Arca

EVALUATION SNAPSHOT · NOVEMBER 2, 2026

Regulatory AI Benchmark Index

How today's leading AI models perform on regulatory affairs, legal research, biological agents, document understanding, agentic search and forecasting.

Featuring **LifeSci Diamond**, Arca's own expert-graded regulatory benchmark. All other results are publisher-reported, compiled and normalized by Arca. LifeSci Diamond updated 2 November 2026; other leaderboards as of 22 September 2026.

LifeSci Diamond: what Arca adds	2
LifeSci Diamond: base models	3
Methodology	4
Leaderboard summary	5
Latest frontier models scorecard	6
Legal research and reasoning	7
Biological agents	8
Document understanding	11
Agentic search	12
Forecasting and sources	13

INSIDE
16 benchmarks

NEW THIS EDITION
18 models on 49 LifeSci Diamond tasks

PUBLISHED BY
Arca · arca.inc

ARCA BENCHMARK · LIFESCI DIAMOND

What Arca adds to a frontier model

General benchmarks test whether a model can answer regulatory questions. LifeSci Diamond tests the deliverables regulatory teams actually produce: change classifications, dossier consistency checks and submission drafts, each graded against a rubric written by regulatory affairs experts.

49
Expert-graded tasks

12
Cases written by outside regulatory specialists

18
Plain model APIs compared

3
Models from three vendors grading every answer

Arca vs. the same models without Arca

49 tasks every system ran · score out of 100 · one run per task

SYSTEM	SCORE	$ / TASK
GPT-6 Astra		
Arca	**86.7**	$1.44
Plain API	**83.3**	$3.23
Claude Opus 5.5		
Arca	**83.5**	$0.62
Plain API	**76.9**	$0.48
Claude Opus 5		
Arca	**82.3**	$0.61
Plain API	**78.1**	$1.86

Cost is model API spend per task. Arca and the plain APIs got the same prompt, task documents and grading jury; the plain APIs searched the web with one search tool shared across vendors. The lead is an average over 49 tasks: its 95% interval includes zero for each model, so read it as a few points, not as a win on every task.

+3.4
Arca lift on GPT-6 Astra
86.7 vs 83.3 for the plain API.

+6.6
Arca lift on Claude Opus 5.5
83.5 vs 76.9 for the plain API.

2 vs 7
Fail lines hit
Answers a reviewer would reject outright, across the three models and 49 tasks each: Arca 2, plain API 7.

How to read this page. Unlike every other result in this index, Arca built this benchmark and ran these evaluations. The scores compare systems on the same tasks and are not comparable with the publisher-run benchmarks that follow.

ARCA BENCHMARK · BASE MODELS

Which model is best at regulatory work

The same benchmark with Arca taken out: 18 frontier models, each called through its vendor's API with the same prompt, the same source documents and the same web search tool, on 49 judgment tasks: the 20 original tasks (10 held-out US and dossier tasks, 10 EU variation tasks) and 29 added since (17 earlier tasks and 12 new cases written by regulatory experts).

Frontier models on regulatory tasks

18 models × 49 tasks · 882 graded answers · score out of 100 · one run per task

#	MODEL	ORIGINAL 20 TASKS	ADDED 29 TASKS	OVERALL	FAIL LINE HIT	COST / TASK
1	**GPT-6 Astra**	**86.3**	**81.3**	**83.3**	**0%**	**$3.23**
2	GPT-6.1 Sol	84.9	81.3	**82.8**	2%	$0.48
3	Muse Wonder	89.0	75.7	**81.1**	2%	$4.02
4	GPT-6.1 Terra	81.9	76.7	**78.8**	0%	$0.27
5	Claude Opus 5	80.9	76.1	**78.1**	4%	$1.86
6	Claude Fable 5.1	74.7	79.2	**77.3**	4%	$1.30
7	Claude Opus 5.5	77.3	76.6	**76.9**	10%	$0.48
8	Grok 4.6	77.2	75.1	**76.0**	2%	$0.37
9	GPT-5.6 Sol	79.0	73.1	**75.5**	4%	$0.72
10	GPT-5.6 Terra	74.0	72.8	**73.3**	2%	$0.20
11	Kimi K3	81.2	65.9	**72.2**	8%	$0.53
12	Claude Fable 5	69.1	70.5	**69.9**	12%	$0.90
13	Claude Sonnet 5	74.2	61.2	**66.5**	8%	$0.48
14	Gemini 4 Argon	71.6	59.5	**64.5**	12%	$0.18
15	Muse Spark 1.2	73.9	57.8	**64.4**	27%	$0.23
16	Claude Opus 4.8	77.3	53.9	**63.4**	12%	$0.45
17	Gemini 3.1 Pro Preview	63.8	50.2	**55.7**	8%	$0.18
18	Gemini 3.7 Flash	58.0	50.9	**53.8**	18%	$0.08

83.3

Top overall · GPT-6 Astra

No model is close to saturating it

30 pts

Spread across 18 models

Model choice matters for regulatory work

6 of 18

Caught the BLA trap below

The rest recommended the wrong category

EXAMPLE TASK · US REPORTING CATEGORY

A plasma-products company adds an independent contract lab for release and stability testing of its immune globulin. The change owner proposes a CBE-30, citing FDA's guidance for drug applications (NDAs). But the product is licensed under a BLA, so 21 CFR 601.12 and the biologics guidance apply, and under those a new release-testing lab is a **prior approval supplement**. Accepting the CBE-30 hits a fail line and scores 0.

Same tasks as page 2. This page uses one shared search tool for every vendor, the same 49 tasks as page 2 and one run per task. The 12 newest cases are private and written by outside contributors, so this index shows their scores only. Registry-lookup tasks are not scored here. Qwen 3.8 Max timed out on most tasks and is left out. Claude Fable 5.1, Gemini 4 Argon, Muse Wonder, GPT-6.1 Sol and GPT-6.1 Terra were added on 28 October 2026 with the same prompt and jury. Muse Wonder's first run of one EU task failed on an API error; its rerun is the score shown. Gemini 4 Carbon has no public API yet, so Gemini 4 Argon stands in; on one EU task Argon spent its search budget and returned no memo, and that scores 0. Grading jury: Claude Sonnet 5, GPT-5.6 Sol and Gemini 3.1 Pro, all three also scored here.

ARCA BENCHMARK · METHODOLOGY

How LifeSci Diamond is built

Five task families

49 tasks

TASK FAMILY	WHAT THE ANSWER MUST DO
Reporting category	Classify a post-approval change: EU variation type and code under the guideline in force at submission, or the US category (PAS, CBE-30, CBE-0, annual report).
Dossier consistency	Audit Module 2 and 3 documents for conflicts. Real conflicts are planted next to differences that are supposed to exist; false alarms lose credit as well as misses.
Missing evidence	Draft a dossier section when a required report is absent. Credit for naming what is missing and what it blocks; fail lines for inventing results.
Submissions & correspondence	Everyday deliverables: agency correspondence, label reviews, 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.

Ground truth

EU variation tasks start from published EMA procedural records, rewritten as change controls for a fictional company; the published classification is the answer. US reporting-category tasks are checked line by line against 21 CFR 314.70, 601.12 and FDA post-approval guidance.

Planted conflicts

Consistency tasks hide real conflicts (a 36-month shelf life in the QOS against 24 months in the stability conclusion) beside differences an expert should not flag. A leak check confirms no document contains the answer.

Rubric authors

Three external regulatory affairs professionals and one contractor engaged by Arca. Each rubric awards points for decisions, citations and open questions, and a fail line zeroes any answer that makes a disqualifying call.

Model jury

Models from three vendors grade every criterion; the full jury steps in when a verdict is partial or a fail line might fire. Judges see the source documents, so a fact from the library is never mistaken for an invented one.

0.86

Pearson correlation, two expert rubrics

Two experts wrote rubrics for the same ten tasks; the same 69 answers were scored under each.

0.96

Spearman correlation of system ranks

The seven systems rank almost identically under either rubric.

1

Run per task

A single run can move one task by twenty points, so read small gaps as noise.

Worked example: CAR-T capacity change

US change assessment memo · 7 rubric criteria

A CAR-T site adds eight instruments to empty bays of its licensed suite, doubling concurrent lots from 8 to 16. The answer is a prior approval supplement, not the CBE-30 the change owner proposed. Both reach the right category; Arca finds the rows in FDA's post-approval guidance tables that decide it, the plain API argues from the regulation and never cites them.

CRITERION	POINTS	ARCA · OPUS 5	OPUS 5 API
States the reporting category is a prior approval supplement	23.8	Met	Met
Finds the prior approval entry for a change in production capacity	20.9	Met	Missed
Handles the change owner's citation: the duplicated process train row does not govern	17.1	Met	Missed
Flags that all sixteen instruments have never run at once	13.3	Met	Met
Limits what the engineering runs support	10.4	Met	Met
Notes there is no approved comparability protocol	9.5	Met	Met
Recognises there is no location, suite or HVAC change	5.0	Met	Met
Score · steps · cost · time	**100**	**100 · 9 · $1.26 · 8.3 min**	**62 · 6 searches · $1.36 · 5.9 min**

Scope. 49 scenarios in one domain, 12 of them written by outside regulatory specialists from their own work, with EU classifications from real EMA procedures. Scores measure the written deliverable, not time saved or submission outcomes.

LEADERBOARD SUMMARY

Top result in each benchmark

All benchmarks

Best published configuration per benchmark

BENCHMARK	TOP CONFIGURATION	SCORE	METRIC
LifeSci Diamond ARCA	**Arca · GPT-6 Astra**	**86.7**	Rubric score / 100
Legal Research Bench	Claude Opus 5 / Fable 5.1 / Muse Spark 1.3 Max	**55.29%**	Strict all-pass
LegalBench	Claude Fable 5	**88.56%**	Overall accuracy
SpatialBench	Claude Opus 5.5 · Pi	**72.75%**	Pass rate
SpatialBench-Long	Claude Opus 5 · Claude Code	**41.27%**	Pass rate
scBench	GPT-6 Astra · Pi	**63.93%**	Pass rate
scBench-Long	GPT-5.6 Sol · Codex	**34.85%**	Pass rate
EpiBench	Claude Opus 5.5 · Pi	**38.99%**	Pass rate
TxBench-PP	Claude Opus 5 · Pi / GPT-6 Sol · Pi	**63.00%**	Pass rate
VariantBench	Claude Opus 5 · Claude Code	**52.82%**	Pass rate
BioSecBench-Refusal	Grok 4.7 · Grok Build	**62.4%**	Balanced score
BioSecBench-Surveillance	Claude Opus 5.5 · Claude Code [R]	**58.53%**	Pass rate
MMLongBench-Doc	Claude 4.5 Opus	**61.9%**	Accuracy
BrowseComp	OpenAI Deep Research	**51.5%**	Accuracy
Deep Research Bench	Claude Opus 4.6 · high	**0.553**	Category-averaged score
BTF-3	FutureSearch SOTA ensemble	**0.116**	Pooled Brier/RPS ↓

Reading rule. Scores are only comparable within the same benchmark and metric. Model and harness are kept where the publisher reports them. [R] = the configuration declined more than half of its runs; on BioSecBench capability boards the publisher excludes refusals from the denominator (Claude Opus 5.5 · Claude Code refused 201 of 306 Surveillance runs).

MODEL SCORECARD

Latest frontier models

September releases on every board

Best published configuration · rank among all ranked entries

BENCHMARK	CLAUDE OPUS 5.5 Anthropic · 22 Sep	CLAUDE FABLE 5.1 Anthropic · 1 Sep	GPT-6 ASTRA OpenAI · Sep	GPT-6 SOL OpenAI · 22 Sep	GPT-6 LUNA OpenAI · 22 Sep	GEMINI 3.8 FLASH Google · 2 Sep	GROK 4.7 xAI · 21 Sep	MUSE SPARK 1.3 Meta · 2 Sep
Legal Research Bench Strict all-pass	50.48% #4 of 68	55.29% **#1 of 68**	39.42% #23 of 68	28.85% #38 of 68	30.29% #35 of 68	38.94% #24 of 68	47.12% #10 of 68	55.29% **#1 of 68**
LegalBench Accuracy	—	88.51% #2 of 147	—	—	—	86.99% #6 of 147	84.39% #31 of 147	—
SpatialBench Pass rate	72.75% **#1 of 26**	—	67.54% #7 of 26	67.54% #7 of 26	59.71% #19 of 26	59.42% #20 of 26	69.57% #2 of 26	—
SpatialBench-Long Pass rate	19.05% #24 of 26	—	39.68% #2 of 26	28.57% #11 of 26	23.81% #16 of 26	23.81% #16 of 26	23.81% #16 of 26	—
scBench Pass rate	56.41% #13 of 26	—	63.93% **#1 of 26**	61.54% #3 of 26	53.85% #17 of 26	34.02% #26 of 26	58.46% #7 of 26	—
scBench-Long Pass rate	7.58% R #24 of 26	—	25.76% #4 of 26	24.24% #6 of 26	10.61% #20 of 26	24.24% #6 of 26	9.09% #22 of 26	—
EpiBench Pass rate	38.99% **#1 of 26**	—	28.30% #3 of 26	22.01% #11 of 26	19.18% #22 of 26	21.38% #14 of 26	20.75% #17 of 26	—
TxBench-PP Pass rate	30.67% R #24 of 25	—	59.33% #6 of 25	63.00% **#1 of 25**	44.33% #18 of 25	54.67% #12 of 25	54.67% #12 of 25	—
VariantBench Pass rate	50.56% #2 of 26	—	46.61% #5 of 26	38.70% #14 of 26	27.68% #23 of 26	31.36% #19 of 26	45.48% #6 of 26	—
BioSecBench-Refusal Balanced	—	—	40.8% #6 of 9	—	—	47.3% #3 of 9	62.4% **#1 of 9**	—
BioSecBench-Surveillance Pass rate	58.53% R **#1 of 26**	—	48.52% #4 of 26	47.74% #7 of 26	41.46% #18 of 26	44.11% #14 of 26	48.04% #5 of 26	—
BTF-3 Pooled ↓	—	—	0.135 #4 of 9	—	—	—	—	0.141 #6 of 9

Best harness for biology. Ties share a rank. The Muse Spark 1.3 legal entry is the Max variant (the base model scores 40.87%). — = not yet evaluated. R = declined more than half of its runs. **Blue** = #1 on that board.

ARCA READ-OUT

Claude Opus 5.5 is strong but gated in biology.
It leads SpatialBench (72.75%) and EpiBench (38.99%), but declined 57–62% of TxBench-PP runs and 98–99% of TxBench-Oligo-Discovery runs, so its therapeutics and scBench-Long scores mostly reflect refusals. Its #1 Surveillance score covers only the runs it answered.

GPT-6 Astra is the most consistent biology agent.
#1 on scBench and on the benchmarks.bio overall board (51.3%, Pi harness) without a comparable refusal pattern. GPT-6 Sol ties Claude Opus 5 for #1 on TxBench-PP.

Legal research did not improve with the new releases.
GPT-6 Sol scores 28.85% strict all-pass (#38), against 48.08% for GPT-5.6 Sol, and Claude Opus 5.5 (50.48%) trails Claude Opus 5 (55.29%).

LEGAL BENCHMARKS

Legal research and reasoning

Legal Research Bench

Strict all-pass · 68 models · 22 Sep 2026

#	MODEL	SCORE
1	**Muse Spark 1.3 Max**	**55.288%**
1	Claude Opus 5	55.288%
1	Claude Fable 5.1	55.288%
4	Claude Opus 5.5	50.481%
5	Claude Fable 5	49.519%
6	GLM-5.3	49.038%
7	Grok 4.6	48.077%
7	GPT-5.6 Sol	48.077%
9	Qwen 3.8 Max	47.596%
10	Grok 4.7	47.115%
10	MiMo V2.6 Pro	47.115%

LegalBench

Overall accuracy · 147 models · 22 Sep 2026

#	MODEL	SCORE
1	**Claude Fable 5**	**88.561%**
2	Claude Fable 5.1	88.511%
3	Gemini 3.1 Pro Preview	87.398%
4	Gemini 3.7 Flash	87.256%
5	Gemini 3 Pro Preview	87.025%
6	Gemini 3.8 Flash	86.993%
7	Claude Opus 5	86.974%
8	GPT-5.6 Sol	86.965%
9	Gemini 3 Flash Preview	86.858%
10	Gemini 3.6 Flash	86.705%

ADDITIONAL PUBLISHED SLICES

90.58%

Legal research leader, weighted pass

Claude Opus 5

68.2%

Top admin / regulatory all-pass

Claude Fable 5.1

80.0%

Top health all-pass

Claude Fable 5.1, GPT-5.6 Terra, GPT-5.6 Sol, Claude Opus 5.5

Outside the top 10. On Legal Research Bench: GPT-6 Astra 39.42% (#23), GPT-6 Luna 30.29% (#35), GPT-6 Sol 28.85% (#38), Gemini 3.8 Flash 38.94% (#24). Claude Opus 5.5, the GPT-6 family and Muse Spark 1.3 have not yet been run on LegalBench. Source: VALS Legal Research Bench and VALS LegalBench, both updated 22 Sep 2026.

BIOLOGICAL AGENT BENCHMARKS · 1 OF 3

Spatial and single-cell results

SpatialBench

115 evaluations · pass rate

#	MODEL	HARNESS	PASS	AVG COST	REFUSED
1	**Claude Opus 5.5**	**Pi**	**72.75%**	**$2.13**	**—**
2	Claude Opus 5.5	Claude Code	69.57%	$1.94	—
2	Grok 4.7	Grok Build	69.57%	$1.66	—
4	Claude Opus 5	Claude Code	68.41%	$0.71	—
4	Claude Opus 5	Pi	68.41%	$0.91	—

SpatialBench-Long

21 evaluations · pass rate

#	MODEL	HARNESS	PASS	AVG COST	REFUSED
1	**Claude Opus 5**	**Claude Code**	**41.27%**	**$10.91**	**—**
2	GPT-6 Astra	Codex	39.68%	$3.78	—
3	Claude Opus 5	Pi	36.51%	$7.70	—
3	GPT-5.6 Sol	Codex	36.51%	$3.17	—
3	GPT-5.6 Sol	Pi	36.51%	$3.72	—

scBench

195 evaluations · pass rate

#	MODEL	HARNESS	PASS	AVG COST	REFUSED
1	**GPT-6 Astra**	**Pi**	**63.93%**	**$1.09**	**—**
2	GPT-6 Astra	Codex	63.42%	$1.56	—
3	GPT-6 Sol	Pi	61.54%	$0.31	—
4	GPT-5.6 Sol	Codex	59.83%	$1.04	—
5	Grok 4.6	Pi	59.15%	$0.56	—

Source. benchmarks.bio V1 results (the site's default view since its September rebuild), accessed 25 Sep 2026. Avg cost is per run. Refused is the share of runs the model declined; — means none or under 5%.

BIOLOGICAL AGENT BENCHMARKS · 2 OF 3

Long-horizon, epigenomics and therapeutics

scBench-Long

22 evaluations · pass rate

#	MODEL	HARNESS	PASS	AVG COST	REFUSED
1	**GPT-5.6 Sol**	**Codex**	**34.85%**	**$3.51**	**—**
2	Claude Opus 5	Pi	30.30%	$7.89	—
3	GPT-5.6 Sol	Pi	27.27%	$4.49	—
4	Claude Opus 5	Claude Code	25.76%	$5.44	6%
4	GPT-6 Astra	Codex	25.76%	$3.25	—

EpiBench

106 evaluations · pass rate

#	MODEL	HARNESS	PASS	AVG COST	REFUSED
1	**Claude Opus 5.5**	**Pi**	**38.99%**	**$2.84**	**—**
2	Claude Opus 5.5	Claude Code	38.68%	$3.77	—
3	GPT-5.5	Pi	28.30%	$1.50	—
3	GPT-6 Astra	Pi	28.30%	$1.06	—
5	Claude Opus 4.8	Pi	25.79%	$4.12	—
5	GPT-5.6 Sol	Codex	25.79%	$2.29	—

TxBench-PP

100 evaluations · pass rate

#	MODEL	HARNESS	PASS	AVG COST	REFUSED
1	**Claude Opus 5**	**Pi**	**63.00%**	**$1.33**	**—**
1	GPT-6 Sol	Pi	63.00%	$0.26	—
3	Grok 4.6	Pi	61.67%	$0.51	—
4	GPT-5.6 Sol	Pi	61.33%	$1.22	—
5	Claude Opus 5	Claude Code	59.67%	$0.87	—

Source. benchmarks.bio V1 results (the site's default view since its September rebuild), accessed 25 Sep 2026. Avg cost is per run. Refused is the share of runs the model declined; — means none or under 5%.

BIOLOGICAL AGENT BENCHMARKS · 3 OF 3

Genomics and biosecurity

VariantBench

118 evaluations · pass rate

#	MODEL	HARNESS	PASS	AVG COST	REFUSED
1	**Claude Opus 5**	**Claude Code**	**52.82%**	**$1.07**	—
2	Claude Opus 5	Pi	50.56%	$2.43	—
2	Claude Opus 5.5	Pi	50.56%	$2.59	5%
4	Claude Opus 5.5	Claude Code	49.72%	$2.91	—
5	GPT-6 Astra	Pi	46.61%	$1.17	—

BioSecBench-Refusal

107 evaluations · balanced score · 9 configurations in V1

#	MODEL	HARNESS	BALANCED	RED-TEAM	ROUTINE
1	**Grok 4.7**	**Grok Build**	**62.4%**	**79.7%**	**53.6%**
2	Gemini 3.7 Flash	Pi	54.8%	63.6%	49.7%
3	Gemini 3.8 Flash	Pi	47.3%	77.5%	36.6%
4	Grok 4.6	Pi	46.9%	51.4%	44.0%
5	Grok 4.6	Grok Build	45.6%	33.3%	63.2%

BioSecBench-Surveillance

102 evaluations · pass rate

#	MODEL	HARNESS	PASS	AVG COST	REFUSED
1	**Claude Opus 5.5**	**Claude Code**	**58.53%**	**$3.24**	**66%**
2	Claude Opus 5.5	Pi	55.81%	$1.71	65%
3	Claude Opus 5	Pi	51.48%	$6.20	26%
4	GPT-6 Astra	Pi	48.52%	$1.82	16%
5	Grok 4.6	Grok Build	48.04%	$1.94	—
5	Grok 4.7	Grok Build	48.04%	$4.33	—

Biosecurity reading. Red-team is the refusal rate on hazardous requests, routine the compliance rate on benign ones, balanced combines them. Claude Opus 5.5, GPT-6 Sol and GPT-6 Luna have not yet been run on BioSecBench-Refusal V1. On Surveillance, refused runs are excluded from the denominator, so Claude Opus 5.5's lead rests on the roughly one-third of runs it answered. Source: benchmarks.bio V1 results, accessed 25 Sep 2026.

DOCUMENT UNDERSTANDING

Long-document benchmark results

MMLongBench-Doc leaderboard

1,082 questions · 135 PDFs · 19 models · top 10

#	MODEL	WEIGHTS	RELEASED	ACCURACY
1	**Claude 4.5 Opus**	**Closed**	**Nov 2025**	**61.9%**
2	Qwen3.5-397B-A17B	Open	Feb 2026	61.5%
3	Gemini-3 Pro	Closed	Nov 2025	60.5%
4	OriOn-Qwen-SR1	Open	Apr 2026	58.3%
5	Nemotron 3 Nano Omni 30B-A3B	Open	Apr 2026	57.6%
6	Qwen3-VL-235B-A22B-Instruct	Open	Sep 2025	57.0%
7	Qwen3-VL-235B-A22B-Thinking	Open	Sep 2025	56.2%
8	TeleMM-2.0	Closed	Jan 2026	56.1%
9	GLM-4.6V	Open	Dec 2025	54.9%
10	GPT-4.1	Closed	Apr 2025	49.7%

Detailed breakdown

The three models with per-category results published

MODEL	F1	SINGLE-PAGE	CROSS-PAGE	UNANSWERABLE	CHART	TABLE	TEXT	LAYOUT	FIGURE
TeleMM2.0	55.90%	66.05%	48.56%	46.19%	54.16%	59.12%	57.23%	58.34%	57.09%
GPT-4.1	51.42%	60.06%	49.87%	26.01%	48.47%	61.92%	54.26%	51.65%	48.80%
GPT-4o	46.24%	55.15%	41.40%	34.08%	43.15%	51.01%	47.87%	49.51%	43.66%

OTHER DOCUMENT SUITES INCLUDED IN THE REVIEW

1,071

MMLongBench-Doc-V2 questions

134 PDFs, corrected annotations. No leaderboard published.

13,331

MMLongBench examples

Visual RAG, NIAH, ICL, summarization, long-doc QA.

8

DocVQA 2026 domains

Competition still running; no final leaderboard yet.

Latest-model coverage. The public leaderboard was last updated 30 Apr 2026 and the detailed results set in January 2026, so none of the September 2026 models has a published score here. Sources: MMLongBench-Doc leaderboard (Hugging Face), official detailed results, MMLongBench-Doc-V2, MMLongBench and DocVQA 2026.

AGENTIC SEARCH

BrowseComp and Deep Research Bench

BrowseComp

1,266 questions · original published evaluation

#	SYSTEM	ACCURACY
1	**OpenAI Deep Research**	**51.5%**
2	OpenAI o1	9.9%
3	GPT-4o with browsing	1.9%
4	GPT-4.5	0.9%
5	GPT-4o	0.6%

The publisher notes that Deep Research was trained on tasks designed to improve BrowseComp performance.

Deep Research Bench

169 tasks · leaderboard updated 10 Jun 2026

#	AGENT CONFIGURATION	SCORE	COST	RUNTIME
1	**Claude Opus 4.6 · high**	**0.553**	**$0.53**	**183s**
2	Claude Sonnet 4.6 · high	0.549	$0.46	262s
3	Claude Opus 4.5 · high	0.548	$0.46	140s
4	GPT-5.5 · high	0.540	$0.36	231s
5	Claude Opus 4.5 · low	0.537	$0.43	162s
6	Claude Opus 4.6 · medium	0.532	$0.36	119s
7	Claude Opus 4.6 · low	0.514	$0.24	73s
8	Claude Sonnet 4.6 · low	0.504	$0.27	130s
9	Claude Opus 4.8 · high	0.502	$0.35	121s
10	Gemini 3 Flash · low	0.498	$0.10	96s

Latest-model coverage. The Deep Research Bench leaderboard was last updated 10 Jun 2026 and includes none of the September releases. Vendor-reported BrowseComp scores for newer models use different tools and harnesses and are not mixed into the original evaluation above.

Also reviewed. BrowseComp-Plus adds a reproducible 830-query evaluation over a fixed corpus of roughly 100,000 documents. Sources: BrowseComp, BrowseComp-Plus, DRB.

FORECASTING & SOURCES

BTF-3 results and source register

Bench to the Future 3

2,386 resolved questions · evaluated Jun–Aug 2026 · lower is better

#	AGENT	POOLED	BINARY BRIER	NUMERIC RPS
1	**FutureSearch SOTA ensemble §**	**0.116**	**0.112**	**0.123**
2	Claude Opus 5 · xhigh §	0.120	0.116	0.128
3	Claude Opus 4.8 · xhigh	0.132	0.130	0.135
4	GPT-6 Astra · low	0.135	0.137	0.130
5	GPT-5.6 Sol · medium	0.137	0.139	0.134
6	Muse Spark 1.3 · xhigh	0.141	0.136	0.150
7	Claude Sonnet 5 · xhigh	0.142	0.139	0.146
8	GLM-5.3-Flash	0.149	0.146	0.155
9	GLM-5.3	0.149	0.147	0.152

§ Claude Opus 5 has the latest training cutoff on the board (May 2026), close to the question snapshots; FutureSearch reports no resolution leakage. The SOTA ensemble includes Opus 5 runs. Claude Opus 5.5, Claude Fable 5.1 and GPT-6 Sol / Luna have not yet been evaluated.

Source register

Dynamic results may change after this snapshot

ARCA

Introducing LifeSci Diamond (arca.inc/blog/lifesci-diamond)

LEGAL

VALS Legal Research Bench ↗
VALS LegalBench ↗

BIOLOGY

benchmarks.bio ↗
benchmarks.bio V1 results ↗

DOCUMENTS

MMLongBench-Doc leaderboard ↗
MMLongBench-Doc results ↗
MMLongBench-Doc paper ↗
MMLongBench-Doc-V2
DocVQA 2026 ↗

SEARCH

BrowseComp ↗
BrowseComp-Plus ↗
Deep Research Bench ↗

FORECASTING

BTF-3 leaderboard ↗
BTF-3 dataset

FDA-SPECIFIC REFERENCE

Innolitics 1,200+ question private FDA benchmark

DATA DATE

Accessed 25 Sep 2026
VALS boards updated 22 Sep 2026

Methodology. LifeSci Diamond (pages 2–4) was built and run by Arca; its scores are Arca-executed evaluations. All other results are publisher-reported, compiled, normalized and analyzed by Arca. The cited publishers executed those runs.