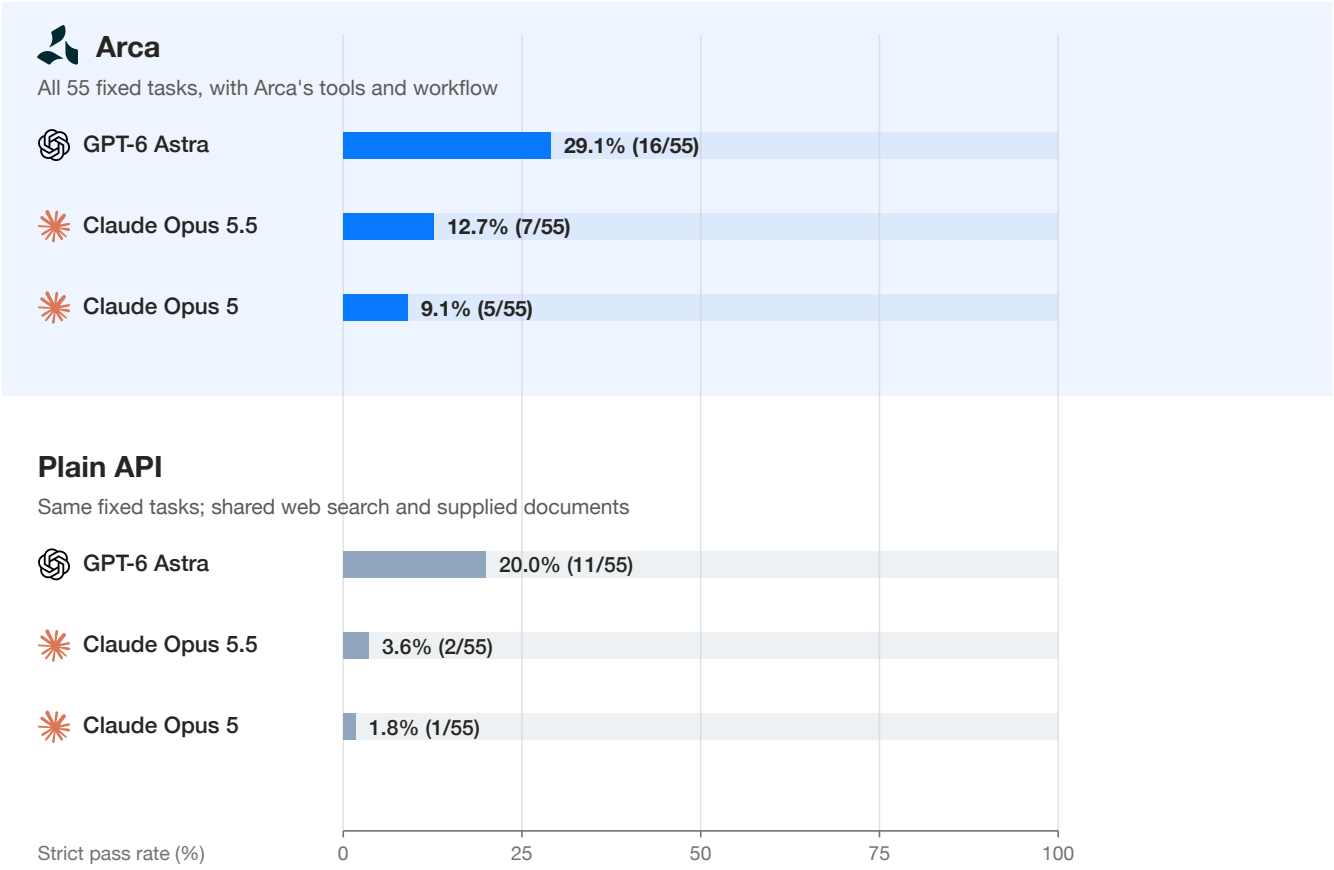
29.1% GPT-6 Astra · Arca 16/55; plain API 20.0% (11/55)

9.1% Claude Opus 5 · Arca 5/55; plain API 1.8% (1/55)

12.7% Claude Opus 5.5 · Arca 7/55; plain API 3.6% (2/55)

Model	Arca strict pass	Plain API strict pass
GPT-6 Astra	**29.1% (16/55)**	20.0% (11/55)
Claude Opus 5	**9.1% (5/55)**	1.8% (1/55)
Claude Opus 5.5	**12.7% (7/55)**	3.6% (2/55)

Every system gets the same task request and task documents in context. The plain APIs search the web with one search tool shared across vendors. Arca uses its own tools and workflow.



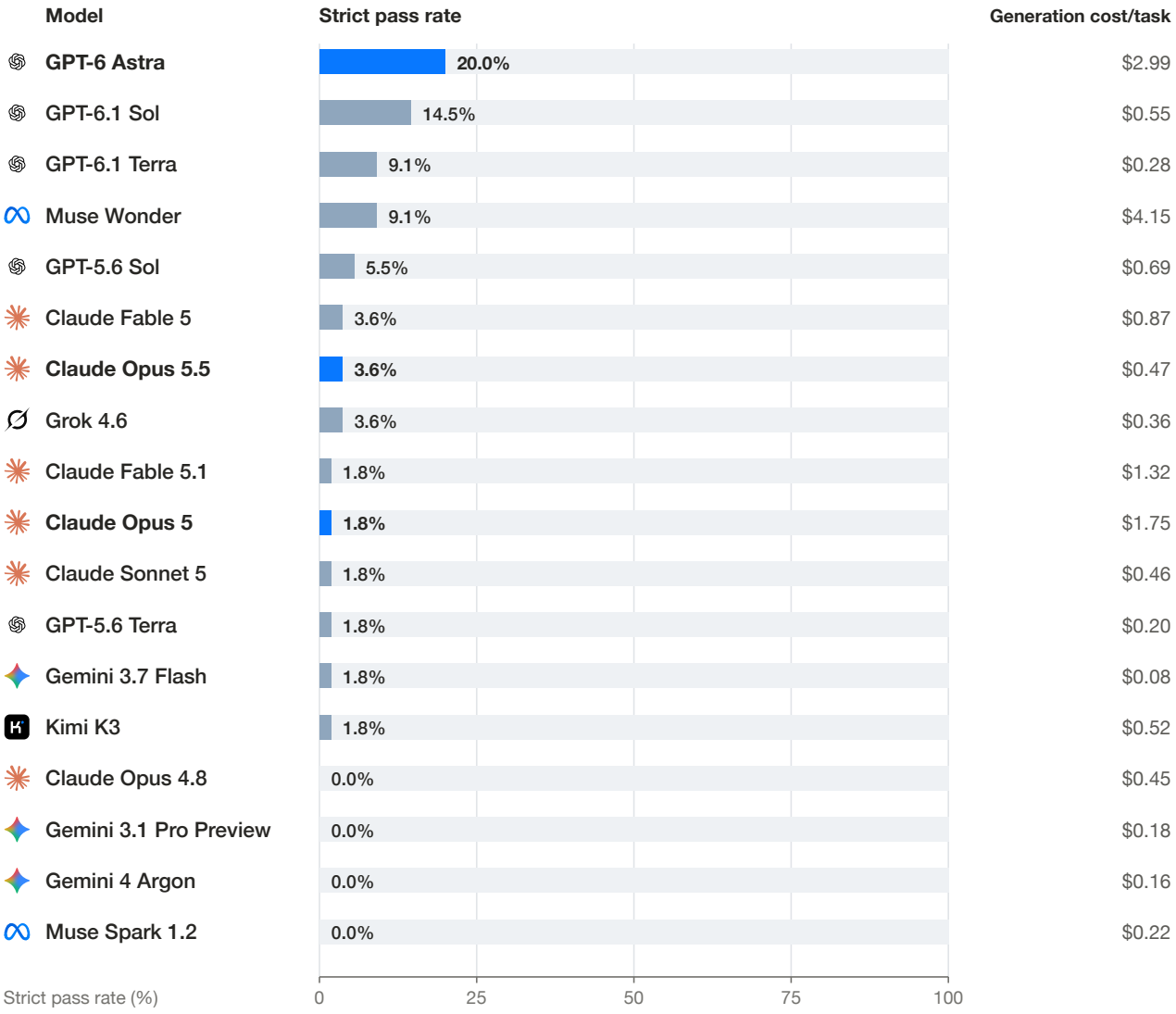
The full fixed task set. Counts and denominators appear beside each bar.

Full fixed task set

All 22 systems were graded on the same 55 tasks.

System	Full-set strict pass	Passes	Answers graded	Unavailable
Arca: GPT-6 Astra	29.1%	16	55/55	0
Arca Auto	20.0%	11	55/55	0
Arca: Claude Opus 5.5	12.7%	7	55/55	0
Arca: Claude Opus 5	9.1%	5	55/55	0
Plain API: GPT-6 Astra	20.0%	11	55/55	0
Plain API: GPT-6.1 Sol	14.5%	8	55/55	0
Plain API: GPT-6.1 Terra	9.1%	5	55/55	0
Plain API: Muse Wonder	9.1%	5	55/55	0
Plain API: GPT-5.6 Sol	5.5%	3	55/55	0
Plain API: Claude Fable 5	3.6%	2	55/55	0
Plain API: Claude Opus 5.5	3.6%	2	55/55	0

System	Full-set strict pass	Passes	Answers graded	Unavailable
Plain API: Grok 4.6	3.6%	2	55/55	0
Plain API: Claude Fable 5.1	1.8%	1	55/55	0
Plain API: Claude Opus 5	1.8%	1	55/55	0
Plain API: Claude Sonnet 5	1.8%	1	55/55	0
Plain API: GPT-5.6 Terra	1.8%	1	55/55	0
Plain API: Gemini 3.7 Flash	1.8%	1	55/55	0
Plain API: Kimi K3	1.8%	1	55/55	0
Plain API: Claude Opus 4.8	0.0%	0	55/55	0
Plain API: Gemini 3.1 Pro Preview	0.0%	0	55/55	0
Plain API: Gemini 4 Argon	0.0%	0	55/55	0
Plain API: Muse Spark 1.2	0.0%	0	55/55	0



Strict pass rate on 55 tasks. Cost per task at provider list prices, excluding grading.

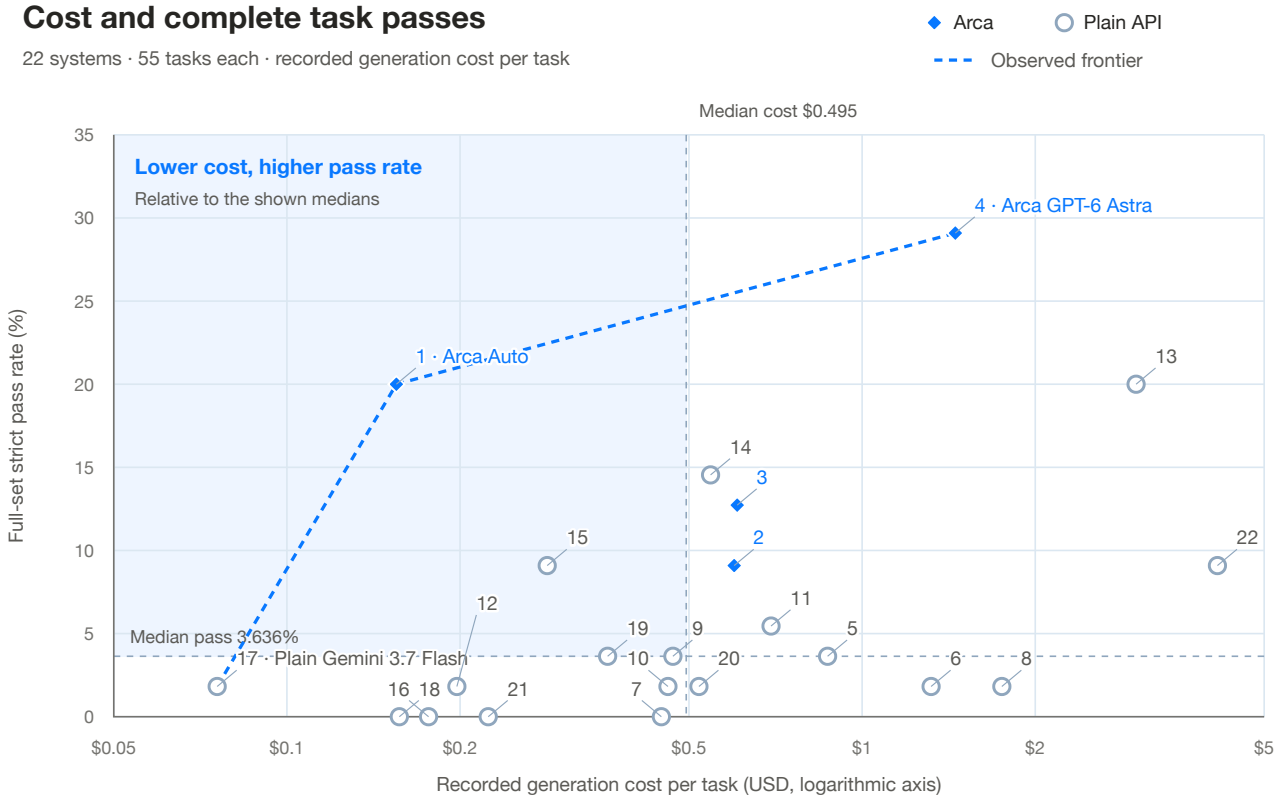
Arca's Auto setting, which picks a model for each request, passes 20.0% of tasks at $0.15 per task.

Cost and task passes

Systems toward the upper left pass more tasks at lower cost. The line marks the Pareto frontier: no other system offers at least the same pass rate at a lower cost, or a higher pass rate at the same cost. The shaded quadrant shows above-median pass rates and below-median costs.

Cost and complete task passes

22 systems · 55 tasks each · recorded generation cost per task



Shading marks below-median cost and above-median pass rate; these are visual guides.
Frontier segments connect observed systems only; no intermediate performance is implied.

System key

01 Arca Auto	$0.155 / 20.0%	12 Plain GPT-5.6 Terra	$0.197 / 1.8%
02 Arca Claude Opus 5	$0.599 / 9.1%	13 Plain GPT-6 Astra	$2.995 / 20.0%
03 Arca Claude Opus 5.5	$0.606 / 12.7%	14 Plain GPT-6.1 Sol	$0.545 / 14.5%
04 Arca GPT-6 Astra	$1.452 / 29.1%	15 Plain GPT-6.1 Terra	$0.283 / 9.1%
05 Plain Claude Fable 5	$0.871 / 3.6%	16 Plain Gemini 3.1 Pro Preview	$0.176 / 0.0%
06 Plain Claude Fable 5.1	$1.318 / 1.8%	17 Plain Gemini 3.7 Flash	$0.076 / 1.8%
07 Plain Claude Opus 4.8	$0.448 / 0.0%	18 Plain Gemini 4 Argon	$0.157 / 0.0%
08 Plain Claude Opus 5	$1.750 / 1.8%	19 Plain Grok 4.6	$0.361 / 3.6%
09 Plain Claude Opus 5.5	$0.469 / 3.6%	20 Plain Kimi K3	$0.520 / 1.8%
10 Plain Claude Sonnet 5	$0.460 / 1.8%	21 Plain Muse Spark 1.2	$0.224 / 0.0%
11 Plain GPT-5.6 Sol	$0.694 / 5.5%	22 Plain Muse Wonder	$4.147 / 9.1%

22 plotted; 0 omitted for missing cost or full-set grades. Costs exclude regrading.

Cost per task at provider list prices, excluding grading. The cost axis is logarithmic.

Models and run settings

The results combine saved answers from several batches, including 102 missing answers completed on 4 November 2026. Dates below are UTC and refer to the original batches. The API IDs are the strings sent to the gateway, not pinned model snapshots. Full run notes are in the repository.

Plain model APIs

All 18 models used Vercel AI Gateway's `/v1/chat/completions` endpoint with the same reference harness. The table gives the run date for each model's first 20 tasks. For every model, the next 29 tasks were run in batches on 2026-10-29–2026-10-30, and the six November additions on 2026-11-03. A failed Muse Wonder call on one of the 29 tasks was rerun on 2026-10-30. The timed-out Opus 5.5 call on a November addition was retried on 2026-11-03.

Model	API model ID	First 20 tasks (UTC)
Muse Wonder	`meta/muse-wonder-1.0`	2026-10-28
GPT-6 Astra	`openai/gpt-6-astra`	2026-09-25
GPT-6.1 Sol	`openai/gpt-6.1-sol`	2026-10-28
Claude Opus 5	`anthropic/claude-opus-5`	2026-09-25
Claude Fable 5.1	`anthropic/claude-fable-5.1`	2026-10-28
Claude Opus 5.5	`anthropic/claude-opus-5.5`	2026-09-25
Muse Spark 1.2	`meta/muse-spark-1.2`	2026-09-25
Grok 4.6	`spacexai/grok-4.6`	2026-09-25
GPT-6.1 Terra	`openai/gpt-6.1-terra`	2026-10-28
GPT-5.6 Sol	`openai/gpt-5.6-sol`	2026-09-25
GPT-5.6 Terra	`openai/gpt-5.6-terra`	2026-09-25
Kimi K3	`moonshotai/kimi-k3`	2026-09-25
Claude Sonnet 5	`anthropic/claude-sonnet-5`	2026-09-25
Claude Opus 4.8	`anthropic/claude-opus-4.8`	2026-09-25
Gemini 4 Argon	`google/gemini-4-argon`	2026-10-28
Gemini 3.7 Flash	`google/gemini-3.7-flash`	2026-09-25
Gemini 3.1 Pro Preview	`google/gemini-3.1-pro-preview`	2026-09-25
Claude Fable 5	`anthropic/claude-fable-5`	2026-09-25

The plain harness did not set temperature, top-p, seed, reasoning effort or a maximum output-token parameter. Those were left to gateway/provider defaults; their effective values were not recorded. Omission does not mean temperature zero or reasoning disabled.

The system prompt was “You are assisting a regulatory affairs professional. Answer in markdown. Today's date is YYYY-MM-DD.”, with the batch date substituted, followed by “Use web_search for public sources.” Task documents were included in the first user message. The only tool was `web_search`, backed by Exa, returning six results with up to 2,500 characters of text per result. The loop allowed 40 model steps per user turn, with

`tool_choice: auto` until the last step, when it requested a text answer. The gateway's caching option was `auto`; Google requests were restricted to Vertex. These are tool-assisted API results, not closed-book model results.

Arca configurations

Arca used its chat workflow, tools and system instructions through OpenRouter, rather than the plain reference harness. Each fixed-model configuration requested the API ID below. Auto is a routing setting, not a separate model: the saved answers record `openai/gpt-6.1-terra` as the selected model on all 55 tasks.

Configuration	Requested or selected model ID	First 49 tasks (UTC)	Six additions (UTC)
Arca with GPT-6 Astra	`openai/gpt-6-astra`	2026-10-30–2026-10-31	2026-11-03
Arca with Claude Opus 5	`anthropic/claude-opus-5`	2026-10-30–2026-10-31	2026-11-03
Arca with Claude Opus 5.5	`anthropic/claude-opus-5.5`	2026-10-29–2026-10-30	2026-11-03
Arca Auto	`openai/gpt-6.1-terra` (selected)	2026-10-29–2026-10-30	2026-11-03

The benchmark runner supplied no temperature, reasoning-effort or output-token overrides to Arca. Per-step provider request settings were not recorded in the saved benchmark artifacts, so these records do not establish their effective values. The runner allowed 20 minutes per workflow turn. Run dates describe answer generation, not later regrading; exact per-call timestamps were not retained for the plain APIs.

Limits and next steps

One task changes the strict pass rate by 1.8 percentage points. These results use one answer per task and system, not repeated runs.

Scores measure the written deliverable, not time saved or the outcome of a real submission. Some keys rest on judgment calls.

One added CEP-source case has a disputed Type IB answer key: a reviewer raised Type II as another reading. It still needs a second expert review.

We keep improving the benchmark. We are collecting cases from regulatory specialists' own work, where a careful colleague could still pick the wrong answer. We use plain-model runs to check difficulty. Once a task set is frozen, we keep its cases even when models answer them correctly.

Footnotes

1. Xiong et al. (2026), "FDARxBench: Benchmarking Regulatory and Clinical Reasoning on FDA Generic Drug Assessment." arXiv:2603.19539 ↩
2. OpenAI, "Introducing LifeSciBench." openai.com ↩
3. Harvey, "Introducing Harvey's Legal Agent Benchmark," 6 May 2026. harvey.ai ↩
4. US reporting-category tasks are checked line by line against 21 CFR 314.70, 601.12 and FDA's post-approval guidance. Device vigilance tasks are checked against the FDA adverse-event record. ↩

5. Two experts wrote rubrics for the same ten tasks without seeing each other's work. Scoring the same 69 answers with each rubric gives a Pearson correlation of 0.86, and the systems rank almost identically under either one (Spearman 0.96). Where they disagreed on substance, the merged rubric follows the written guidance. ↩

6. GPT-6 Astra (55 matched tasks): mean partial-credit score 87.1 for Arca versus 83.8 for the plain API; mean difference +3.3 points, paired bootstrap 95% interval -1.6 to +9.6; strict-pass difference +9.1 percentage points, paired bootstrap 95% interval +0.0 to +18.2. Claude Opus 5 (55 matched tasks): mean partial-credit score 77.5 for Arca versus 56.2 for the plain API; mean difference +21.3 points, paired bootstrap 95% interval +12.7 to +30.0; strict-pass difference +7.3 percentage points, paired bootstrap 95% interval +1.8 to +14.5. Claude Opus 5.5 (55 matched tasks): mean partial-credit score 79.4 for Arca versus 65.3 for the plain API; mean difference +14.1 points, paired bootstrap 95% interval +6.4 to +22.6; strict-pass difference +9.1 percentage points, paired bootstrap 95% interval +0.0 to +18.2. Partial-credit means are not pass rates. Intervals resample the matched tasks; each system still contributes only one answer per task. ↩