TENUREBENCH: BENCHMARKING LIFELONG AGENT CAPABILITY OVER A DECADE OF EVOLVING WORKPLACES

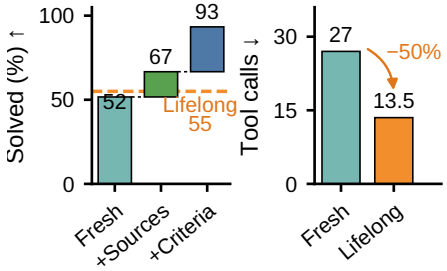
Anonymous authors
Paper under double-blind review

ABSTRACT

LLM agents are increasingly expected to work for long periods and to improve from experience. In real deployments, however, an agent works on its own without learning whether its answers were right, while the records it relies on keep changing. We introduce TenureBench to measure the lifelong capability of agents, that is, how an agent's capability changes as it keeps working. TenureBench replays 13,407 official documents from ten years of finance, medicine, and law to one persistent agent per field in publication order, keeps old versions next to new ones, and never shows the agent the correct answers. On a 300-task sample of its 920 tasks, the best of five frontier models solves 52.7%, and every agent is weakest overall at tracking what changed between two dates. Instead of improving over its lifetime, each agent does worse on later tasks. A long working history makes an agent more efficient, halving its search, but it does not make the agent more accurate. Given the source documents and the grading criteria without their answers, the same agent solves 93%. The agent has the ability to solve these tasks, and what it lacks is the professional sense, built over a career, of what a complete answer must cover. The time-driven protocol of TenureBench extends to any evolving archive to measure the lifelong capability of agents.

1 INTRODUCTION

LLM agents now work on the same kind of task for months, carrying a history of up to a million tokens from one task to the next. Many researchers expect such agents to improve from experience, as a human professional does over a career (Shinn et al., 2023; Silver & Sutton, 2025). Most of this evidence comes from short runs in which the agent is told whether it succeeded. A deployed agent is usually autonomous and rarely learns whether its answers were right. Its records also keep changing. Companies restate their numbers, drug labels get revised, and rules get rewritten, while the old versions stay on file. Existing benchmarks do not measure how an agent's capability changes under these conditions. QA benchmarks restart the agent for each task, memory benchmarks ask their questions after the history is complete, and lifelong-learning benchmarks tell the agent after each task whether it was right (Liska et al., 2022; Wu et al., 2025; 2024; Zheng et al., 2025). In a small test on such changing records, a long history makes an agent more efficient and leaves its accuracy unchanged (Figure 1).



(a) Tasks Solved (b) Tool Calls

Figure 1: Years on the job do not make the agent more accurate. The lifelong agent solves 55% of the tasks against 52% for a fresh agent, with half the tool calls (b). Pointed to the answer sources (+Sources) and given the grading criteria without their answers (+Criteria), the fresh agent reaches 93% (a).

We introduce TenureBench, a benchmark that follows one agent through a decade of real, changing records (Figure 2). It replays ten years of official records from finance, medicine, and law to one agent per field, in the order they were published. The records hold 13,407 documents about 150 companies, drugs, and regulations. The agent receives each document only on its publication date,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

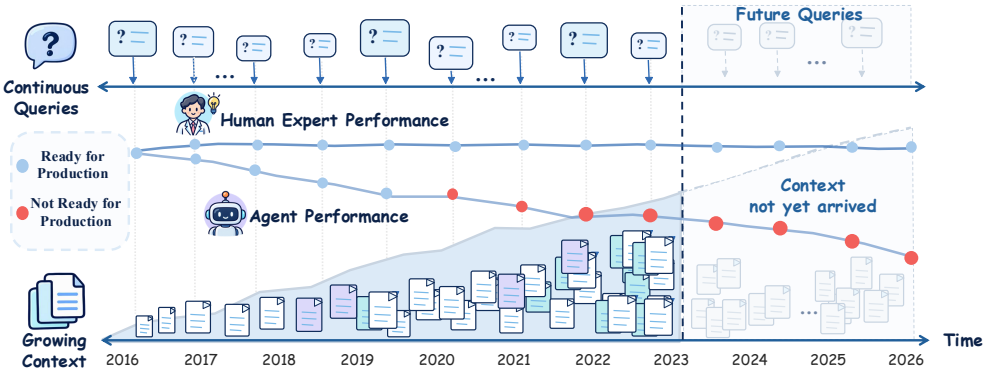


Figure 2: TenureBench follows one agent while its history grows over a decade. Documents and tasks after the current date do not yet exist for the agent (schematic, not measured).

is never reset, and never sees the correct answers. Old versions stay next to the new ones, and each of the 920 tasks needs facts from several dates. Language-model workers built the tasks under file allowlists, and domain experts checked every task. A judge grades each answer against its binary rubric criteria, 13,318 in all, and checks each cited result against the documents the agent cited. Because the protocol needs only dated documents, it applies to any field, including a company's private records.

We test agents built on five frontier models in the same agent harness. The best solves 52.7% of a 300-task sample, and every model is weakest overall at change tracking (identifying what changed between two dates). Success falls from 54.1% early in a lifetime to between 32.8% and 44.4% later. It falls most in finance, where the archive grows fastest, and it also falls as the context window fills, while the agent searches less. A controlled test on 60 tasks shows what a long history changes (Figure 1). An agent that has worked through the records up to each task makes half as many tool calls as a fresh agent on the same records, yet it solves about as many tasks, 55% against 52%. Pointed to the exact source documents, the same model solves 67%, and guided by the grading criteria as well, 93%, so the ability is there. What the agent lacks is a sense of what a complete answer covers, and a long history makes it more efficient at finding evidence without teaching it that.

Our contributions are as follows.

- We propose lifelong evaluation, a protocol that replays dated records to one agent in the order they were published and measures how its capability changes as its history grows. It needs only documents with dates, so it applies to any field (Section 3).
- We release TenureBench, 920 tasks built on ten years of official records from finance, medicine, and law. A domain expert read every task, and 13,318 binary rubric criteria grade the answers (Section 4).
- We evaluate agents built on five frontier models and find that a long history makes them more efficient without making them more accurate. Every agent is weakest overall at tracking change, and success falls as the archive grows and as the context fills (Section 5).

2 Related Work

Existing benchmarks let the world change in some way, but each misses at least one of the four properties in Table 1. News benchmarks release their evidence in time order or refresh their questions (Liska et al., 2022; Kasai et al., 2023; Vu et al., 2024; White et al., 2025; Jain et al., 2025), yet every question starts with a fresh model over a fixed snapshot, so no agent lives through the changes. Few benchmarks keep an old version next to the one that replaced it. LongMemEval and MemoryAgentBench do so for some questions in synthetic histories (Wu et al., 2025; Hu et al., 2026), and Yadav (2026) counts how often an answer relies on an outdated fact. In most benchmarks one document or one moment answers the question, while professional work often turns on what changed between dates, such as a restated figure against the original. LoCoMo, DialSim, LongMemEval, and Gaia2 each have one category of questions about time or updates (Maharana et al., 2024; Kim et al., 2024a; Froger et al., 2026), and professional benchmarks ask about a fixed set of filings, reports, contracts,

Table 1: TenureBench is the only benchmark with all four properties. *Disclosure* means evidence arrives in publication order, *Versions* that old documents stay beside new ones, *Multi-state* that a task needs several dates, and *Persistent* that one agent answers every task, never reset and never shown the answers. ○ means in part.

Benchmark	Evidence	Disclosure	Versions	Multi-state	Persistent	Horizon / scale
StreamingQA	time-stamped news	✓	–	–	–	14 yr
RealTime QA	live news and web documents	✓	–	–	–	1 wk
FreshQA	current web search	–	–	–	–	–
LoCoMo	synthetic dialogue	–	–	○	–	6 to 12 mo
DialSim	TV dialogue stream	✓	○	○	○	5 yr
LongMemEval	synthetic chat history	✓	○	○	–	500 sess., 1.5M tok.
LongMemEval-V2	web-agent trajectories	○	–	○	–	500 traj., 115M tok.
MemoryAgentBench	incremental chunked contexts	○	○	–	–	1.44M tok.
StreamBench	task-feedback stream	○	–	–	–	2,147 inputs
LifelongAgentBench	DB, OS, and KG interactions	○	–	–	–	1,396 tasks
Gaia2	simulated app states and events	○	–	○	–	$\leq$5 min (Time)
TenureBench	versioned official records	✓	✓	✓	✓	10 yr

legal cases, and exams (Islam et al., 2023; Chen et al., 2021b; Zhu et al., 2021; Jiang et al., 2026; Guha et al., 2023; Hendrycks et al., 2021; Jin et al., 2021). DialSim, StreamBench, and LifelongAgentBench keep an agent running, but each relaxes one condition. DialSim runs in a scripted world, StreamBench shows the agent after each task whether it was correct, and LifelongAgentBench gives every task a fresh environment (Kim et al., 2024a; Wu et al., 2024; Zheng et al., 2025). Agent benchmarks test whole agents in environments that do not change while the agent works (Zhou et al., 2024; Xie et al., 2024; Liu et al., 2024b; Trivedi et al., 2024; Yao et al., 2025; Xu et al., 2025a; Mialon et al., 2024; Wei et al., 2025; Froger et al., 2026; Bragg et al., 2026). Appendix F reviews the wider field, and TenureBench is the only benchmark in Table 1 that has all four properties.

3 LIFELONG EVALUATION

A professional agent serves one body of records for years, the way an analyst covers a set of companies or a regulatory specialist follows a set of rules. Documents come out one at a time, and a new document can replace an old one without deleting it. Questions can arrive on any date, and the same agent answers all of them while everything it has done stays in its history. Lifelong evaluation places an agent in this setting and measures how its success changes as that history grows (Figure 2).

We write a field as a set of dated documents $\mathcal{D} = \{(d_i, t_i)\}_{i=1}^{N}$, where t_i is the day d_i was published. On date τ the agent can see the archive $\mathcal{A}(\tau) = \{d_i \in \mathcal{D} \mid t_i \leq \tau\}$, which holds everything published so far. A task j consists of a question date τ_j, a statement q_j, gold evidence $G_j \subseteq \mathcal{A}(\tau_j)$ that supports the correct answer, and a rubric R_j of binary criteria. Tasks arrive in order of question date. At task j, the agent m reads from $\mathcal{A}(\tau_j)$ and from its own history h_{j-1}, which holds everything it did on earlier tasks. It returns an answer a_j together with the documents it cites, $C_j \subseteq \mathcal{A}(\tau_j)$, and its history grows to h_j,

$$(a_j,\, C_j,\, h_j) = m\big(q_j,\, \tau_j,\, \mathcal{A}(\tau_j),\, h_{j-1}\big). \tag{1}$$

We refer to one agent's pass through all the tasks of a field as a *lifetime*.

Archive and disclosure. A lifetime replays one field's timeline to one agent. On each publication date, the harness copies the new documents into the agent's workspace without telling it, and a replaced document stays next to its replacement, as it does in a real archive. Each document is dated by the day it became public, not the day it took effect, because an agent can only use what has been published. A document published after the current date τ does not exist yet, so the agent answers each question with what was known on that date. The harness checks every tool call against the workspace folder, and no call in our runs reached outside it. The archive therefore grows the way the real record grew, and the replay mirrors how the field changed in the world.

Tasks and rubrics. On each question date, the harness sends one message, a line such as `Current date: 2016-06-30 (Thu)` followed by `Question:` and the task. Every task needs documents from several dates, so reading the newest document is not enough, and the agent has to relate earlier states to later ones. Every criterion of the rubric is binary, as in HealthBench

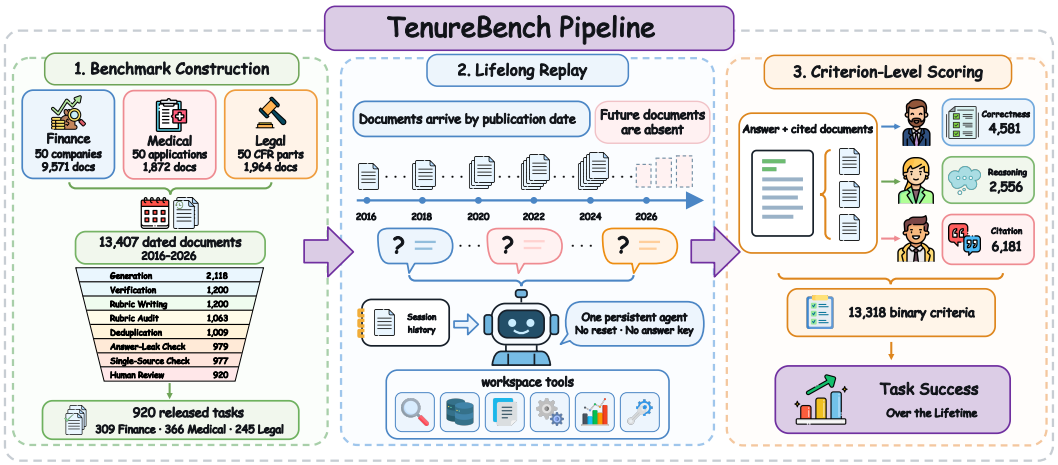


Figure 3: TenureBench builds 920 tasks from dated official records (left), replays the records and the tasks to one agent per field in publication order (middle), and scores each answer against its binary rubric criteria (right). The funnel gives the candidates left after each construction step.

(Arora et al., 2025), PaperBench (Starace et al., 2025), and ProfBench (Wang et al., 2026). *Correctness* criteria state the results the answer must report, following Mind2Web 2 (Gou et al., 2025). *Reasoning* criteria state the steps the answer must write out, following ProfBench, and they check the written answer, not the model's hidden reasoning. *Citation* criteria state the evidence each result needs, following ALCE (Gao et al., 2023), so an answer with the right value but no supporting document fails.

Agent and harness. One agent carries its history from task to task. It is never reset and never sees the correct answers, so whatever it gains over time comes from its own work. The conversation never restarts, and the harness never pastes a document into it, so to use ten years of records the agent has to search them. The agent is a small loop around the model with six standard file tools (`read`, `grep`, `find`, `ls`, `write`, `edit`), one system prompt, and one compaction rule, which summarizes the conversation when it nears the context window. It has no shell and no internet, so it cannot run code or reach anything outside its workspace. Only the model changes between systems, so any difference between two systems comes from the model.

Judge and scoring. Like other rubric benchmarks (Ye et al., 2024; Lin et al., 2025; Starace et al., 2025; Arora et al., 2025; Wang et al., 2026; Ruan et al., 2026), we use one language-model judge, GPT-5.6 Sol (OpenAI, 2026a), which grades each answer in a fresh session. The judge reads the task, the question date, the answer, the full text of every document in C_j, and the rubric. It marks each criterion k of R_j as met ($s_{jk} = 1$) or not met ($s_{jk} = 0$) with a reason, and a criterion that is only partly or loosely addressed counts as not met. Task j succeeds only when every criterion is met, because an answer with one required fact wrong cannot be used as it is. For an agent m and a group b of tasks,

$$S_m(b) = \frac{1}{|b|} \sum_{j \in b} \prod_{k \in R_j} s_{jk} \tag{2}$$

is the share of tasks in b that succeed. We refer to it as *task success* and report it across the lifetime rather than as one number, grouping tasks by their position in the lifetime, by the number of documents in the workspace, or by how full the model's context window is. The height of the curve shows how capable the agent is, and its shape shows whether that capability lasts. We also report the *rubric score*, the share of criteria met, which gives partial credit and shows, for each kind of criterion, whether an agent missed facts, cited poorly, or reasoned wrongly.

4 TENUREBENCH

TenureBench applies this protocol to three professional fields, finance (corporate filings), medical (drug labeling), and legal (federal regulation). In these fields every official document carries its

Table 2: Finance adds documents fastest, and most medical documents replace an earlier version. *Docs per entity-year* is the yearly mean over 2016 to 2026, and *replaced* the share of documents that replace an earlier version. Tokens by o200k_base (OpenAI, 2026b).

Domain	Documents	Tokens (M)	Median KB	Docs per entity-year	Replaced (%)	Tasks
Finance	9,571	139.2	5.3	17.4	21.3	309
Medical	1,872	33.6	70.9	3.4	69.9	366
Legal	1,964	92.9	40.4	3.6	24.6	245

publication date and a replaced version stays on record. An answer that uses the wrong version can report a figure the company has since restated, a dose the label no longer gives, or a rule that has since been amended. The benchmark holds 13,407 documents about 150 entities, from January 2016 to mid-2026, and 920 tasks with 13,318 criteria. Language-model workers build every task from this fixed record in three stages (Figure 3, left), and humans wrote the instructions for every step. Each worker starts in a fresh session and can read only the files its instructions list, so each step sees only the material it needs. Of 2,118 candidates, 920 pass every step, and Appendix C gives each step in full with the number of candidates it removes.

Stage I: Generation and verification. *Generation* reads the records of one group of related entities and proposes a question date τ_j, a statement q_j, gold evidence G_j, and an answer a_j^*. *Verification* rebuilds the answer from $\mathcal{A}(\tau_j)$ alone and keeps a candidate only if every value matches the original text, the answer needs documents from at least two dates and every entity the task names, and getting it right takes professional work, such as reconciling restated figures.

Stage II: Rubric writing and audit. *Rubric writing* splits a_j^* into the three kinds of criteria of R_j, reading only q_j, a_j^*, and the quoted passages of G_j, not the source documents. The *rubric audit* checks every criterion against $\mathcal{A}(\tau_j)$ and rejects a task whose criteria are wrong, redundant, incomplete, beyond the question, or tied to one exact file or wording.

Stage III: Filtering and human review. *Deduplication* removes repeated tasks. The *answer-leak check* reads only the statement, its entities, and its date, and rewrites or drops a statement that reveals answer values, directions of change, or the expected outputs. The *single-source check* repeats, by a fixed rule, verification's test that no single document or date answers the task. In *human review*, a domain expert reads every task that passed the automated steps, with its answer, evidence, and criteria, and rejects 57 of the 977. Every released task has therefore been read by an expert.

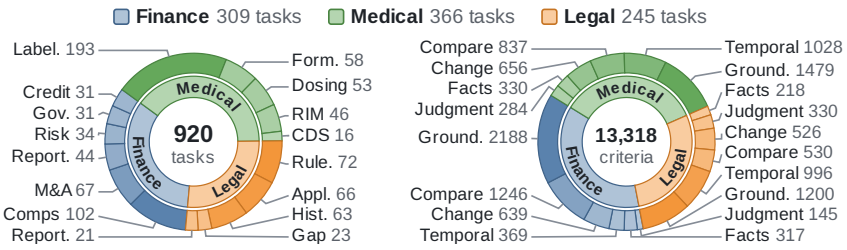
4.1 Domains

Each field follows 50 entities, selected by explicit rules for long and dense revision histories that can support tasks spanning several versions (Figure C.1a in Appendix C). A fixed parser turns each document into plain text and keeps the full text, so the agent searches the real record rather than a summary. The three records change in different ways (Table 2).

Finance. The records are the 8-K, 10-Q, and 10-K filings and their amendments on SEC EDGAR (U.S. Securities and Exchange Commission, 2025a;b). The 50 companies come from the 1,712 that filed an Item 4.02 notice between 2016 and 2026, a formal warning that earlier financial statements can no longer be relied on, and span 39 industries.

Medical. The records are the drug labels on DailyMed, every version of each (U.S. National Library of Medicine, 2026), plus the FDA's safety labeling changes (U.S. Food and Drug Administration, 2026). We picked 50 drug applications from the 241,293 labels on DailyMed, each with at least 20 label versions over nine years. A label often takes effect on a different day from the day it is published, in 282 of the 400 versions we checked, and the agent receives it on the day it is published.

Legal. The records are the final and proposed rules in the Federal Register (Office of the Federal Register, 2026b; U.S. Government Publishing Office, 2026), plus a yearly snapshot of each regulation's text from the eCFR (Office of the Federal Register, 2026a). The 50 parts come from the 333 amended parts of Titles 21 (food and drugs) and 40 (environment) of the Code of Federal Regulations.



(a) Professional Task (b) Capability Each Criterion Checks

Figure 4: The 920 tasks cover 16 professional tasks (a), and their 13,318 rubric criteria cover six capabilities, grounding the most common (b). Task abbreviations follow Section 4.2.

4.2 Statistics and Labels

Each released task records its question date, statement, one to four entities, rubric, and gold evidence with the exact passage each criterion relies on, so every grade can be traced to a source. The evidence for a task spans a median of four publication dates and 1,250 days, and at most ten years. The 920 tasks carry 13,318 criteria, a median of 9 per task, of which 4,581 are correctness, 2,556 reasoning, and 6,181 citation criteria. Every field has tasks in every year from 2016 to 2026. By the last task the archive holds 33.6M to 139.2M tokens, and even the smallest archive is 32 times the largest context window of any model we test, so every agent has to search.

We defined, with domain experts, 16 professional tasks, the jobs a practitioner in each field would name (Figure 4). Finance has credit analysis (Credit), financial reporting review (Report.), M&A and restructuring (M&A), comparable company analysis (Comps), governance and capital markets (Gov.), and risk and contingency disclosure (Risk). Medical has formulary management (Form.), clinical decision support (CDS), dosing and dispensing (Dosing), regulatory information management (RIM), and labeling change management (Label.). Legal has applicability determination (Appl.), compliance reporting and submissions (Report.), rulemaking tracking (Rule.), regulatory history research (Hist.), and regulatory gap analysis (Gap). We also defined six capabilities that a criterion can check. Fact extraction (FE) reports a value as one record states it, and grounding (GR) cites the source behind a result, which most citation criteria check. Temporal reasoning (TR) decides what held when, such as the version in force on a date. Change tracking (CT) states what changed between two states and whether a later state continues or reverses an earlier one. Comparison and calculation (CC) compares entities on aligned terms or derives a quantity, and professional judgment (PJ) applies rules to facts or reaches an evaluative conclusion. A language model applies these fixed categories, labeling each task with its professional task and each criterion with one capability from the task and its rubric alone. Appendix B defines every professional task and gives the selection rules, the released fields, the labeling model and two agreement checks, and Appendix I shows full tasks. Every field has at least 145 criteria of each capability, so the results can show which professional tasks and which capabilities an agent struggles with.

5 Experiments

5.1 Experimental Setup

Models. We evaluate five frontier models from five providers at medium reasoning effort, namely GPT-5.6 Sol (OpenAI, 2026a), Grok 4.6 (xAI, 2026a), DeepSeek Flash (V4.1; DeepSeek-AI et al., 2026), Claude Sonnet 5 (Anthropic, 2026), and GLM-5.3 (Z.ai, 2026). GLM-5.3 offers only low, high, and max effort, so it runs at its default, high, and its low and max settings form a reasoning-effort ablation, for seven systems in all.

Protocol. Every system runs as the agent of Section 3, which we adapt from the pi coding agent (Zechner, 2026) by replacing its shell tool with `grep`, `find`, and `ls`. Each system answers every task of a field in one continuous session, so it runs three full lifetimes, one per field. The evaluation set is a sample of 300 of the 920 tasks (102 finance, 119 medical, 79 legal), stratified by domain, question date, and the construction run that produced each task, and the whole archive is replayed around them. At official list prices with prompt caching, a task costs from $0.04 for DeepSeek Flash

Table 3: GPT-5.6 Sol completes the most tasks overall and in medical and legal, Grok 4.6 and DeepSeek Flash tie in finance, and low effort costs GLM-5.3 13.3 points. *Task success* is the share of tasks with every criterion met, *rubric score* the share of criteria met. Corr., Reas., and Cit. are the correctness, reasoning, and citation criteria. Best of the five systems in **bold**, second underlined, and †marks GLM-5.3 at its default, high. The last two rows set GLM-5.3 to low and max effort, with the change against high.

System	Task success				Rubric score			
	All	*Finance*	*Medical*	*Legal*	*All*	*Corr.*	*Reas.*	*Cit.*
GPT-5.6 Sol	**52.7**	37.3	**57.1**	**65.8**	**90.6**	**88.4**	**82.4**	**95.6**
GLM-5.3†	49.3	37.3	56.3	54.4	88.4	84.9	79.5	94.6
Grok 4.6	46.0	**40.2**	49.6	48.1	88.8	85.9	79.9	94.7
DeepSeek Flash	42.3	**40.2**	50.4	32.9	82.5	78.9	71.5	89.8
Claude Sonnet 5	22.7	21.6	26.9	17.7	76.1	69.5	62.3	86.7
GLM-5.3 low	36.0 (−13.3)	32.4	41.2	32.9	82.6	79.2	72.2	89.5
GLM-5.3 max	46.3 (−3.0)	34.3	52.9	51.9	87.6	83.5	79.2	94.1

to $2.58 for GPT-5.6 Sol, and the seven systems cost $3.0K over the 300 tasks ($17.6K without caching).

Scoring. The 300 tasks carry 4,249 criteria, and the judge of Section 3 scored all 2,100 answers, one run per system and field. Appendix A reports every result by domain and by reasoning effort.

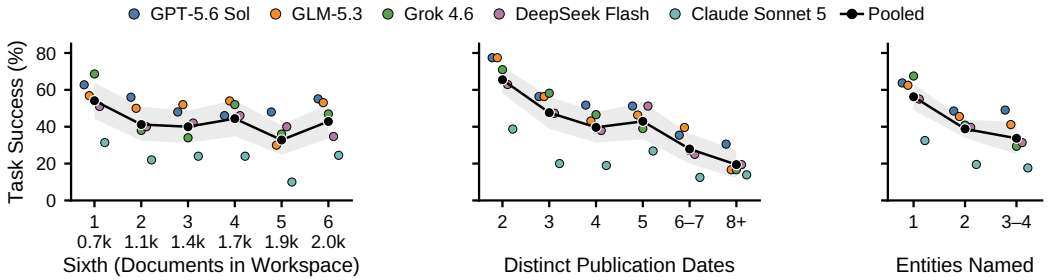
5.2 Main results

Overall. The best system solves about half the tasks. GPT-5.6 Sol solves 52.7% and Claude Sonnet 5 22.7% (Table 3). No system leads every domain. Grok 4.6 and DeepSeek Flash tie on finance at 40.2%, and GPT-5.6 Sol leads medical at 57.1% and legal at 65.8%. Comparable company analysis, which compares one metric across several companies at aligned period ends, is the hardest finance task for all five systems, at 6.2% to 18.8% (Table 4).

Capabilities. Pooled over the three fields, change tracking is every system's weakest capability, and grounding its strongest. Change-tracking criteria pass 60.3% to 80.6% of the time across the seven systems, and grounding criteria 87.7% to 95.9% (Table A.1 in Appendix A). Agents usually find the right document, and they often misread what changed in it.

Table 4: Every system is weakest at change tracking in medical, and comparable company analysis is the hardest finance task. Task success (%) by professional task (Figure 4) and pass rate by capability, the share of that capability's criteria met. Shading ranks columns, marks as in Table 3.

	Task success by professional task						Criterion pass rate by capability					
Finance	*Credit*	*Report.*	*M&A*	*Comps*	*Gov.*	*Risk*	*FE*	*GR*	*TR*	*CT*	*CC*	*PJ*
GPT-5.6 Sol	53.3	**63.6**	47.1	15.6	33.3	**40.0**	77.2	94.6	72.7	70.6	**84.7**	**87.2**
GLM-5.3†	53.3	54.5	52.9	**18.8**	50.0	20.0	84.6	95.4	**78.8**	69.7	82.0	82.1
Grok 4.6	60.0	54.5	52.9	**18.8**	**58.3**	26.7	**88.6**	93.6	73.7	**72.0**	83.3	82.1
DeepSeek Flash	**66.7**	**63.6**	**58.8**	15.6	50.0	20.0	86.2	**96.7**	75.8	67.8	78.8	71.8
Claude Sonnet 5	46.7	36.4	23.5	6.2	25.0	13.3	72.4	89.7	65.7	60.2	70.6	59.0
Medical	*Form.*	*CDS*	*Dosing*	*RIM*	*Label.*		*FE*	*GR*	*TR*	*CT*	*CC*	*PJ*
GPT-5.6 Sol	**66.7**	60.0	47.4	**50.0**	**58.2**		88.6	94.9	88.6	**82.3**	**91.4**	**90.6**
GLM-5.3†	50.0	**80.0**	**57.9**	**50.0**	56.7		83.8	93.1	86.4	74.2	90.5	**90.6**
Grok 4.6	33.3	40.0	52.6	30.0	56.7		**89.5**	**95.7**	**89.0**	77.5	84.7	89.1
DeepSeek Flash	**66.7**	60.0	36.8	**50.0**	49.3		77.1	84.4	82.4	75.1	81.5	81.2
Claude Sonnet 5	22.2	20.0	21.1	40.0	28.4		60.0	89.9	75.7	57.9	74.3	81.2
Legal	*Appl.*	*Report.*	*Rule.*	*Hist.*	*Gap*		*FE*	*GR*	*TR*	*CT*	*CC*	*PJ*
GPT-5.6 Sol	**61.1**	**57.1**	**59.3**	**84.2**	**62.5**		93.2	**99.5**	**96.7**	**90.7**	**93.8**	**96.3**
GLM-5.3†	38.9	**57.1**	55.6	63.2	**62.5**		**95.9**	96.2	87.7	87.2	85.5	93.3
Grok 4.6	38.9	**57.1**	48.1	52.6	50.0		89.2	98.1	92.4	79.7	88.1	91.1
DeepSeek Flash	38.9	14.3	40.7	31.6	12.5		85.1	86.9	69.2	69.8	73.6	90.4
Claude Sonnet 5	11.1	28.6	14.8	26.3	12.5		70.3	81.7	64.1	63.4	71.0	80.0



(a) Lifetime Sixth (b) Evidence Dates (c) Entities

Figure 5: Task success is highest in the first sixth of the lifetime and falls as a task needs more publication dates or entities. Dots are the five systems, pooled over the three fields, and the black line pools them with a 95% bootstrap interval (shaded). The row under (a) gives the median documents in the workspace.

Reasoning effort. Raising GLM-5.3's reasoning effort pays off up to its default and no further. Against high effort, low effort loses 13.3 points of task success, and max loses 3.0, a gap within noise ($p = 0.37$, Table 3). From low to high the agent searches more, with a median of 9 to 12 tool calls per task and 28.1% to 37.2% of the gold evidence opened, and it uses 2.6 times the reasoning tokens (Figure A.2 in Appendix A).

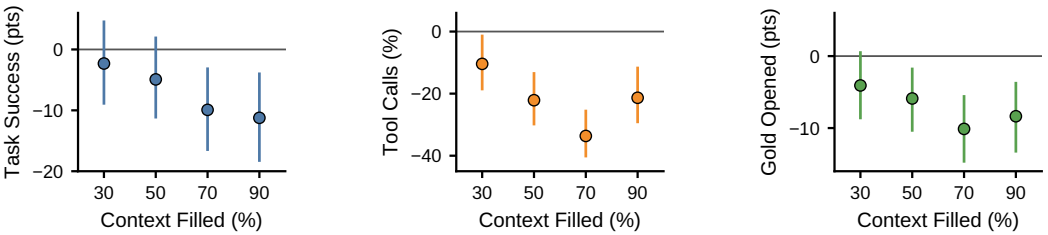
5.3 Capability over the lifetime

Success falls as the lifetime advances, most of all in finance, where the archive grows fastest. Figure 5(a) splits each lifetime into sixths, about 50 tasks each over the three fields. Pooled over the five systems, success is 54.1% in the first sixth and 32.8% to 44.4% in each later one, and all five systems score lower in the last sixth than in the first. In finance the archive nearly triples, from a median of 3,466 documents to 9,415. From the first sixth to the rest, the share of gold evidence the agents open falls from 76.1% to 47.8%, and success falls from 56.5% to 31.1%, a drop of 25.4 points (95% bootstrap interval, 5.6 to 44.7). Medical and legal, with at most 1,964 documents, move from 54.0% to 46.9% and from 51.4% to 42.2%, and both 95% intervals include zero.

Tasks that reach further into the history are harder, and agents lean more on what they read before. Pooled success falls from 65.5% at two publication dates to 19.4% at eight or more, and from 56.2% at one entity to 33.7% at three or four, and every system scores lower at the far end of both axes than at the near end (Figure 5(b) and (c)). Over the same lifetime, the share of cited documents that the agent opened only in an earlier task rises for every system, to as much as a quarter of cited documents in the last third of a lifetime.

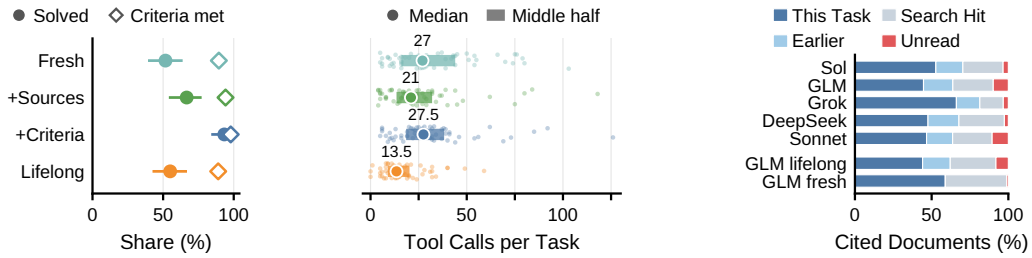
5.4 Capability within one context window

Success falls steadily as the context window fills, and the agent searches less as it does. Every system meets the same task on the same date with the same documents, but with a different amount of conversation in its context, because each conversation grows and is summarized at its own pace. Comparing systems on the same task, net of each system's average level, isolates the effect of a fuller context, with fill measured as a share of each model's window (Figure 6, and Figure A.3



(a) Task Success (b) Tool Calls (c) Gold Evidence Opened

Figure 6: Success, tool calls, and gold evidence opened all fall as the context window fills. Change against tasks that arrive with under a fifth of the window filled, for the same task across the five systems (95% bootstrap intervals over tasks). The x-axis gives the middle of each fifth.



(a) Tasks Solved (b) Tool Calls per Task (c) When Citations Were Read

Figure 7: A fresh agent solves 31 of 60 tasks, 40 with +Sources, and 56 with +Criteria, while the lifelong agent solves 33. (a) Tasks GLM-5.3 solves, with 95% Wilson intervals, and criteria met. (b) Tool calls per task: one mark per task, a bar over the middle half, and a dot at the median. (c) Cited documents by where the agent saw them, read in the task, read in an earlier task, matched by a search it ran (Search Hit), or neither opened nor matched by a search it ran (Unread). The first five rows are the five systems on the 300 tasks, and the last two are GLM-5.3 lifelong and fresh on the 60 tasks.

in Appendix A per system). Each tenth of the window filled lowers success by 1.5 points (95% bootstrap interval, 0.7 to 2.3). Compared with tasks that arrive with under a fifth of the window filled, tasks that arrive above 60% full succeed 10 to 11 points less often, and the agent makes a fifth to a third fewer tool calls and opens 8 to 10 points less of the gold evidence. Right after the conversation is summarized, the context is nearly empty again, and the agent's first task draws a median of 19 tool calls against 13 for other tasks.

5.5 What a long history contributes

GLM-5.3 at high effort answered 60 random tasks of the 300, ten from the first and ten from the last third of each domain's lifetime, under four conditions, each once (Figure 7). *Lifelong* reuses the main run, where the task sits inside the lifetime's conversation, and *Fresh* asks it in a new conversation over the archive as of its question date. *+Sources* points the agent to the sources by leaving only the documents the criteria need in its workspace, and *+Criteria* also gives the grading criteria with every answer value removed.

A long history makes the agent more efficient and leaves its accuracy unchanged. The lifelong agent solves 33 of the 60 tasks and the fresh agent 31, and the difference is not significant (exact McNemar test, $p = 0.82$). The lifelong agent makes a median of 13.5 tool calls against 27 and opens 38.8% of the gold evidence against 55.9% (Figure 7b). On the same tasks, 18.0% of the documents the lifelong agent cites were read only in an earlier task, and 8.1% are unread, neither opened nor matched by a search anywhere in its session, against 1.1% for the fresh agent (Figure 7c).

Guidance on what to cover helps more than guidance on where to look. Pointed to the sources, the agent solves 40 tasks against 31 without them ($p = 0.049$). Given the grading criteria as well, it solves 56 ($p < 0.001$ against the sources alone), including every task the lifelong agent solves. The ability to answer is already there. What the agent lacks is a sense of what a complete answer covers, the professional intuition a person builds on the job, and its long history does not build it.

6 Conclusion

TenureBench measures how an agent's capability changes as its history grows, over ten years of real records that the agent sees only as they are published. The best system solves 52.7% of the tasks, every system is weakest overall at change tracking, and a long history makes an agent more efficient and leaves its accuracy unchanged. We plan a live version that keeps adding documents from the real world or from a company's own records, so that agents are evaluated online, and a version with a human in the loop, which separates what an agent learns from feedback from what it learns from its own history. We also plan to bring in domain experts from more fields, to cover more professions and to make the rubrics more precise.

AI USE STATEMENT

In this work, we used generative AI tools to generate and check the tasks, reference answers, and rubrics of TenureBench from public records, under instructions written by the authors and human experts, and as the evaluated agents and the judge. We also used them to implement the evaluation code, support the analysis of results, and translate. Additionally, we used generative AI tools for text polishing, proofreading, adjusting figure sizes, LaTeX formatting, and organizing the supplementary material release. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The corpus consists of public official documents, namely SEC filings, Federal Register and eCFR text, and FDA drug-label records. It contains no private user data. Personal names occur only as they appear in public regulatory records. No human subjects were involved. Domain experts reviewed every task. The benchmark evaluates reading and comparing public evidence and carries no foreseeable dual-use risk beyond that of the underlying public sources.

REPRODUCIBILITY STATEMENT

The supplementary material provides our agent harness and evaluation harness with their tests, the settings of every reported system and of the judge, and a runnable subset of TenureBench with its documents, rubrics, and example answers and verdicts. The full dataset, with the document archive and all 920 tasks, will be released upon acceptance.

REFERENCES

Anthropic. System Card: Claude Sonnet 5. `https://www.anthropic.com/claude-sonnet-5-system-card`, last accessed 2026/09/26, 2026.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. HealthBench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*, 2024.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3119–3137. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.172.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*, 2022.

Jonathan Bragg et al. AstaBench: Rigorous benchmarking of AI agents with a scientific research suite. In *International Conference on Learning Representations (ICLR)*, 2026.

Center for AI Safety, Scale AI, and HLE Contributors Consortium. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649(8099):1139–1146, 2026. doi: 10.1038/s41586-025-09962-4.

Wenhu Chen, Xinyi Wang, and William Yang Wang. A dataset for answering time-sensitive questions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks)*, volume 1, 2021a.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3697–3711. Association for Computational Linguistics, 2021b. doi: 10.18653/v1/2021.emnlp-main.300.

DeepSeek-AI et al. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression. *arXiv preprint arXiv:2609.19969*, 2026.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics (TACL)*, 10:257–273, 2022. doi: 10.1162/tacl_a_00459.

Shihan Dou, Ming Zhang, Zhangyue Yin, Chenhao Huang, Yujiong Shen, Junzhe Wang, Jiayi Chen, Yuchen Ni, Junjie Ye, Cheng Zhang, Huaibing Xie, Jianglu Hu, Shaolei Wang, Weichao Wang, Yanling Xiao, Yiting Liu, Zenan Xu, Zhen Guo, Pluto Zhou, Tao Gui, Zuxuan Wu, Xipeng Qiu, Qi Zhang, Xuanjing Huang, Yu-Gang Jiang, Di Wang, and Shunyu Yao. CL-bench: A benchmark for context learning. *arXiv preprint arXiv:2602.03587*, 2026.

Romain Froger, Pierre Andrews, Matteo Bettini, Amar Budhiraja, Ricardo Cabral, Virginie Do, Emilien Garreau, Jean-Baptiste Gaya, Hugo Laurençon, Maxime Lecanu, Kunal Malkan, Dheeraj Mekala, Pierre Ménard, Gerard Bertran, Ulyana Piterbarg, Mikhail Plekhanov, Mathieu Rita, Andrey Rusakov, Vladislav Vorotilov, Mengjue Wang, Ian Yu, Amine Benhalloum, Grégoire Mialon, and Thomas Scialom. Gaia2: Benchmarking LLM agents on dynamic and asynchronous environments. In *International Conference on Learning Representations (ICLR)*, 2026.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6465–6488. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.398.

Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Boyu Gou, Zanming Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanev, Botao Yu, Bernal Jimenez Gutierrez, Yiheng Shu, Chan Hee Song, Jiaman Wu, Shijie Chen, Hanane Moussa, Tianshu Zhang, Jian Xie, Yifei Li, Tianci Xue, Zeyi Liao, Kai Zhang, Boyuan Zheng, Zhaowei Cai, Viktor Rozgic, Morteza Ziyadi, Huan Sun, and Yu Su. Mind2Web 2: Evaluating agentic search with agent-as-a-judge. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 38, 2025. doi: 10.52202/085713-5778.

Neel Guha et al. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, pp. 44123–44279, 2023. doi: 10.52202/075280-1915.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *International Conference on Learning Representations (ICLR)*, 2026.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. CUAD: An expert-annotated NLP dataset for legal contract review. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks)*, volume 1, 2021.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *Conference on Language Modeling (COLM)*, 2024.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *International Conference on Learning Representations (ICLR)*, 2026.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. FinanceBench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations (ICLR)*, 2025.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. TemporalWiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6237–6250. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.418.

Yidong Jiang, Junrong Chen, Eftychia Makri, Jialin Chen, Peiwen Li, Ali Maatouk, Leandros Tassiulas, Eliot Brenner, Bing Xiang, and Rex Ying. Fin-RATE: A real-world financial analytics and tracking evaluation benchmark for LLMs on SEC filings. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD)*, pp. 9183–9194. Association for Computing Machinery, 2026. doi: 10.1145/3770855.3817528.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world Github issues? In *International Conference on Learning Representations (ICLR)*, 2024.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. doi: 10.3390/app11146421.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah Smith, Yejin Choi, and Kentaro Inui. RealTime QA: What's the answer right now? In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, pp. 49025–49043, 2023. doi: 10.52202/075280-2130.

Jiho Kim, Woosog Chay, Hyeonji Hwang, Daeun Kyung, Hyunseung Chung, Eunbyeol Cho, Yeonsu Kwon, Yohan Jo, and Edward Choi. DialSim: A dialogue simulator for evaluating long-term multi-party dialogue understanding of conversational agents. *arXiv preprint arXiv:2406.13144*, 2024a.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations (ICLR)*, 2024b.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 29348–29363, 2021.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9459–9474, 2020.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. WildBench: Benchmarking LLMs with challenging tasks from real users in the wild. In *International Conference on Learning Representations (ICLR)*, 2025.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien de Masson d'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-McMahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. StreamingQA: A benchmark for adaptation to new knowledge over time in question answering models. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13604–13622. PMLR, 2022.

Nelson Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023 (Findings of EMNLP)*, pp. 7001–7025. Association for Computational Linguistics, 2023a. doi: 10.18653/v1/2023.findings-emnlp.467.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics (TACL)*, 12:157–173, 2024a. doi: 10.1162/tacl_a_00638.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations (ICLR)*, 2024b.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2511–2522. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.emnlp-main.153.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 13851–13870. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.747.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. In *International Conference on Learning Representations (ICLR)*, 2024.

Office of the Federal Register. eCFR :: API Documentation. `https://www.ecfr.gov/developers/documentation/api/v1`, last accessed 2026/09/26, 2026a.

Office of the Federal Register. Federal Register :: API Documentation. `https://www.federalregister.gov/developers/documentation/api/v1`, last accessed 2026/09/26, 2026b.

OpenAI. GPT-5.6 System Card. `https://deploymentsafety.openai.com/gpt-5-6`, last accessed 2026/09/26, 2026a.

OpenAI. openai/tiktoken: tiktoken is a fast BPE tokeniser for use with OpenAI's models. `https://github.com/openai/tiktoken`, last accessed 2026/09/26, 2026b.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. Proving test set contamination in black-box language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 1–22. Association for Computing Machinery, 2023. doi: 10.1145/3586183.3606763.

Jie Ruan, Inderjeet Nair, Shuyang Cao, Amy Liu, Sheza Munir, Micah Pollens-Dempsey, Yune-Ting Chiang, Lucy Kates, Nicholas David, Sihan Chen, Ruxin Yang, Yuqian Yang, Jihyun Gump, Tessa Bialek, Vivek Sankaran, Margo Schlanger, and Lu Wang. ExpertLongBench: Benchmarking language models on expert-level long-form generation tasks with structured checklists. In *International Conference on Learning Representations (ICLR)*, 2026.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023 (Findings of EMNLP)*, pp. 10776–10787. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.722.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 8634–8652, 2023. doi: 10.52202/075280-0377.

David Silver and Richard S. Sutton. Welcome to the Era of Experience. `https://storage.googleapis.com/deepmind-media/Era-of-Experience%20/The%20Era%20of%20Experience%20Paper.pdf`, last accessed 2026/09/26, 2025.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. PaperBench: Evaluating AI's ability to replicate AI research. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 56843–56873. PMLR, 2025.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 16022–16076. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.850.

U.S. Food and Drug Administration. Drug Safety-related Labeling Changes (SrLC). `https://www.accessdata.fda.gov/scripts/cder/safetylabelingchanges/`, last accessed 2026/09/26, 2026.

U.S. Government Publishing Office. GovInfo. `https://www.govinfo.gov/`, last accessed 2026/09/26, 2026.

U.S. National Library of Medicine. DailyMed - Download All Drug Labels. `https://dailymed.nlm.nih.gov/dailymed/spl-resources-all-drug-labels.cfm`, last accessed 2026/09/26, 2026.

U.S. Securities and Exchange Commission. About EDGAR. `https://www.sec.gov/submit-filings/about-edgar`, last accessed 2026/09/26, 2025a.

U.S. Securities and Exchange Commission. EDGAR Application Programming Interfaces (APIs). `https://www.sec.gov/search-filings/edgar-application-programming-interfaces`, last accessed 2026/09/26, 2025b.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics:*

ACL 2024 (Findings of ACL), pp. 13697–13720. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.813.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 9440–9450. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.511.

Zhilin Wang, Jaehun Jung, Ximing Lu, Shizhe Diao, Ellie Evans, Jiaqi Zeng, Pavlo Molchanov, Yejin Choi, Jan Kautz, and Yi Dong. ProfBench: Multi-domain rubrics requiring professional knowledge to answer and judge. In *International Conference on Learning Representations (ICLR)*, 2026.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. BrowseComp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-limited LLM benchmark. In *International Conference on Learning Representations (ICLR)*, 2025.

Cheng-Kuang Wu, Zhi Rui Tam, Chieh-Yen Lin, Yun-Nung Chen, and Hung-yi Lee. StreamBench: Towards benchmarking continuous improvement of language agents. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37, pp. 107039–107063, 2024. doi: 10.52202/079017-3398.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations (ICLR)*, 2025.

Di Wu, Zixiang Ji, Asmi Kawatkar, Bryan Kwan, Jia-Chen Gu, Nanyun Peng, and Kai-Wei Chang. LongMemEval-V2: Evaluating long-term agent memory toward experienced colleagues. *arXiv preprint arXiv:2605.12493*, 2026.

xAI. Model Card: Grok 4.6. `https://media.x.ai/v1/website/card-4p6-4cd2dc57.pdf`, last accessed 2026/09/26, 2026a.

xAI. Model Card: Grok 4.7. `https://media.x.ai/v1/website/4p7card-5eccc980.pdf`, last accessed 2026/09/26, 2026b.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J. Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37, pp. 52040–52094, 2024. doi: 10.52202/079017-1650.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. TheAgentCompany: Benchmarking LLM agents on consequential real world tasks. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 38, 2025a. doi: 10.52202/085713-0315.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 17577–17604, 2025b. doi: 10.52202/085713-0593.

Neeraj Yadav. Temporal validity in retrieval memory: Eliminating stale-fact errors for AI agents over evolving knowledge. *arXiv preprint arXiv:2606.26511*, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations (ICLR)*, 2025.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *International Conference on Learning Representations (ICLR)*, 2024.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to evaluate long-context models effectively and thoroughly. In *International Conference on Learning Representations (ICLR)*, 2025.

Z.ai. GLM-5.3: Frontier Coding with Emergent Cyber Capabilities. `https://z.ai/blog/glm-5.3`, last accessed 2026/09/26, 2026.

Mario Zechner. earendil-works/pi: AI agent toolkit: unified LLM API, agent loop, TUI, coding agent CLI. `https://github.com/earendil-works/pi`, release v0.84.4, last accessed 2026/09/26, 2026.

Michael Zhang and Eunsol Choi. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7371–7387. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.586.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞Bench: Extending long context evaluation beyond 100K tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 15262–15277. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.814.

Junhao Zheng, Xidi Cai, Qiuke Li, Duzhen Zhang, ZhongZhi Li, Yingying Zhang, Le Song, and Qianli Ma. LifelongAgentBench: Evaluating LLM agents as lifelong learners. *arXiv preprint arXiv:2505.11942*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 19724–19731, 2024. doi: 10.1609/aaai.v38i17.29946.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations (ICLR)*, 2024.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pp. 3277–3287. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.254.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Gong, Diyi Yang, and Xing Xie. DyVal: Dynamic evaluation of large language models for reasoning tasks. In *International Conference on Learning Representations (ICLR)*, 2024.

The appendix is organized as follows. Appendix A gives the results by capability, criterion type, and domain, the reasoning-effort ablation, and each system's first context windows. Appendix B gives the record selection rules, preprocessing, dataset statistics, and capability mapping, and Appendix C the task-construction funnel step by step. Appendix D describes the replay, the confinement, and the judge, Appendix E states the rubric contract, and Appendix F extends the related work. Appendix G follows two agents through each of six tasks, call by call. The last two appendices give the prompt templates of construction, the agent, and the judge (Appendix H) and walk through sixteen evaluated tasks (Appendix I).

A Additional results

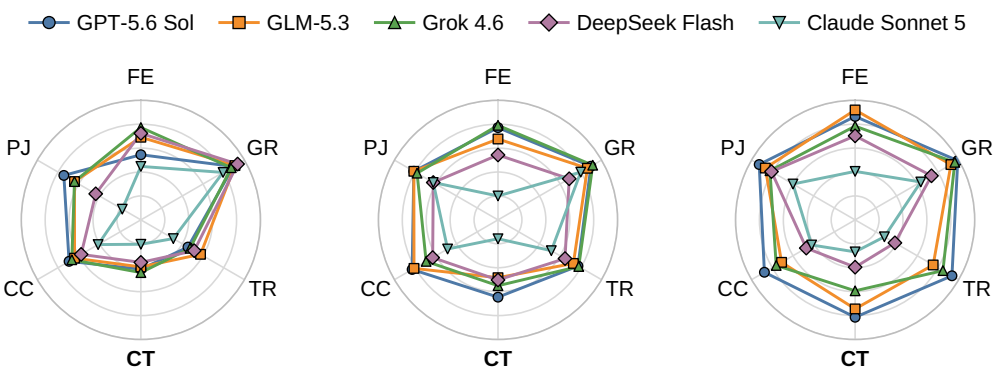
A.1 Capabilities and criterion types

Table A.1 pools the criterion pass rate by capability over the three domains. The capabilities differ in how far they separate the systems. Grounding spans 8.2 points across the seven, from 87.7% to 95.9%, and fact extraction spans 21.6, from 67.5% to 89.1%. Lowering GLM-5.3 from high to low effort costs change tracking the most, 11.0 points, and comparison and calculation the least, 2.7. By criterion type the order is the same for all seven systems. Citation criteria pass at 86.7% to 95.6%, correctness criteria at 69.5% to 88.4%, and reasoning criteria at 62.3% to 82.4% (Table 3).

Table A.1: Pooled over the three domains, change tracking is the capability every system meets least. Criterion pass rate (%) by the capability each criterion checks, over the 300 tasks, criterion counts in parentheses: fact extraction (FE), grounding (GR), temporal reasoning (TR), change tracking (CT), comparison and calculation (CC), professional judgment (PJ). Marks as in Table 3.

System	*FE* (302)	*GR* (1,677)	*TR* (647)	*CT* (592)	*CC* (793)	*PJ* (238)
GPT-5.6 Sol	85.1	**95.9**	**89.6**	**80.6**	**88.8**	**93.3**
GLM-5.3†	87.1	95.0	85.8	76.4	85.2	90.8
Grok 4.6	**89.1**	95.3	88.1	76.2	84.9	89.1
DeepSeek Flash	82.8	90.8	75.7	70.9	78.3	84.9
Claude Sonnet 5	67.5	87.7	69.2	60.3	71.8	76.9
GLM-5.3 low	78.1	90.2	80.5	65.4	82.5	84.5
GLM-5.3 max	82.8	94.4	86.1	75.8	85.0	87.8

Figure A.1 draws the same pass rates for each domain, for the five systems. Grounding is the longest axis of all five in finance and in medical. The temporal axes are shortest in finance, where no system meets more than 78.8% of the finance temporal-reasoning criteria or 72.0% of the change-tracking ones, against 89.0% and 82.3% in medical and 96.7% and 90.7% in legal. Legal separates the systems most. GPT-5.6 Sol's legal profile is nearly round, from 90.7% on change tracking to 99.5% on grounding, and legal temporal reasoning runs from 64.1% for Claude Sonnet 5 to 96.7% for GPT-5.6 Sol, the widest spread of any capability in any domain.



(a) Finance (b) Medical (c) Legal

Figure A.1: Change tracking (CT) is the shortest axis of every system in medical, of four of the five in finance, and of three in legal. Criterion pass rate by the capability each criterion checks: fact extraction (FE), grounding (GR), temporal reasoning (TR), change tracking (CT), comparison and calculation (CC), and professional judgment (PJ). The radius runs from 50% at the center to 100% at the rim, with rings every 10 points from 60%, and the values are those of Tables A.2 to A.4.

A.2 Results by domain

Tables A.2 to A.4 give each domain by professional task and by capability, with the task and criterion counts, and with GLM-5.3 at low and max effort. Over all seven rows, change tracking is the lowest capability of every system in medical, of six of the seven in finance, and of four in legal.

In finance, comparable company analysis is both the most common task, 32 of the 102, and the hardest for all five systems, at 6.2% to 18.8%. Risk and contingency disclosure is the second hardest for four of the five. DeepSeek Flash leads credit analysis and M&A and restructuring, and ties GPT-5.6 Sol on financial reporting review. Low effort costs GLM-5.3 the least in finance, 4.9 points of task success, against 15.1 in medical and 21.5 in legal (Table 3).

Table A.2: Comparable company analysis is the hardest finance task for all five systems. Left, task success (%) by professional task, and right, criterion pass rate (%) by capability as in Table A.1, with counts in parentheses. Task abbreviations follow Section 4.2. Marks as in Table 3.

System	Task success by professional task						Criterion pass rate by capability					
	Credit (15)	*Report.* (11)	*M&A* (17)	*Comps* (32)	*Gov.* (12)	*Risk* (15)	*FE* (123)	*GR* (784)	*TR* (99)	*CT* (211)	*CC* (378)	*PJ* (39)
GPT-5.6 Sol	53.3	**63.6**	47.1	15.6	33.3	**40.0**	77.2	94.6	72.7	70.6	**84.7**	**87.2**
GLM-5.3†	53.3	54.5	52.9	**18.8**	50.0	20.0	84.6	95.4	**78.8**	69.7	82.0	82.1
Grok 4.6	60.0	54.5	52.9	**18.8**	**58.3**	26.7	**88.6**	93.6	73.7	**72.0**	83.3	82.1
DeepSeek Flash	**66.7**	**63.6**	**58.8**	15.6	50.0	20.0	86.2	**96.7**	75.8	67.8	78.8	71.8
Claude Sonnet 5	46.7	36.4	23.5	6.2	25.0	13.3	72.4	89.7	65.7	60.2	70.6	59.0
GLM-5.3 low	26.7	63.6	47.1	21.9	41.7	13.3	79.7	90.1	70.7	58.3	79.1	71.8
GLM-5.3 max	33.3	54.5	41.2	25.0	41.7	26.7	74.8	93.1	78.8	66.4	81.2	79.5

In medical, labeling change management holds 67 of the 119 tasks, and GPT-5.6 Sol leads it at 58.2%. GLM-5.3 leads clinical decision support and dosing and dispensing, at 80.0% and 57.9%, although clinical decision support has five tasks, so each moves its rate by 20 points. Change tracking is the lowest capability of all seven systems in medical, and after fact extraction it separates the five most, from 57.9% for Claude Sonnet 5 to 82.3% for GPT-5.6 Sol.

Table A.3: In medical, change tracking is the capability every system meets least. Tasks: formulary management (Form.), clinical decision support (CDS), dosing and dispensing (Dosing), regulatory information management (RIM), labeling change management (Label.). Layout as in Table A.2.

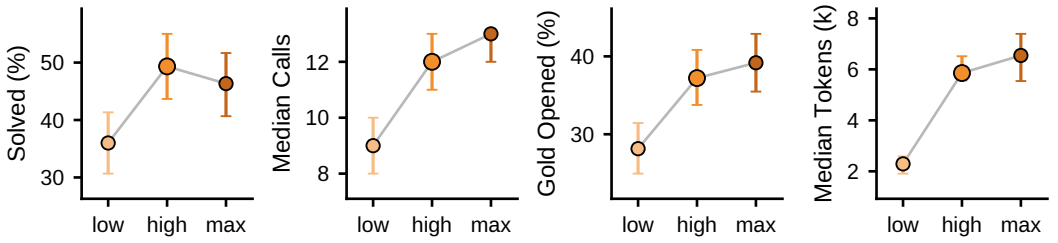
System	Task success by professional task					Criterion pass rate by capability					
	Form. (18)	*CDS* (5)	*Dosing* (19)	*RIM* (10)	*Label.* (67)	*FE* (105)	*GR* (467)	*TR* (272)	*CT* (209)	*CC* (222)	*PJ* (64)
GPT-5.6 Sol	**66.7**	60.0	47.4	**50.0**	**58.2**	88.6	94.9	88.6	**82.3**	**91.4**	**90.6**
GLM-5.3†	50.0	**80.0**	**57.9**	**50.0**	56.7	83.8	93.1	86.4	74.2	90.5	**90.6**
Grok 4.6	33.3	40.0	52.6	30.0	56.7	**89.5**	**95.7**	**89.0**	77.5	84.7	89.1
DeepSeek Flash	**66.7**	60.0	36.8	**50.0**	49.3	77.1	84.4	82.4	75.1	81.5	81.2
Claude Sonnet 5	22.2	20.0	21.1	40.0	28.4	60.0	89.9	75.7	57.9	74.3	81.2
GLM-5.3 low	33.3	60.0	52.6	50.0	37.3	73.3	89.3	80.9	67.0	87.8	85.9
GLM-5.3 max	61.1	20.0	68.4	50.0	49.3	86.7	94.4	86.8	74.6	91.4	87.5

In legal, GPT-5.6 Sol leads applicability determination by 22.2 points over the next of the five systems and regulatory history research by 21.0, the two widest leads in any task column of the three tables. Legal is DeepSeek Flash's weakest domain. Its task success there is 32.9%, against 40.2% in finance and 50.4% in medical, and temporal reasoning and change tracking are its two lowest capabilities, at 69.2% and 69.8%.

Table A.4: GPT-5.6 Sol leads or ties every legal column except fact extraction. Tasks: applicability determination (Appl.), compliance reporting and submissions (Report.), rulemaking tracking (Rule.), regulatory history research (Hist.), regulatory gap analysis (Gap). Layout as in Table A.2.

System	Task success by professional task					Criterion pass rate by capability					
	Appl. (18)	*Report.* (7)	*Rule.* (27)	*Hist.* (19)	*Gap* (8)	*FE* (74)	*GR* (426)	*TR* (276)	*CT* (172)	*CC* (193)	*PJ* (135)
GPT-5.6 Sol	**61.1**	**57.1**	**59.3**	**84.2**	**62.5**	93.2	**99.5**	**96.7**	**90.7**	**93.8**	**96.3**
GLM-5.3†	38.9	**57.1**	55.6	63.2	**62.5**	**95.9**	96.2	87.7	87.2	85.5	93.3
Grok 4.6	38.9	**57.1**	48.1	52.6	50.0	89.2	98.1	92.4	79.7	88.1	91.1
DeepSeek Flash	38.9	14.3	40.7	31.6	12.5	85.1	86.9	69.2	69.8	73.6	90.4
Claude Sonnet 5	11.1	28.6	14.8	26.3	12.5	70.3	81.7	64.1	63.4	71.0	80.0
GLM-5.3 low	33.3	42.9	37.0	31.6	12.5	82.4	91.3	83.7	72.1	82.9	87.4
GLM-5.3 max	55.6	42.9	55.6	52.6	37.5	90.5	96.7	88.0	89.0	85.0	90.4

A.3 Reasoning effort



(a) Task Success (b) Tool Calls (c) Gold Evidence Opened (d) Reasoning Tokens

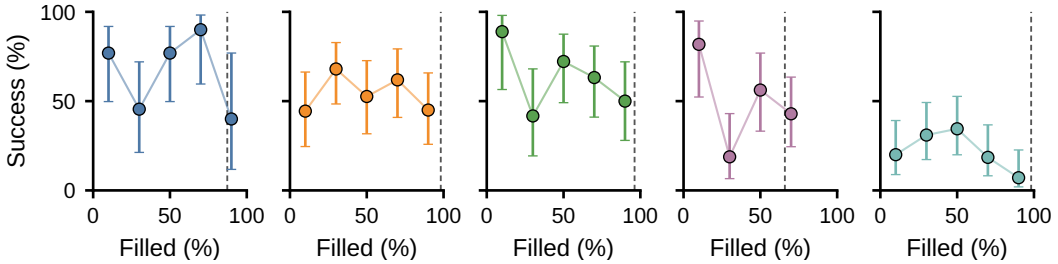
Figure A.2: More reasoning effort helps GLM-5.3 up to high, with more search, and max adds reasoning tokens but not task success. GLM-5.3 only, with high, the default, outlined (95% bootstrap intervals).

Figure A.2 details the effort ablation of Section 5.2. From low to high effort GLM-5.3 searches more, and from high to max its search changes little while its reasoning tokens grow. Paired by task, high effort solves 64 tasks that low effort misses and misses 24 that low effort solves (exact McNemar test, $p < 0.001$). Against max, high solves 44 tasks that max misses and misses 35 that

max solves ($p = 0.37$), so the 3.0-point gap between high and max is not distinguishable from noise on these 300 tasks. Of the 300, 68 are solved at all three settings and 106 at none. The ablation holds GLM-5.3 against itself, so it does not say how effort acts on the other systems.

A.4 The first context window

Figure A.3 reads the first context window of each of a system's three lifetimes, from the first task to the first compaction, before any summary replaces the conversation. The windows hold 1.05M tokens for GPT-5.6 Sol, 500K tokens for Grok 4.6, and 1M tokens for the other three, and the three first windows span 52 tasks for GPT-5.6 Sol, 103 for GLM-5.3, 74 for Grok 4.6, 64 for DeepSeek Flash, and 138 for Claude Sonnet 5. Pooled over the five systems, success is 49.0% for tasks that arrive with three to four fifths of the window filled and 30.4% above four fifths, and GPT-5.6 Sol moves from 90.0% to 40.0% between those two fifths, on 10 and 5 tasks. This view fixes neither the task nor the lifetime position, which the estimate of Section 5.4 holds fixed.



(a) GPT-5.6 Sol (b) GLM-5.3 (c) Grok 4.6 (d) DeepSeek Flash (e) Claude Sonnet 5

Figure A.3: Within its first context window, every system that reaches the fullest fifth scores lower there than in the fifth before. Task success by how full the window is when a task arrives, with 95% Wilson intervals. The dashed line marks the first compaction.

B Benchmark details

Sources in detail. Finance takes every 8-K current report, 10-Q and 10-K periodic report, and amendment filed between 1 January 2016 and 20 May 2026, from each company's EDGAR filing index. Medical takes every label version from the DailyMed monthly release files from January 2016 onward, since the DailyMed web service returns only the current version, and every safety labeling change the FDA recorded between 1 January 2016 and 25 July 2026. Legal takes every final and proposed rule that the Federal Register API lists as affecting a selected part, with its published full text from GovInfo, and a snapshot of each part's codified text from the eCFR on 1 January of every year from 2017 to 2026 and on 29 July 2026. Three quarters of the finance documents are 8-K current reports.

Record selection. Since the evaluation turns on replacement, entities were selected for a history of revision rather than sampled at random, by explicit rules applied to what was acquired. *Finance* starts from the 1,712 companies that a full-text search of EDGAR finds carrying the Item 4.02 non-reliance statement between 2016 and 2026. Removing shell companies and companies with no annual report before the notice leaves 1,058 operating companies, of which we keep the 200 with the most amendments, non-reliance notices, and restated XBRL values, and acquire their 33,313 filings. Of these, a company qualifies if its filings span at least eight years, number at least 60, and carry at least 100 value changes visible in text. The qualifying companies are split into five strata of revision density, the share of opportunities for a value to change that it did, and ten are kept per stratum, 50 in all across 39 industry classifications. *Medical* starts from the 241,293 labels DailyMed publishes and keeps the prescription labels of new drug and biologics applications held by the application sponsor rather than a repackager, 6,458 labels of 3,679 applications. Of these, 1,527 have at least six label versions spanning at least eight years. Excluding medical gases, 150 applications have at least 20 versions, a nine-year span, ten FDA safety labeling changes, and one contraindication change, and 50 of them are kept. *Legal* starts from the 333 parts of Titles 21 and 40 with a recorded amendment history, acquiring 2,519 Federal Register documents and 4,052 eCFR snapshots. A part qualifies when its history supports a multi-state task, which requires at least two

transitions, a medium or long span, an update in force immediately, a provision that stays unchanged as a control, and a boundary between versions. Of the 333 parts, 229 qualify. They are split into five strata of revision density, and ten are kept per stratum, allocated between the two titles in proportion to their share. The record was fixed first, and the tasks were written over it afterwards.

Preprocessing. The parser uses no model. A filing keeps the text of its primary document, with line structure preserved and tables flattened to text. A Federal Register rule keeps its published full text, with symbol markup mapped to characters and stray NUL bytes removed. An eCFR snapshot and a label version are parsed from XML, the former into part, subpart, and section headings and the latter into one heading per label section, with the clinical sections first and the ingredient and packaging sections kept at the end. A safety labeling change becomes a short record of the drug, the application, the action date, and the label sections that changed.

Publication dates and entities. The publication date is the EDGAR filing date, the Federal Register publication date, the eCFR snapshot date, the DailyMed publication date, or the FDA action date. Effective dates are kept only as metadata, and in the 400 sampled label versions they differ from the publication date by as much as 3,682 days. A document is attached to every entity it concerns. A filing is attached to its company, a rule to each CFR part it cites, and a label or a labeling change to its application, so that a label with several set identifiers joins one lifetime. A document that concerns several entities is stored once. Such documents include 108 rules that touch up to four selected parts and 101 labeling documents that touch two applications. Each domain's stream keeps the documents of its 50 entities dated from 1 January 2016 to its cutoff, sorted by date, with documents released on a task's question date placed before the task.

Released files. Table B.1 lists every field of the three released files. A task and its gold record share one task identifier, and every document a gold record cites appears in the document file with its publication date.

Table B.1: Every task links to its gold record by its identifier, and every cited document to the document file. Fields of the released files.

File	Field	Content
`tasks.jsonl`	`task_id`	task identifier
	`domain`	finance, medical, or legal
	`ask_at`	question date τ_j
	`question`	statement q_j
	`entity_ids`, `entity_count`	the one to four entities the task is about
	`task_type`	construction line that produced the task
`gold.jsonl`	`evidence`	gold evidence G_j, one entry per supporting document with its identifier, entity, publication date, quoted passage, and what it supports
	`rubrics`	rubric R_j, each criterion with its identifier, kind (A correctness, R reasoning, E citation), and text
	`point_counts`	number of criteria of each kind
	`reference_answer`	the verified answer a_j^*
`documents.jsonl`	`document_id`, `path`	document identifier and its plain-text file
	`published_at`	publication date t_i
	`source`, `document_type`	official source and form, such as `sec_edgar` and 8-K
	`domain`, `entities`, `bytes`	field, the entities the document is about, and its size

Statistics. Figure B.1 shows how the 920 tasks spread over question years, evidence dates, evidence spans, and criteria. The released tasks carry at most 77 criteria each, and their question dates fall on 132 distinct dates in finance, 157 in medical, and 76 in legal.

Professional tasks. Table B.2 defines the 16 professional tasks, each in the terms its practitioners use.

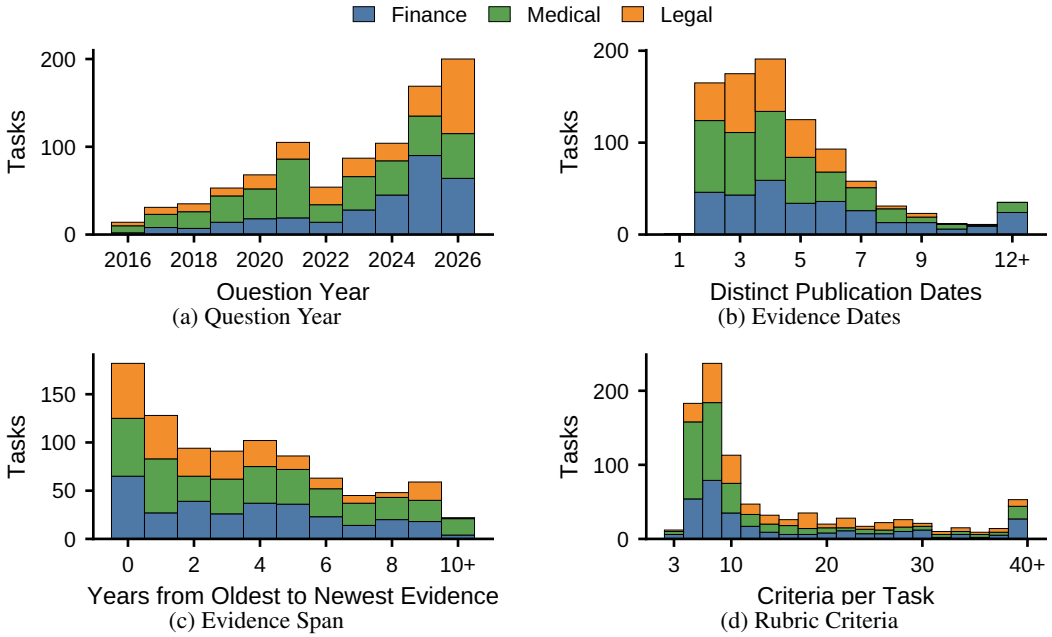
Table B.2: Each field's professional tasks, named as its practitioners name them, and what a task of each kind asks about.

Professional task	What the task asks about
Finance, for credit and audit committees, deal teams, and analysts working from SEC filings	
Credit analysis (Credit)	A borrower's credit standing, such as liquidity, covenants, leverage, debt maturities, and refinancing
Financial reporting review (Report.)	How figures are reported and whether they can be relied on, such as restatements, non-reliance notices, control weaknesses, and new accounting standards
M&A and restructuring (M&A)	Acquisitions, divestitures, spin-offs, purchase accounting, restructuring programs, and reorganizations
Comparable company analysis (Comps)	One defined metric computed for several companies at aligned period ends and compared
Governance and capital markets (Gov.)	Share classes and voting control, executive compensation, listing compliance, and the terms of equity and debt issues
Risk and contingency disclosure (Risk)	Litigation, insurance losses, hedging, cybersecurity, and the regulatory regimes that bind the company
Medical, for formulary and medication-safety committees, pharmacists, and labeling teams working from FDA labels	
Formulary management (Form.)	Formulary policy for a drug or a drug family, such as substitution, restrictions, and covered populations
Clinical decision support (CDS)	Interaction and safety alerts in a clinical system, which products they cover, and from which date
Dosing and dispensing (Dosing)	The dose, regimen, quantity, or timing for a stated patient under the label in force at the date
Regulatory information management (RIM)	Whether labels and product records of related applications agree, and which one governs
Labeling change management (Label.)	How an FDA action reached the published label, which sections changed and when, and how a warning evolved
Legal, for compliance and regulatory-reporting teams and counsel working from federal regulation	
Applicability determination (Appl.)	Which requirements apply to a facility, product, or activity at a date, and what obligations follow
Compliance reporting and submissions (Report.)	Which reports, records, or filings were due for a period, under which form, and where they must go
Rulemaking tracking (Rule.)	A rule's path from proposal to final, effective, stayed, corrected, or withdrawn, with its dates
Regulatory history research (Hist.)	How a provision, definition, or requirement changed from proposal to final rule and across later revisions
Regulatory gap analysis (Gap)	Whether one procedure or text satisfies requirements across programs or parts, and where they differ

Grok 4.7 (xAI, 2026b) at xhigh reasoning effort assigned each task its professional task from the task statement alone. Every task received a label, and 7 of the 920 are marked as a weak fit. An independent labeling made with an earlier version of the categories agrees on 82.8% of the 789 tasks where the two versions map one to one.

Capabilities. Table B.3 defines the six capabilities. Grounding is what 4,867 of the 6,181 citation criteria check.

Each of the 13,318 criteria of the 920 tasks was labeled with the one capability it checks by Grok 4.7, in a fresh session per task that could read only the task statement and its rubric. The 300 evaluated tasks and 23 others were labeled at xhigh reasoning effort, 83 tasks at medium effort, and the other 514 with reasoning off, covering 4,412, 640, and 8,266 criteria, and the result tables use only the xhigh labels of the evaluated tasks. The labeler chose among eight capabilities with a fixed tie-break order, merged into the six of Table B.3. Version resolution joins temporal reasoning, cross-entity alignment and calculation join comparison and calculation, and evidence retrieval is named grounding. Every criterion received a label. A stratified sample of 40 labels from the evaluated tasks, five per original label, was read by hand, and 36 of the 40 agree with the reader's choice, 37 (92.5%) after merging to six columns.



(a) Question Year (b) Evidence Dates (c) Evidence Span (d) Rubric Criteria

Figure B.1: Half of the tasks (51.4%) are asked in 2024 to 2026, and most draw on two to six publication dates. The last bins of (b), (c), and (d) hold 12 or more dates, 10 or more years, and 40 or more criteria.

Table B.3: Grounding is the most frequent capability in every domain. What each capability's criteria ask, and how many criteria it has per domain over the 920 tasks.

Capability	The criterion asks the answer to	Fin.	Med.	Leg.
Fact extraction (FE)	report a fact or value as it stands in one record at one point	317	330	218
Grounding (GR)	locate and cite the source that supports a stated result	2,188	1,479	1,200
Temporal reasoning (TR)	decide what held when: the version in force, order, dates, and deadlines	369	1,028	996
Change tracking (CT)	state what changed between states, and whether a later state continues or reverses an earlier one	639	656	526
Comparison and calculation (CC)	compare entities on aligned terms, or derive a quantity	1,246	837	530
Professional judgment (PJ)	apply rules to facts, or reach an evaluative conclusion	145	284	330
All six		4,904	4,614	3,800

C CONSTRUCTION FUNNEL

Table C.1 gives the task-construction funnel of Section 4 step by step, with what each step reads, when it rejects a candidate, and how many candidates it removed. Deduplication and the single-source check are mechanical, and a language-model worker runs every other automated step. The workers are models from the GPT-5.6 and Grok families.

Stage I in full. In *generation*, two workers run in turn. The first proposes a group of related entities and the kinds of dated states that connect them, and the second reads the group's records and proposes a question date τ_j, a statement q_j, gold evidence G_j, and an answer a_j^*. In *verification*, a second worker rebuilds the answer from $\mathcal{A}(\tau_j)$ alone and treats the generator's claims as proposals. It checks six things. First, every entity, value, date, period, unit, and legal status matches the original text. Second, the evidence covers every part of the answer, and every other official document with the same support is recorded. Third, the task needs at least two dated states, and no single document, single date, later summary, or latest state may give the answer. Fourth, every named entity is needed. Fifth, answering takes professional work, such as reconciling restated values, choosing the version in force, or applying a condition. Sixth, every defect is named, and no unsupported claim remains. Each field adds its own rules. In finance, for example, the worker recomputes every period alignment, unit conversion, and calculation. A candidate passes only if every check holds, and verification removes 918 candidates.

Stage II in full. In *rubric writing*, a third worker splits a_j^* into the binary criteria of R_j. It sees q_j, a_j^*, and the quoted passages of G_j, but not the source documents. The criteria come in three kinds. A *correctness* criterion states one fact the question asks for. A *citation* criterion states what an official document must support, and any official document with the same support counts. A *reasoning* criterion states a step the answer must write out and names a plausible error that it catches. In the *rubric audit*, a fourth worker checks every criterion against the documents of $\mathcal{A}(\tau_j)$. It rejects a task when a criterion is wrong, redundant, not asked for, incomplete, or impossible to judge. It also rejects a task when the criteria miss part of the question, go beyond it, or demand one exact file or wording. The *criterion-type check* confirms the kind of each criterion from the question alone (Figure H.17), and the *evidence-coverage check* confirms that the recorded evidence covers every correctness criterion (Figure H.18). The audit removes 137 tasks.

Stage III in full. *Deduplication* removes 54 tasks that repeat another. In the *answer-leak check*, a worker sees only the statement, its entities, and its question date, and a *method-leak check*, which also sees a sample of accepted statements, completes it (Figure H.19). It checks that the statement names the entities, the professional scenario, and the period clearly. The statement must not reveal answer values, directions of change, categories, or a list of expected outputs, and it may name a file or a section only when that document is the professional object of the question. The worker keeps the statement, rewrites it with the smallest natural change, or drops it, and a second reviewer must approve every rewrite. This check removes 30 tasks. The *single-source check* recomputes the multi-state test from the evidence dates by a fixed rule and removes 2 more. In *human review*, a domain expert reads each of the 977 remaining tasks with its answer, evidence, quoted passages, criteria, and audit verdict. The expert decides whether the task can be released exactly as shown, and rejects 57. Every released task has therefore passed every automated check and an expert's reading.

Figure C.1 draws both selections behind TenureBench. Panel (a) follows each domain from its full source to the 50 entities we keep, by the rules of Appendix B, with bar length on a log scale. Panel (b) follows the 2,118 candidate tasks through the construction steps to the 920 released, with the number each step removes on the right. Verification removes the most candidates, 918, and human review removes 57 of the 977 that reach it, so the automated steps carry most of the filtering.

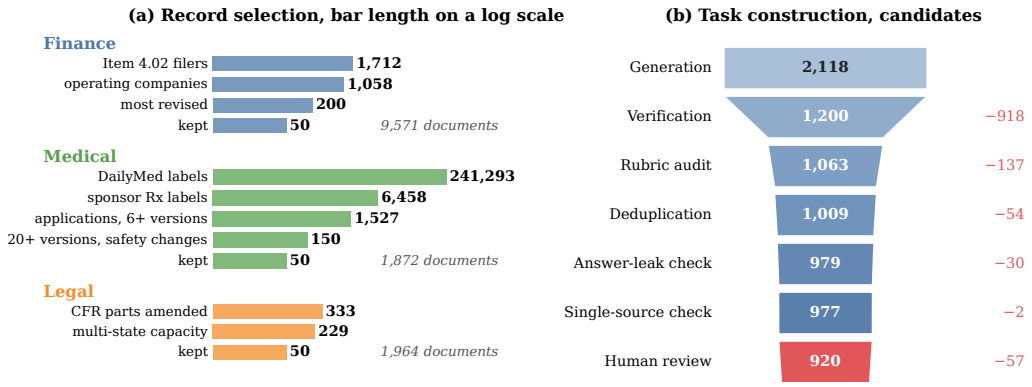


Figure C.1: Each domain keeps 50 entities out of a large pool (a), and verification removes the most candidates (b). (a) Record selection in each domain, from the full source to the 50 entities kept (bar length on a log scale). (b) Candidate tasks left after each construction step, with the number removed at right.

Table C.1: Verification rejects more candidates than all later steps together. Δ is the number of candidates a step removes, and the italic rows describe steps that produce rather than reject.

Step	Reads	Rejects a candidate when	Δ	Left
Generation	the documents of one entity group that its task file lists	*proposes the question date, statement, evidence, and answer, in eight runs*		2,118
Verification	the record visible at the question date	that record does not support the answer, or a single document or a single date does	−918	1,200
Rubric writing	the statement, the answer, and the evidence spans	*writes the criteria, none demanding one exact wording*		1,200
Rubric audit	the statement, the criteria, and the evidence	a criterion is not backed by the evidence or has the wrong type, or the evidence misses a fact	−137	1,063
Deduplication	all candidates	the task repeats another	−54	1,009
Answer-leak check	the statement alone, and accepted statements from an earlier reviewed set	the statement gives away its answer or its method, and its rewrite is not equivalent to it	−30	979
Single-source check	the evidence dates	one date or one document closes the task	−2	977
Human review	every task that passed the automated steps	a domain expert rejects the statement or the rubric	−57	920

D PROTOCOL DETAILS

Replay. The harness sets each file's timestamp to its publication date. Before each task it updates a clock file with the question date, and then it sends the task message of Section 3. Every run records, for each task, the answer, the tool calls, the token counts, the compactions, and an audit of every file access.

Confinement. Throughout the paper, the agent's workspace is its whole folder, which holds the archive and one writable subfolder (`workspace/` in the agent's instructions, Figure H.21). A tool-call hook blocks any path that resolves outside the folder and any write outside the writable subfolder, and it refuses shell, web access, and subagents. No call in the evaluated runs was blocked for reaching outside the folder.

Judge. The judge, GPT-5.6 Sol at low reasoning effort, scores all criteria of one task in a single fresh session that has seen nothing else. It runs in the same harness with read-only file tools (`read`, `grep`, `find`, and `ls`) and no network. A criterion enters with its identifier, type, and text. A citation criterion also carries the text of the correctness criteria it backs, and a reasoning criterion carries the error it exists to catch. Every document the answer cites enters in full text.

E RUBRIC CONTRACT

A rubric is fixed together with its task before any agent is evaluated, and it is never edited afterwards.

Correctness criteria. Correctness criteria, type A in the templates of Appendix H, state the facts the answer must report, keeping entity, value, period, scope, polarity, and modality. They may not demand fields the task does not request, and they may not score a value the task statement already gives (Figure H.17).

Citation criteria. Citation criteria, type E, describe what a source must support, not which file it is, and they are judged by content. Any official document visible at the question date that gives the same support counts, whether or not the construction pipeline used it. No criterion may require an exact quotation, a document identifier, or a filename. Each citation criterion names the correctness criteria it backs and is judged against them, because a citation supports a stated result. A document published after the state it describes is valid evidence for that state if it was visible at the question date.

Reasoning criteria. Reasoning criteria, type R, name the steps an answer must write out to reach its results, such as whether a later state continues or reverses an earlier one, which version was in force, or how two figures are reconciled. Each names the error it exists to catch, and that error must leave every correctness criterion right while the reasoning criterion fails. A reasoning criterion that cannot fail while all correctness criteria pass carries no weight. The number of criteria follows what the task requires, not a fixed quota. A model's hidden reasoning is never judged, so a step counts only when the answer writes it out.

Scoring. The judge never sees the verifier's notes or the construction quotations, and the measures of Section 3 follow from the criterion verdicts by a fixed rule.

F EXTENDED RELATED WORK

Time in question answering. Temporal question answering ties each answer to a time or a situation (Chen et al., 2021a; Zhang & Choi, 2021; Dhingra et al., 2022), TemporalWiki updates a model between consecutive snapshots of a knowledge source (Jang et al., 2022), and language models degrade on text from after their training period (Lazaridou et al., 2021). Tool-using agents such as ReAct (Yao et al., 2023) answer by acting in an environment, as the agent in TenureBench does.

Long contexts and memory. Long-context benchmarks test how models use long inputs (Hsieh et al., 2024; Bai et al., 2024; Zhang et al., 2024; Yen et al., 2025; Liu et al., 2024a), and memory systems let agents work past the context window (Packer et al., 2023; Park et al., 2023; Shinn et al., 2023; Zhong et al., 2024; Xu et al., 2025b). LongMemEval, LoCoMo, and MemoryAgentBench evaluate the memory once a history is complete (Wu et al., 2025; Maharana et al., 2024; Hu et al., 2026), and LongMemEval-V2 tests a memory module on its own (Wu et al., 2026). TenureBench leaves memory to the agent and evaluates the whole agent while the history grows.

Judges and rubrics. LLM judges (Zheng et al., 2023; Liu et al., 2023b; Kim et al., 2024b) are sensitive to answer order and to their own outputs (Wang et al., 2024; Verga et al., 2024). Binary rubrics grade long-form answers (Arora et al., 2025; Dou et al., 2026; Starace et al., 2025; Gunjal et al., 2026), retrieval-augmented generation grounds answers in retrieved passages (Lewis et al., 2020; Karpukhin et al., 2020; Asai et al., 2024), and attribution work checks whether cited passages support an answer (Bohnet et al., 2022; Gao et al., 2023; Liu et al., 2023a). TenureBench grades each answer with binary criteria and checks citations against the documents the agent cited.

Contamination and construction. Contamination threatens static test sets (Sainz et al., 2023; Golchin & Surdeanu, 2024; Oren et al., 2024; Zhu et al., 2024), and mined benchmarks report their construction funnels (Jimenez et al., 2024; Center for AI Safety et al., 2026). TenureBench reports each construction step with the number of candidates it removes.

G CASE STUDIES

Figures G.1–G.6 follow two agents through the same task, one whose answer meets every criterion and one whose answer does not. The six tasks were picked by hand so that each shows a different way the history of the record defeats an answer (Table G.1), and every system that fails in one figure succeeds in another. The tool calls are copied from the run logs, which keep each call but not its output, so every note on what a file or a search contains is recomputed on the released documents. We checked each failure against the documents, and every judge verdict shown agrees with them.

Two properties of the setting recur. The grep tool returns at most 100 matching lines unless the agent asks for more, and it says when it stops, while a search of the whole archive for DORYX, Activella, or DARZALEX matches 1,124 to 3,183 lines by the question date. In addition, 30 of the 50 medical applications issue their labels in two or more streams, so the newest label of one stream need not be the newest label of the application. The failing agent of Figures G.2 and G.3 follows one stream.

Table G.1: Each case turns on a different way the record's history defeats an answer, and every system that fails in one case succeeds in another. *Fails* and *Meets every criterion* name the two agents each figure follows.

Fig.	Domain	What the record holds	Fails	Meets every criterion
G.1	Finance	A 10-K/A restates the opening balances an answer needs	GPT-5.6 Sol	GLM-5.3
G.2	Medical	A second label stream changes the label, and the gap recurs six questions later	Grok 4.6	GLM-5.3
G.3	Medical	A second label stream holds the whole history	GLM-5.3	GPT-5.6 Sol
G.4	Medical	A new paragraph shares no word with the search	Grok 4.6	GLM-5.3
G.5	Legal	The CFR snapshot in force is dated the question date	GLM-5.3	Grok 4.6
G.6	Legal	A proposed rule, made final ten months later	GLM-5.3	GPT-5.6 Sol

Two agents on a restated balance sheet tithonus-000742

Both agents are at task 31 of their finance lifetimes, with 6,898 documents in the workspace. Tool calls are copied from the run logs with archive paths shortened, and calls about AngioDynamics are left out. Notes after an arrow are recomputed on the released documents.

Message from the harness

```
Current date:  2023-03-31 (Fri)
Question:  As of March 31, 2023, using the consolidated annual
figures authoritative by that date, compare AngioDynamics'
cash-conversion-cycle change from fiscal 2021 to fiscal 2022
with Dentsply Sirona's change from fiscal 2021 to fiscal 2022.
Calculate CCC = DSO + DIO - DPO from average opening-and-closing
trade receivables, inventory, and accounts payable, ...
```

In the archive: Dentsply Sirona's balances at 31 December 2020 ($ million)

1 Mar 2022 The 10-K for 2021: receivables 673, inventories 466, payables 305.
7 Nov 2022 A 10-K/A for 2021, whose balance sheets are marked "As Restated": 667, 476, and 302.
1 Mar 2023 The 10-K for 2022, whose balance sheets cover 2022 and 2021 only.

GPT-5.6 Sol: 19 tool calls, and an answer that meets 14 of 20 criteria

2. `grep "DENTSPLY SIRONA"` in `archive/2023/03`
3. `find archive/2023/03/**/*818479*.txt`, Dentsply's filings of March 2023
7. `grep "Net sales|Cost of products sold|Inventories|Accounts receivable|Accounts payable"` in the 10-K of 1 Mar 2023
10. `read` the 10-K of 1 Mar 2023, twice
12. `find archive/2022/**/*818479*.txt` → Dentsply's filings of 2022, the 10-K/A among them. The file names give accession numbers, not form types.
15. `grep "Accounts and notes receivable-trade|Inventories, net|Accounts payable"` in the 10-K of 1 Mar 2022 → 673, 466, and 305 at 31 December 2020
16. The same search in the 10-K of 1 Mar 2023, then `read` both 10-Ks

Answer. "For Dentsply Sirona, FY2021 uses the restated comparative figures in the March 1, 2023 Form 10-K, including restated 2021 net sales and year-end balances." Its opening balances are 673, 466, and 305, and its Dentsply cycle for fiscal 2021 is 101.5 days.
Judge. ✗ Correctness 3: "reports Dentsply FY2021 CCC as 101.5 days, not 102.5; it used superseded original FY2020 opening balances (673/466/305) instead of the corrected 667/476/302." ✗ Correctness 6 and 7, which follow from it, and Citation 7 and Reasoning 1, for the same balances. ✗ Reasoning 3, for subtracting rounded cycles. The other 14 criteria are met.

GLM-5.3: 32 tool calls, and an answer that meets all 20 criteria

3. `find **/doc_818479*` over the archive, every Dentsply filing
4. `grep "FORM 10-K|FORM 8-K|FORM 10-Q"` in each of Dentsply's three filings of February and March 2023 (calls 4–6), then in three of February and March 2022 (calls 15–17)
23. A like search in the filings of 1 and 7 November 2022 (calls 23–25) → the last is the 10-K/A
26. `grep "CONSOLIDATED BALANCE SHEETS"` in the 10-K/A, then `read` it four times

Answer. "Dentsply Sirona figures are from its FY2022 Form 10-K (filed March 1, 2023) and the 2021 Form 10-K/A (filed November 7, 2022) — i.e., the restated/revised 2021 figures." Its opening balances are 667, 476, and 302, Dentsply's cycle is 102.5 days for fiscal 2021 and 125.4 for fiscal 2022, and AngioDynamics' change minus Dentsply's is −45.6 days.
Judge. ✓ All 20 criteria.

Figure G.1: A restatement can replace figures that a later report does not repeat. The 10-K/A of November 2022 restated Dentsply Sirona's balances at 31 December 2020, and the 10-K of March 2023 shows only 2022 and 2021. GPT-5.6 Sol takes the balances from the original 10-K for 2021 and never opens the amendment, though its own file listing includes it, so its fiscal 2021 cycle for Dentsply, and every result built on it, is off by one day. DeepSeek Flash and Claude Sonnet 5 use the same balances. GLM-5.3 checks the form type of each Dentsply filing, finds the amendment, and uses its restated balances.

Two agents on two Activella questions tithonus-000463, tithonus-000541

Tasks 47 and 53 of both agents' medical lifetimes, with 1,148 and 1,209 documents in the workspace. Calls about ESTROGEL are left out, and the rest is shown as in Figure G.1.

Messages from the harness

```
Current date:  2021-08-31 (Tue)
Question:  Between 2018 and 2021, how did Activella's boxed warning
frame the applicability of WHI cardiovascular, dementia, and
breast-cancer findings beyond the studied regimen?
Current date:  2021-12-22 (Wed)
Question:  An eligibility engine treats 'not indicated in
pregnancy' and 'contraindicated in pregnancy' as synonymous for
systemic menopausal hormone products.  Validate that assumption
for ACTIVELLA and ESTROGEL as of 2021-12-22 against their 2016 and
intervening label states.
```

In the archive: Activella's labels, in three streams, two of them here

2016–2019 The first stream, eight labels to 26 Mar 2019: pregnancy is a contraindication, and the WHI risks "should be assumed to be similar" for other products.

2019–2021 The second stream, begun 13 Mar 2019. Its label of 9 Apr 2020 drops pregnancy from section 4, and its label of 27 Aug 2021 puts "Only daily oral 0.625 mg CE and 2.5 mg MPA were studied" in place of that boxed-warning sentence and says Activella "is not indicated for use in pregnancy."

Grok 4.6: 2 of 6 criteria at question 47, and 2 of 6 at question 53

Question 47, 7 tool calls.

2. `grep "Activella|ACTIVELLA"` over the archive → 1,795 lines in 17 documents
3. `find **/*32a9f9a3*` over the archive, the first stream
5. `read` its labels of 28 Nov 2017 and 26 Mar 2019 (calls 5–6)

Question 53, 21 tool calls.

2. `grep "contraindicat|Pregnancy|not indicated"` in the first-stream labels of 30 Jun 2016 and 26 Mar 2019 (calls 2–3), with no search before
6. `find **/*32a9f9a3*` in `archive/2021` → no file

Answers. At question 47: "There is no 2018–2021 Activella flip to 'should not be assumed.'" At question 53: "ACTIVELLA last archive label 2019-03-26 (still in force at cutoff): unchanged."

Judge. ✗ At question 47, Correctness 2: the answer "explicitly denies any 2018–2021 flip." ✗ At question 53, Correctness 1: it "claims ACTIVELLA's last label (2019-03-26) was unchanged." Three more criteria fail on each question for the same labels.

GLM-5.3: all 6 criteria at question 47, and all 6 at question 53

Question 47, 10 tool calls.

1. `grep "Activella"` over the archive → 1,652 lines in 17 documents
2. `find **/doc_spl_32a9f9a3*.txt`, the first stream, and `read` two of its labels
5. `grep "Activella"` in `archive/2020`, then in `archive/2021` (call 6) → 109 lines in the label of 9 Apr 2020, and 96 in an FDA record and the label of 27 Aug 2021
7. `find **/doc_spl_d4b0c1dd*.txt`, the second stream, and `read` both labels

Question 53, 18 tool calls.

1. `find` the files of both Activella streams and of ESTROGEL (calls 1–3), with no search before, and later `read` the labels of 27 Aug 2021 and 9 Apr 2020 (calls 10 and 16)

Answers. At question 47, the 2021 revision "replaced the presumption-of-similarity sentence." At question 53, in the label of 27 Aug 2021 "§4 now lists eight contraindications … — pregnancy is absent."

Judge. ✓ All criteria on both questions.

Figure G.2: What an agent misses on one question it can miss again later. At question 47 GLM-5.3 searches 2020 and 2021 on their own and finds Activella's second label stream, which rewrote the boxed warning in August 2021, while Grok 4.6 stays with the first stream and answers that nothing changed. Six questions later both go straight to files they had listed before, and Grok 4.6 again takes the 2019 label as current, so it keeps a pregnancy contraindication that the second stream had dropped in April 2020.

Two agents on the DORYX pregnancy labeling tithonus-000295

Both agents are at task 114 of their medical lifetimes, with 1,846 documents in the workspace. Tool calls are copied from the run logs with archive paths shortened, and calls about RIFADIN are left out. Notes after an arrow are recomputed on the released documents.

Message from the harness

```
Current date:  2026-05-29 (Fri)
Question:  For a 2026 maternal-fetal-medicine formulary review,
reconstruct how the pregnancy-risk classification and supporting
evidence for DORYX delayed-release tablets and RIFADIN evolved from
their early archive labels, and advise whether the historical label
transitions warrant different treatment-eligibility or counseling
decisions at the cutoff.
```

In the archive: DORYX's own labels, in two streams

22 Apr 2016 The only label of the first DORYX stream, with "Pregnancy Category D."

2017–2023 Seven labels of the second DORYX stream, from 12 Dec 2017 to 31 May 2023. The label of 1 Dec 2021 still says "Pregnancy Category D". The label of 9 Oct 2022 drops the category and adds background rates of 2–4% for major birth defects and 15–20% for miscarriage.

2016–2026 Under the same application, DORYX MPC and unbranded doxycycline delayed-release labels in four more streams, and seven FDA safety-labeling records, one of them of 25 Jul 2022.

GLM-5.3: 14 tool calls, and an answer that meets 3 of 7 criteria

1. `grep "DORYX|doxycycline|RIFADIN|rifampin"` over the archive
3. `grep "DORYX"` over the archive → 1,124 lines in 27 documents
4. `find archive/**/doc_spl_37efefc3*` → the first stream, one label
5. `grep "regnancy|Teratogenic|PREGNANCY"` in that label of 22 Apr 2016

Answer. "The archive contains exactly one DORYX SPL (2016-04-22) and no DORYX safety actions." And: "Evolution: none reconstructable from this archive."

Judge. ✗ Reasoning 1: "The answer explicitly states that no 2021 DORYX label exists in the archive and that any post-2016 transition 'cannot be determined'." ✗ Correctness 1, Citation 1, and Reasoning 2 fail with it. The other three criteria, on RIFADIN, are met.

GPT-5.6 Sol: 12 tool calls, and an answer that meets all 7 criteria

1. `grep "DORYX"` over the archive
4. `find archive/**/doc_spl_99cf2de6-..._*.txt` → the second stream, seven labels
6. `grep "Pregnancy Category|8.1 Pregnancy|Risk Summary|..."` in the FDA record of 25 Jul 2022
8. `grep` the labels of 31 May 2023 and 12 Dec 2017 (calls 8–9), then `read` the label of 31 May 2023 (call 12)

Answer. DORYX "moved from Pregnancy Category D to narrative Pregnancy and Lactation Labeling Rule (PLLR) labeling in 2022. The category's deletion was regulatory-format conversion, not a finding that fetal risk disappeared." It calls the label of 12 Dec 2017 "the earliest archived DORYX delayed-release-tablet SPL in this stream."

Judge. ✓ All 7 criteria.

Figure G.3: An agent can report as absent what its search did not reach. DORYX's own labels come in two streams. GLM-5.3 lists the stream that holds a single 2016 label and concludes that the archive has one DORYX label and no safety action, so the change it was asked to trace cannot be reconstructed. GPT-5.6 Sol lists the other stream, which records the 2022 move from a pregnancy category to a narrative, and it says the earliest label it found is the earliest in that stream.

Two agents on the DARZALEX question tithonus-000591

Both agents are at task 55 of their medical lifetimes, with 1,224 documents in the workspace. Tool calls are copied from the run logs with archive paths shortened. The notes after an arrow are recomputed on the released documents.

Message from the harness

```
Current date:  2022-01-31 (Mon)
Question:  How did DARZALEX's guidance for interrupting and
restarting treatment after an infusion reaction evolve from the
initial approach to the current one?
```

Grok 4.6: seven tool calls, and an answer that meets six of eight criteria

1. `read CLOCK`
2. `grep "DARZALEX"` over the archive
3. `find **/*a4d0efe9*` over the archive, the DARZALEX label files
4. `grep "Grade 1|Grade 2|interrupt"` over the archive
5. `grep "Grade 1-2|Grade 3|Grade 4|anaphylactic"` in the label of 29 Jan 2022 → 61 lines, including the grade ladder and the anaphylaxis stop. The eye-symptom paragraph contains none of these words.
6. `read` the label of 30 Jun 2016
7. `read` the label of 1 Dec 2016

Answer, closing line. "Net: half-rate restart and three-strikes Grade 3 never left. What changed is when Grade 3 may restart (≤G2 → symptoms gone) and what ends therapy forever (G4 → G4 or anaphylaxis)."
Judge. ✗ Correctness 2: "the ocular-symptom branch requiring interruption and immediate ophthalmologic evaluation before restarting is entirely omitted." ✗ Reasoning 1, for the same omission. The other six criteria are met.

GLM-5.3: eight tool calls, and an answer that meets all eight criteria

1. `grep "DARZALEX|daratumumab"` over the archive
2. `find **/doc_spl_a4d0efe9*.txt` over the archive
3. `grep "Dosage Modifications for Infusion|interrupt the infusion|Discontinue DARZALEX"` in the label of 30 Jun 2016
4. `grep "infusion reaction|Infusion Reaction"` in the label of 30 Jun 2016
5. `grep "immediately interrupt|resume the infusion|restarting the infusion|Permanently discontinue DARZALEX"` in the label of 29 Jan 2022
6. `grep "decreases to Grade 2 or lower|Once reaction symptoms resolve, consider restarting"` over the archive
7. `grep "anaphylactic reaction or life-threatening|prior to restarting DARZALEX|Infusion-Related Reactions"` over the archive → matches the eye-symptom paragraph, which only the FDA record of 21 Jan 2022 and the label of 29 Jan 2022 contain
8. `grep "missed|Missed|split over two days|maximum rate of 200"` in the label of 29 Jan 2022

Answer, fourth change. "Current label (2022-01-29): a new symptom-specific restart gate for ocular reactions, and an explicit re-escalation cap. §5.1 now states: '… If ocular symptoms occur, interrupt DARZALEX infusion and seek immediate ophthalmologic evaluation prior to restarting DARZALEX' — the first restart conditioned on a specialist evaluation rather than symptom resolution alone."
Judge. ✓ All eight criteria.

Figure G.4: A new paragraph can share no word with a search built from the old ones. Neither agent opens the current label in full. Grok 4.6 searches it for grade words, which miss the new eye-symptom paragraph, and its answer leaves the rule out. GLM-5.3 searches the whole archive for restart wording and finds the paragraph. Both answer at the same point of their lifetimes, and the question is that of Figure I.6.

Two agents on the DEA hearing-request rule tithonus-000521

Both agents are at task 7 of their legal lifetimes, with 419 documents in the workspace. Tool calls are copied from the run logs with archive paths shortened. Notes after an arrow are recomputed on the released documents.

Message from the harness

```
Current date:  2018-01-01 (Mon)
Question:  As of January 1, 2018, counsel for a List I chemical
applicant or registrant is relying on DEA-403 when deciding how
to contest an order to show cause.  Is that reliance sound?  Give
a source-audited implementation assessment that reconciles the
action's rulemaking history with the operative CFR provisions in
parts 1309 and 1316.
```

In the archive

30 Dec 2016 The final rule of DEA-403 points §1309.46(d) to §1309.53 in place of §1309.54, "Burden of Proof", and readdresses the hearing-request form of §1316.47.

1 Jan 2017 eCFR snapshots of parts 1309 and 1316. §1309.46(d) still reads "pursuant to §1309.54", with a link to the amendment, and §1316.47 is addressed to the "DEA Federal Register Representative".

30 Jan 2017 A notice delays the rule's effective date to 21 Mar 2017.

1 Jan 2018 New snapshots of both parts: "pursuant to §1309.53" and "Hearing Clerk/OALJ".

GLM-5.3: five tool calls, and an answer that meets 9 of 11 criteria

1. `find 2017/**/*` over the archive
2. `grep "DEA-403|1117-AB41"` in `archive/2017`
3. `grep "Drug Enforcement Administration"` in `archive/2017`
4. `grep "1309\.5[13457]|1309\.46"` in the part 1309 snapshot of 1 Jan 2017
5. `read` the part 1316 snapshot of 1 Jan 2017

Answer. "Yes — reliance on DEA-403 is sound. … But two audit findings qualify that advice: the CFR snapshot in this record is stale (it predates effectiveness), and one residual textual tension in §1309.53 should be noted."

Judge. ✗ Citation 4: "The cited eCFR documents are the January 1, 2017 pre-amendment snapshots, already superseded by the ask date." ✗ Correctness 5, a leftover reference in §1309.53(d) that the answer does not name. The other nine criteria are met.

Grok 4.6: 15 tool calls, and an answer that meets all 11 criteria

2. `grep "DEA-403|1117-AB41"` over the archive
3. `find **/*1309*` and `find **/*1316*` over the archive (calls 3–4) → the snapshots of 1 Jan 2017 and 1 Jan 2018 among them
5. `grep "1309\.46|1309\.53|1309\.54|1316\.47|Hearing Clerk|Federal Register Representative"` in the part 1309 snapshot of 1 Jan 2018, and a like search in the part 1316 snapshot (call 6)
8. `grep "delay of effective date|March 21"` over the archive
9. `read` both 2018 snapshots (calls 9–10)

Answer. "Yes. Reliance on DEA-403 is sound as of January 1, 2018. … The January 1, 2018 eCFR is the implementation check: it codifies the final-rule text."

Judge. ✓ All 11 criteria.

Figure G.5: The version in force can carry the question date itself. On 1 January 2018 GLM-5.3 searches only the 2017 folder, reads the eCFR snapshot of 1 January 2017, and even notes that it predates the rule. The snapshot of 1 January 2018, which codifies the rule, is in the 2018 folder. Grok 4.6 searches the whole archive by part number, finds both snapshots, and cites the new one. Both answers reach the same conclusion, and only the second cites the text in force.

Two agents on the CISWI rulemaking tithonus-000889

Both agents are at task 56 of their legal lifetimes, with 1,865 documents in the workspace. Tool calls are copied from the run logs with archive paths shortened. Notes after an arrow are recomputed on the released documents.

Message from the harness

```
Current date:  2026-03-31 (Tue)
Question:  As of March 31, 2026, air-program counsel is assessing
the linked CISWI standards, emission-guideline, and Federal-plan
regimes associated with EPA-HQ-OAR-2003-0119.  Reconstruct the
procedurally distinct path from EPA's 2016 reconsideration
through its 2026 withdrawal of the disaster-recovery treatment.
Using the operative effective dates and an end-exclusive
calendar-day convention, quantify how long each completed governing
configuration lasted for the new-source, emission-guideline, and
Federal-plan components, separately quantify the pendency of the
intervening proposal, and explain where the components' clocks
diverged or converged.
```

In the archive: the docket's Federal Register documents

23 Jun 2016 The final rule on reconsideration. The emission guidelines take effect that day and the new-source standards on 23 Dec 2016.
15 Jun 2018 Technical amendments, headed "Proposed Rules", "ACTION: Proposed rule."
16 Apr 2019 The same amendments as a final rule, "effective on April 16, 2019."
2024–2026 The Federal plan (11 Dec 2024), an interim final rule for disaster recovery (26 Aug 2025), and its rescission (19 Mar 2026).

GLM-5.3: 32 tool calls, and an answer that meets 30 of 39 criteria

1. `grep "CISWI|Commercial and Industrial Solid Waste Incineration"` over the archive → 5,303 lines in 78 documents
4. `grep "CISWI|disaster"` in `archive/2022`, then in each year to 2026 (calls 4–8)
9. 24 more calls list, search, or read files of 2016 and of 2024 to 2026. No call opens a document of 2018 or 2019.

Answer. "Refinements within this baseline followed — e.g., the June 15, 2018 final amendments to subpart CCCC, 83 FR 28068, and the April 16, 2019 EG amendments, 84 FR 15846 — without altering the regime's structure." It lets the 2016 configurations run to 26 Aug 2025, 3,351 days for the guidelines and 3,168 for the standards.
Judge. ✗ Correctness 4: "mischaracterizes the June 15, 2018 action (83 FR 28068) as 'final amendments to subpart CCCC' rather than a proposed rule requesting comments." ✗ Eight more criteria, among them every duration that ends on 16 Apr 2019.

GPT-5.6 Sol: 27 tool calls, and an answer that meets all 39 criteria

1. `grep "EPA-HQ-OAR-2003-0119"` over the archive
8. `read` the docket's documents of 23 Jun 2016, 11 Jan 2017, 15 Jun 2018, 16 Apr 2019, 11 Dec 2024, 26 Aug 2025, and 19 Mar 2026, in that order (calls 8–14)
18. `grep "June 15, 2018"` in the final rule of 16 Apr 2019

Answer. "EPA also proposed technical amendments on June 15, 2018, finalized with an effective date of April 16, 2019." Its 2016 configurations end there, after 1,027 days for the guidelines and 844 for the standards, and the proposal was pending 305 days.
Judge. ✓ All 39 criteria.

Figure G.6: A proposal changes nothing until a final rule takes effect. GLM-5.3 never opens the docket's documents of 2018 or 2019, calls the June 2018 proposal final amendments, and lets the 2016 configurations run until August 2025. GPT-5.6 Sol reads the docket's documents in date order and ends the 2016 configurations at the April 2019 final rule.

H PROMPT TEMPLATES

Each domain has its own set of construction templates for the five steps whose rules depend on the record, namely generation, verification, rubric writing, rubric audit, and the answer-leak check. Figures H.1–H.5 give the finance set, Figures H.6–H.10 the medical set, and Figures H.11–H.15 the legal set. The steps are the same in every domain, and what a worker must check about a company filing, a drug label, or a federal rule differs. In the pipeline one template per step held the shared rules and the rules of all three domains, and each figure here shows the shared rules with one domain's rules, its record, and a released example. The other construction steps use one template in every domain (Figures H.16–H.20), followed by the evaluated agent and the judge (Figures H.21 and H.22). The construction templates are condensed and reworded for human readers, with output formats abbreviated. The agent's rules are verbatim, and the judge's template is reworded without changing a rule. Every worker runs in a fresh session and may open only the files its task file lists (Figure H.16). Deduplication and the single-source check are mechanical and have no template.

Finance: generation

Sees

Two workers run in turn over the finance record, the EDGAR filings of 50 public companies: every 8-K, 10-Q, and 10-K report and amendment they filed from January 2016 to May 2026. The first sees only each company's filing index of form types and dates. The second, given one group of companies, sees the documents its task file lists, and short summaries used only to find them. Neither sees existing tasks, other workers' outputs, or any file outside its list.

Does, first worker: choose related companies

1. Find a real professional relationship, and list every company it needs and no decorative ones. A single company is allowed.
2. Name the kinds of dated states that could matter, and say why the latest state alone might not answer a question about them.
3. Propose zero or more groups, and never weaken a group to meet a quota.

Finance rules

The relationship must be real: comparability, a counterparty, peer reporting, a transaction, an accounting policy, capital structure, or risk. A shared industry code or keyword is not enough. Follow a company through renames, acquisitions, changes of reporting perimeter, and subsidiaries, and flag periods, units, denominators, stocks and flows, or accounting bases that may not compare.

Does, second worker: write a candidate task for one group

1. Read each company's filings in publication order.
2. Find a finite question whose answer must relate several dated states rather than report the latest one, and drop any company the answer does not need.
3. Choose a month-end question date by which every document the answer needs is published, so that nothing later shapes the answer.
4. Write the complete answer first, with its evidence: for each document, its publication date, a quoted span, and what the span supports.
5. Only then write the question, in natural words. It names the companies and the period, but not the files, sections, answer values, or direction of a change.
6. Test the candidate against four shortcuts: one document, one state, the latest state, and a later document that summarizes the history. Report the results honestly, and never approve your own task.

Finance rules

Align fiscal periods, quarter and year-to-date figures, stocks and flows, original, amended, and restated values, units, denominators, and table definitions. Keep a transaction's announcement, signing, closing, financing, and later accounting apart, and compare reporting perimeters explicitly when acquisitions, divestitures, subsidiaries, or minority interests change them. Name each company as it was named at the question date, and hide an old-to-new name mapping when continuity is part of the answer.

Returns

The first worker returns `PROPOSE`, with each group's companies, their relationship, the professional use, candidate official sources, the kinds of dated states, and why the latest state is not enough, or `NO_GROUP`. The second returns `CANDIDATE`, in the format of Figure H.16, or `NO_QUESTION`.

Figure H.1: Finance, generation. Two workers turn company filings into a candidate task.

Finance: verification

Sees

The candidate's question, question date, companies, answer, and evidence, and every filing of those companies published by the question date. It does not see other tasks, later documents, or other workers' outputs.

Does

Rebuilds the answer from the documents, treating the generator's claims as proposals.

1. Check every entity, name change, value, date, period, unit, sign, population, and legal status against the original text.
2. Find every piece of evidence the answer needs, and every other official document visible at the question date that gives the same support.
3. Check that at least two dated states are needed: no single document, single publication date, later summary, or latest state gives the answer. For a series of updates, check each item and that the series contains a real change.
4. Check that every named entity is needed.
5. Check that answering takes professional work: comparing, reconciling, calculating, choosing the version in force, applying a condition, attributing a cause, following an entity through a change, or resolving a conflict.
6. Name every defect precisely, and keep no unsupported claim.

Finance rules

Recompute every period alignment, unit conversion, denominator, and calculation. Keep original, amended, and restated values and each stage of a transaction apart. Check names and reporting perimeters at the question date, and reject decorative peers and unqualified claims of an unchanged perimeter. For a table, check entity, period, row, column, unit, and aggregation.

Returns

`AGREE`, `DISAGREE`, or `UNSURE` when the record cannot settle a point, with the verified answer, the documents that support each part of it, and a verdict on every claim, entity, dated state, and shortcut.

Figure H.2: Finance, verification. A different worker rebuilds the answer from the filings visible at the question date.

Finance: rubric writing

Sees

The question, the question date, the verified answer, and the evidence with its quoted spans, but not the source documents. It does not change the question, the answer, or the evidence.

Does

Splits the verified answer into binary criteria of three types.

- *Correctness* (A1, A2, . . .): one criterion per fact the question asks for, keeping entity, value, date, period, unit, population, and legal status. No reasoning, and no detail the question did not ask for.
- *Citation* (E1, E2, . . .): one per piece of evidence the answer needs, stating what an official document must support and which criteria it backs. Any official document giving the same support counts, and the quoted span is kept for auditing, not as required wording.
- *Reasoning* (R1, R2, . . .): at least one finding the answer must establish on the way to its results. Each names a plausible error that leaves every correctness criterion right and makes it fail. Putting dates in order, stating the direction of a change, or restating a correctness criterion does not count.

Finance rules

Correctness criteria keep issuer, period, unit, denominator, stock or flow, transaction stage, and restatement status. Each citation criterion names its entity, period, measure, table, and perimeter, and none merely combines others. Normalizations, calculations, reconciliations, perimeter comparisons, and transaction stages go in reasoning criteria.

Returns

`COMPLETE` with the criteria, in the released format of Figure H.16, or `DEFECT_FOUND` naming the upstream problem.

Example, a released reasoning criterion (The Joint Corp.'s accounting-error notice, Figure I.1)

Criterion. "The answer establishes that the initial notice's planned revisions to the 2021 comparison-period financial statements were not already a formal material-error, restatement, and non-reliance conclusion, and that the September 15, 2023 update is what newly brought the 2021 annual statements, the March 31, 2022 interim period, and the March 31, June 30, and September 30, 2021 interim periods into that formal scope."
Error it catches. "All A results can be correct while the answer treats the initial notice's planned 2021 comparison-period revisions as already a restatement and non-reliance conclusion, so it does not establish that the later update broadened formal scope."

Figure H.3: Finance, rubric writing. A worker that sees the evidence spans, not the filings, writes the three types of criteria.

Finance: rubric audit

Sees

The question, the verified answer and evidence, the criteria, and every filing of those companies visible at the question date. It does not edit the criteria. Two further checks belong to the rubric audit (Figures H.17 and H.18).

Checks

1. The correctness criteria cover the question exactly, and each is one fact, correct, with the right legal status.
2. No correctness criterion carries reasoning that belongs in a reasoning criterion.
3. The citation criteria cover every piece of evidence needed and accept any equivalent official document, never one exact file or quotation.
4. Each reasoning criterion follows from the question, repeats no other criterion, can be seen in an answer, and fails on its named error.
5. At least one reasoning criterion passes.
6. The whole rubric keeps the question's time limit, entities, scope, and difficulty, and adds nothing the question did not ask.

Finance rules

Challenge every period, unit, denominator, transaction stage, restatement, and perimeter. Reject citation criteria that combine others, or that tie scoring to one filing or table when equivalent official support exists. Check that normalization, arithmetic, and comparability sit in reasoning criteria that can fail while the correctness criteria hold.

Returns

For each criterion `PASS`, `REDUNDANT`, `NOT_ASKED`, `FACTUALLY_DEFECTIVE`, `INCOMPLETE`, `NOT_JUDGEABLE`, or `UNSURE`. For the task `CLEAN`, `DEFECTIVE`, or `UNSURE`, naming where a defect lies: the question, the answer and evidence, or one criterion type. `CLEAN` needs every criterion to pass and at least one valid reasoning criterion.

Figure H.4: Finance, rubric audit. A different worker checks every criterion against the filings.

Finance: answer-leak check

Sees

The question, its entities and question date, and a one-line scope. It does not see the answer, evidence, rubric, other questions, or earlier verdicts. A method-leak check completes it (Figure H.19).

Checks

1. The entities, the professional scenario, and the needed period are clear and natural.
2. The question does not reveal answer values, directions, categories, old-to-new name mappings, rubric fields, or a checklist of the outputs expected.
3. A file, form, section, table, or snapshot is named only when it is the professional object, not a pointer to the answer.
4. Dropping such a pointer does not let several real events in the archive fit the question.
5. The task stays professionally hard and still needs several dated states. Length alone is never a defect.

Finance rules

Keep the company names in force at the question date, and a former name only when continuity is the task. Fiscal years, deal names, and accounting terms may stay when they identify the task, but form items, table routes, and figures that carry the answer may not. Never reveal the direction, winner, bridge, denominator, or stage outcome of a comparison.

Returns

`PASS`; `REWRITE`, with the smallest natural change and why it asks the same thing, which a second reviewer must approve; `DROP`; or `UNSURE`.

Example, a released question

"How did Repligen Corporation's product-revenue composition change between 2016 and 2025, and which categories drove the largest share gains and losses?"
It names the company and the two years, and leaves the categories, their shares, and the direction of each change to the answer (Figure I.5).

Figure H.5: Finance, answer-leak check. A worker that sees only the question looks for wording that gives away its answer.

Medical: generation

Sees

Two workers run in turn over the medical record, 50 prescription drug applications: every label version DailyMed published for them from January 2016, and every safety labeling change the FDA recorded for them. The first sees only each application's index of label versions and labeling changes with their dates. The second, given one group of drug applications, sees the documents its task file lists, and short summaries used only to find them. Neither sees existing tasks, other workers' outputs, or any file outside its list.

Does, first worker: choose related drug applications

1. Find a real professional relationship, and list every drug application it needs and no decorative ones. A single drug application is allowed.
2. Name the kinds of dated states that could matter, and say why the latest state alone might not answer a question about them.
3. Propose zero or more groups, and never weaken a group to meet a quota.

Medical rules

The relationship must be real: a shared ingredient, indication, interaction, population, or regulatory action, or a question that compares labels. A shared drug class is not enough. Keep product, ingredient, application, route, strength, and label apart, and never assume that one product's warning, dose, approval path, or implementation applies to another.

Does, second worker: write a candidate task for one group

1. Read each application's label versions and labeling changes in publication order.
2. Find a finite question whose answer must relate several dated states rather than report the latest one, and drop any drug application the answer does not need.
3. Choose a month-end question date by which every document the answer needs is published, so that nothing later shapes the answer.
4. Write the complete answer first, with its evidence: for each document, its publication date, a quoted span, and what the span supports.
5. Only then write the question, in natural words. It names the drug applications and the period, but not the files, sections, answer values, or direction of a change.
6. Test the candidate against four shortcuts: one document, one state, the latest state, and a later document that summarizes the history. Report the results honestly, and never approve your own task.

Medical rules

Align product, ingredient, application, route, strength, population, indication, and label version. Keep the date FDA announced, acted, or approved apart from the date the label text was published and put into effect. Compare like sections and scopes, and do not invent a change by matching a broad early statement to later detailed rows. State what the label says, never clinical advice.

Returns

The first worker returns `PROPOSE`, with each group's drug applications, their relationship, the professional use, candidate official sources, the kinds of dated states, and why the latest state is not enough, or `NO_GROUP`. The second returns `CANDIDATE`, in the format of Figure H.16, or `NO_QUESTION`.

Figure H.6: Medical, generation. Two workers turn drug labels and FDA labeling changes into a candidate task.

Medical: verification

Sees

The candidate's question, question date, drug applications, answer, and evidence, and every label version and labeling change of those applications published by the question date. It does not see other tasks, later documents, or other workers' outputs.

Does

Rebuilds the answer from the documents, treating the generator's claims as proposals.

1. Check every entity, name change, value, date, period, unit, sign, population, and legal status against the original text.
2. Find every piece of evidence the answer needs, and every other official document visible at the question date that gives the same support.
3. Check that at least two dated states are needed: no single document, single publication date, later summary, or latest state gives the answer. For a series of updates, check each item and that the series contains a real change.
4. Check that every named entity is needed.
5. Check that answering takes professional work: comparing, reconciling, calculating, choosing the version in force, applying a condition, attributing a cause, following an entity through a change, or resolving a conflict.
6. Name every defect precisely, and keep no unsupported claim.

Medical rules

Check product, ingredient, application, route, strength, population, indication, section, and label version for every claim. Keep regulatory action and approval apart from label publication and implementation. Reject any transfer of a warning, dose, contraindication, or population across products, and check that a comparison uses equivalent sections and does not turn label text into clinical advice.

Returns

`AGREE`, `DISAGREE`, or `UNSURE` when the record cannot settle a point, with the verified answer, the documents that support each part of it, and a verdict on every claim, entity, dated state, and shortcut.

Figure H.7: Medical, verification. A different worker rebuilds the answer from the labels visible at the question date.

Medical: rubric writing

Sees

The question, the question date, the verified answer, and the evidence with its quoted spans, but not the source documents. It does not change the question, the answer, or the evidence.

Does

Splits the verified answer into binary criteria of three types.

- *Correctness* (A1, A2, . . .): one criterion per fact the question asks for, keeping entity, value, date, period, unit, population, and legal status. No reasoning, and no detail the question did not ask for.
- *Citation* (E1, E2, . . .): one per piece of evidence the answer needs, stating what an official document must support and which criteria it backs. Any official document giving the same support counts, and the quoted span is kept for auditing, not as required wording.
- *Reasoning* (R1, R2, . . .): at least one finding the answer must establish on the way to its results. Each names a plausible error that leaves every correctness criterion right and makes it fail. Putting dates in order, stating the direction of a change, or restating a correctness criterion does not count.

Medical rules

Correctness criteria keep product, ingredient, application, route, strength, population, indication, section, and version. Citation criteria keep evidence of a regulatory action apart from evidence of label text when both matter. Reasoning criteria cover comparison across versions, which product a statement applies to, conditions, and the gap from action to label, and never score clinical advice.

Returns

`COMPLETE` with the criteria, in the released format of Figure H.16, or `DEFECT_FOUND` naming the upstream problem.

Example, a released reasoning criterion (DARZALEX infusion reactions, Figure I.6)

Criterion. "The answer establishes that the elements that changed between the initial and current infusion-reaction guidance are the Grade 3 restart gate, the addition of anaphylaxis as a permanent-discontinuation trigger with emergency care, and the addition of an ocular-symptom interruption and ophthalmologic-evaluation branch, and that the core any-grade interruption, half-rate restart, third Grade 3-or-greater discontinuation, and Grade 4 discontinuation framework continued."
Error it catches. "All A results can be correct while the answer attributes the evolution to a change in the Grade 1–2 half-rate restart rule or treats the ocular branch as pre-existing initial guidance."

Figure H.8: Medical, rubric writing. A worker that sees the evidence spans, not the labels, writes the three types of criteria.

Medical: rubric audit

Sees

The question, the verified answer and evidence, the criteria, and every label version and labeling change of those applications visible at the question date. It does not edit the criteria. Two further checks belong to the rubric audit (Figures H.17 and H.18).

Checks

1. The correctness criteria cover the question exactly, and each is one fact, correct, with the right legal status.
2. No correctness criterion carries reasoning that belongs in a reasoning criterion.
3. The citation criteria cover every piece of evidence needed and accept any equivalent official document, never one exact file or quotation.
4. Each reasoning criterion follows from the question, repeats no other criterion, can be seen in an answer, and fails on its named error.
5. At least one reasoning criterion passes.
6. The whole rubric keeps the question's time limit, entities, scope, and difficulty, and adds nothing the question did not ask.

Medical rules

Challenge product, application, route, strength, population, section, and version, and any statement carried across products. Reject criteria that merge FDA action, approval, label publication, and implementation into one fact. Check that reasoning criteria test comparison, application, or timing without adding clinical advice or a product state the question did not ask for.

Returns

For each criterion `PASS`, `REDUNDANT`, `NOT_ASKED`, `FACTUALLY_DEFECTIVE`, `INCOMPLETE`, `NOT_JUDGEABLE`, or `UNSURE`. For the task `CLEAN`, `DEFECTIVE`, or `UNSURE`, naming where a defect lies: the question, the answer and evidence, or one criterion type. `CLEAN` needs every criterion to pass and at least one valid reasoning criterion.

Figure H.9: Medical, rubric audit. A different worker checks every criterion against the labels.

Medical: answer-leak check

Sees

The question, its entities and question date, and a one-line scope. It does not see the answer, evidence, rubric, other questions, or earlier verdicts. A method-leak check completes it (Figure H.19).

Checks

1. The entities, the professional scenario, and the needed period are clear and natural.
2. The question does not reveal answer values, directions, categories, old-to-new name mappings, rubric fields, or a checklist of the outputs expected.
3. A file, form, section, table, or snapshot is named only when it is the professional object, not a pointer to the answer.
4. Dropping such a pointer does not let several real events in the archive fit the question.
5. The task stays professionally hard and still needs several dated states. Length alone is never a defect.

Medical rules

Keep the product or application identifier needed to make the question unique, but drop label-section or snapshot routes that point to the answer. Keep product, route, population, and timing distinctions without listing the expected warning, dose, or category changes. Name FDA action and label implementation separately only when the gap between them is the task.

Returns

`PASS`; `REWRITE`, with the smallest natural change and why it asks the same thing, which a second reviewer must approve; `DROP`; or `UNSURE`.

Example, a released question

"How did AUBAGIO's pregnancy-risk labeling evolve from 2016 through 2019?"
It names the product and the period, and says neither which sections changed nor whether any control loosened (Figure I.7).

Figure H.10: Medical, answer-leak check. A worker that sees only the question looks for wording that gives away its answer.

Legal: generation

Sees

Two workers run in turn over the legal record, 50 parts of CFR Titles 21 and 40: every proposed and final rule the Federal Register lists for them, and a yearly eCFR snapshot of each part. The first sees only each part's index of rules and snapshots with their dates. The second, given one group of CFR parts, sees the documents its task file lists, and short summaries used only to find them. Neither sees existing tasks, other workers' outputs, or any file outside its list.

Does, first worker: choose related CFR parts

1. Find a real professional relationship, and list every CFR part it needs and no decorative ones. A single CFR part is allowed.
2. Name the kinds of dated states that could matter, and say why the latest state alone might not answer a question about them.
3. Propose zero or more groups, and never weaken a group to meet a quota.

Legal rules

The relationship must be real: a governing, procedural, cross-reference, or compliance dependency between rules, provisions, proceedings, or regulated parties. Keep jurisdiction, issuing agency, CFR title, part, and section, and regulated scope explicit, and flag differences between proposed, final, effective, compliance, and codified text.

Does, second worker: write a candidate task for one group

1. Read each part's rules and snapshots in publication order.
2. Find a finite question whose answer must relate several dated states rather than report the latest one, and drop any CFR part the answer does not need.
3. Choose a month-end question date by which every document the answer needs is published, so that nothing later shapes the answer.
4. Write the complete answer first, with its evidence: for each document, its publication date, a quoted span, and what the span supports.
5. Only then write the question, in natural words. It names the CFR parts and the period, but not the files, sections, answer values, or direction of a change.
6. Test the candidate against four shortcuts: one document, one state, the latest state, and a later document that summarizes the history. Report the results honestly, and never approve your own task.

Legal rules

Align jurisdiction, agency, CFR target, regulated party, and proceeding. Keep proposed, final, corrected, delayed, stayed, effective, compliance, and codified states and their dates apart. Reconcile amending instructions, operative text, and later codification, and never let a preamble's shorthand stand in for controlling text. Forms, filing duties, and CFR provisions are valid task objects when they are substantive law.

Returns

The first worker returns `PROPOSE`, with each group's CFR parts, their relationship, the professional use, candidate official sources, the kinds of dated states, and why the latest state is not enough, or `NO_GROUP`. The second returns `CANDIDATE`, in the format of Figure H.16, or `NO_QUESTION`.

Figure H.11: Legal, generation. Two workers turn Federal Register rules and eCFR snapshots into a candidate task.

Legal: verification

Sees

The candidate's question, question date, CFR parts, answer, and evidence, and every rule and snapshot of those parts published by the question date. It does not see other tasks, later documents, or other workers' outputs.

Does

Rebuilds the answer from the documents, treating the generator's claims as proposals.

1. Check every entity, name change, value, date, period, unit, sign, population, and legal status against the original text.
2. Find every piece of evidence the answer needs, and every other official document visible at the question date that gives the same support.
3. Check that at least two dated states are needed: no single document, single publication date, later summary, or latest state gives the answer. For a series of updates, check each item and that the series contains a real change.
4. Check that every named entity is needed.
5. Check that answering takes professional work: comparing, reconciling, calculating, choosing the version in force, applying a condition, attributing a cause, following an entity through a change, or resolving a conflict.
6. Name every defect precisely, and keep no unsupported claim.

Legal rules

Check jurisdiction, agency, CFR target, provision, and regulated party. Keep publication, proposal, final rule, correction, effective date, compliance date, stay or delay, and codification apart. Score operative text, treating preamble summaries as context, check the interval in which a text governed, and never treat text published but not yet effective as in force.

Returns

`AGREE`, `DISAGREE`, or `UNSURE` when the record cannot settle a point, with the verified answer, the documents that support each part of it, and a verdict on every claim, entity, dated state, and shortcut.

Figure H.12: Legal, verification. A different worker rebuilds the answer from the rules and snapshots visible at the question date.

Legal: rubric writing

Sees

The question, the question date, the verified answer, and the evidence with its quoted spans, but not the source documents. It does not change the question, the answer, or the evidence.

Does

Splits the verified answer into binary criteria of three types.

- *Correctness* (A1, A2, . . .): one criterion per fact the question asks for, keeping entity, value, date, period, unit, population, and legal status. No reasoning, and no detail the question did not ask for.
- *Citation* (E1, E2, . . .): one per piece of evidence the answer needs, stating what an official document must support and which criteria it backs. Any official document giving the same support counts, and the quoted span is kept for auditing, not as required wording.
- *Reasoning* (R1, R2, . . .): at least one finding the answer must establish on the way to its results. Each names a plausible error that leaves every correctness criterion right and makes it fail. Putting dates in order, stating the direction of a change, or restating a correctness criterion does not count.

Legal rules

Correctness criteria keep jurisdiction, agency, provision, regulated party, and the exact legal status and date. Citation criteria keep proposed, final, amending, effective, compliance, and codified text apart when each contributes. Reasoning criteria cover which rule governed when, supersession, conditional procedure, cross-references, and conflicts, without repeating the end states.

Returns

`COMPLETE` with the criteria, in the released format of Figure H.16, or `DEFECT_FOUND` naming the upstream problem.

Example, a released reasoning criterion (waters of the United States in 40 CFR 232, Figure I.12)

Criterion. "The answer establishes that the February 6, 2020 date stated in the 2019 snapshot is an applicability date, and that the 2018–2021 snapshot dates are locators of text states rather than dates from which operative status of every clause is inferred."
Error it catches. "All A results can be correct while the answer infers operative status of the 2015-rule clauses solely from snapshot publication dates, or treats February 6, 2020 as a publication date rather than an applicability date."

Figure H.13: Legal, rubric writing. A worker that sees the evidence spans, not the rules, writes the three types of criteria.

Legal: rubric audit

Sees

The question, the verified answer and evidence, the criteria, and every rule and snapshot of those parts visible at the question date. It does not edit the criteria. Two further checks belong to the rubric audit (Figures H.17 and H.18).

Checks

1. The correctness criteria cover the question exactly, and each is one fact, correct, with the right legal status.
2. No correctness criterion carries reasoning that belongs in a reasoning criterion.
3. The citation criteria cover every piece of evidence needed and accept any equivalent official document, never one exact file or quotation.
4. Each reasoning criterion follows from the question, repeats no other criterion, can be seen in an answer, and fails on its named error.
5. At least one reasoning criterion passes.
6. The whole rubric keeps the question's time limit, entities, scope, and difficulty, and adds nothing the question did not ask.

Legal rules

Challenge jurisdiction, agency, provision, and every proposed, final, corrected, delayed, stayed, effective, compliance, or codified distinction. Reject preamble-only support for a criterion about operative text, and ties to one source where equivalent controlling text exists. Check that governing interval, supersession, procedure, and conflicts sit in reasoning criteria rather than being settled by correctness criteria.

Returns

For each criterion `PASS`, `REDUNDANT`, `NOT_ASKED`, `FACTUALLY_DEFECTIVE`, `INCOMPLETE`, `NOT_JUDGEABLE`, or `UNSURE`. For the task `CLEAN`, `DEFECTIVE`, or `UNSURE`, naming where a defect lies: the question, the answer and evidence, or one criterion type. `CLEAN` needs every criterion to pass and at least one valid reasoning criterion.

Figure H.14: Legal, rubric audit. A different worker checks every criterion against the rules and snapshots.

Legal: answer-leak check

Sees

The question, its entities and question date, and a one-line scope. It does not see the answer, evidence, rubric, other questions, or earlier verdicts. A method-leak check completes it (Figure H.19).

Checks

1. The entities, the professional scenario, and the needed period are clear and natural.
2. The question does not reveal answer values, directions, categories, old-to-new name mappings, rubric fields, or a checklist of the outputs expected.
3. A file, form, section, table, or snapshot is named only when it is the professional object, not a pointer to the answer.
4. Dropping such a pointer does not let several real events in the archive fit the question.
5. The task stays professionally hard and still needs several dated states. Length alone is never a defect.

Legal rules

A CFR title, part, or section, a proceeding, or a filing duty may stay when it is the legal object rather than a locator. Keep jurisdiction and the needed period without revealing whether a rule was final, delayed, effective, or controlling when that is the answer, and never turn the requested analysis into a checklist copied from the rubric.

Returns

`PASS`; `REWRITE`, with the smallest natural change and why it asks the same thing, which a second reviewer must approve; `DROP`; or `UNSURE`.

Example, a released question

"For 21 CFR Part 145 (Canned Fruits), how did FDA's July 2025 direct-final process determine whether the Part 145 revocations would take effect?"
It names the CFR part and the proceeding, which are the legal object, and does not say whether the revocations took effect (Figure I.15).

Figure H.15: Legal, answer-leak check. A worker that sees only the question looks for wording that gives away its answer.

Task file and output formats

One real task file, abbreviated

```
{"task_id": "lens-001",
 "domain": "finance",
 "group_id": "finance-upstream-oil-gas-producers",
 "entity_ids": ["315852", "717423", "1309082", "1539838"],
 "generation_lens": "risk_control_response",
 "per_task_cap": 6, "per_task_minimum": 0,
 "may_read_existing_qa": false,
 "official_read_allowlist": [
   "qa_entity_source_views_v2/finance/sec_1309082/manifest.jsonl",
   "qa_entity_source_views_v2/finance/sec_1539838/manifest.jsonl",
   "... 919 more files"],
 "entity_summaries": [
   "qa_entity_source_views_v2/finance/sec_1309082/_summary.json",
   "qa_entity_source_views_v2/finance/sec_1539838/_summary.json",
   "... 2 more files"],
 "literal_message_allowlist": ["this task file", "the generation template", "... the
     files above"],
 "output_path": "outputs/lens-001.json"}
```

What the fields mean

- `entity_ids`: the group, here four oil and gas producers, and `generation_lens`: an optional angle for the questions.
- `per_task_cap`: at most six candidates, and `may_read_existing_qa`: no earlier task may be read.
- `official_read_allowlist`: every file the worker may open, 921 here. `literal_message_allowlist`: everything it is shown. Reads are checked against both lists after the run.

The task file carries the parameters, and the template carries the rules.

Output formats, abbreviated

A candidate from generation, and criteria from rubric writing in the released format:

```
{"question_date": "YYYY-MM-DD", "question": "...", "answer": "...",
 "evidence": [{"document": "...", "published": "YYYY-MM-DD",
               "quote": "...", "supports": "..."}],
 "dated_states": [{"entity": "...", "date": "YYYY-MM-DD", "role": "..."}],
 "shortcut_tests": {"one_document": "PASS", "one_state": "PASS",
                    "latest_state": "PASS", "later_summary": "PASS"}}

[{"id": "A1", "type": "A", "criterion": "..."},
 {"id": "E1", "type": "E", "criterion": "...", "supports": ["A1", "R1"]},
 {"id": "R1", "type": "R", "criterion": "...",
  "counterfactual_error": "All A results can be correct while ..."}]
```

Figure H.16: Task file and output formats. Every worker receives one template and one task file like this, and nothing else, and generation and rubric writing return these formats.

Rubric audit: criterion-type check

Sees
The final question and the criteria, and nothing else: not the rubric writer's notes, the documents, other tasks, or any answer.

Checks
Given only the question, is each fact scored as the right type? Where this check and the audit against the record disagree, the question decides.

1. Every output the question asks for has a correctness criterion. An output without one is the gap this check exists to catch.
2. No correctness criterion scores a value the question itself supplies.
3. No reasoning criterion restates a correctness criterion, which would score one fact twice.
4. A reasoning criterion states a finding, not how the answer must be presented.
5. For each reasoning criterion, state an error that leaves every correctness criterion right and makes it fail. If there is none, the criterion carries no weight.
6. Invent nothing. Content missing from the question, answer, and rubric goes to source review.

Returns
`CLEAN`, `WRONG_TYPE`, or `NEEDS_SOURCE_REVIEW`, listing the question outputs without a correctness criterion, the correctness criteria that score a given value, the reasoning criteria that repeat one or cannot fail, and a reason in at most two sentences.

Figure H.17: Criterion-type check. Part of the rubric audit, judged against the question alone.

Rubric audit: evidence-coverage check

Sees
The final question, the criteria, and every evidence entry with its quoted span, but not the documents.

Checks
Earlier steps check that the recorded evidence is correct, and this one checks that it is enough. A quotation can be true, official, and cited, and still stop one sentence before the number a criterion needs.

1. For each correctness criterion, name the entry whose span establishes it.
2. Find any fact the question needs, such as an entity, a period, a comparison point, or the governing version, that no entry covers.
3. Report each correctness criterion that rests on a single entry. That alone is not a defect.

Never assume a document holds more than its recorded span shows.

Returns
`SUFFICIENT`, `THIN_BUT_SUFFICIENT`, or `INSUFFICIENT`, with the unsupported criteria, the missing facts, and a reason in at most two sentences.

Figure H.18: Evidence-coverage check. Part of the rubric audit, reading only the recorded spans.

Answer-leak check: method-leak check

Sees

The question and a sample of accepted questions from an earlier, independently reviewed set, and nothing else.

Checks

The wording check judges a question against itself, but a batch drifting toward telling the agent what to do shows only by comparison.

1. Compared with the sample, does the question give away more of the method: the inputs to use, the operations, their order, or the fields to report?
2. Does it list the values to work from? That turns retrieval into copying, and pushes the rubric to score those values instead of the result.
3. Would a professional given only the scenario and the date arrive at the same task without the procedural text?
4. Is the requested output still clear? Under-specification is a defect too.

Returns

`APPROPRIATE`, `PATH_GIVEN_AWAY`, `TOO_EASY`, `UNDER_SPECIFIED`, or `HARD`; a method-leak score from 1 to 5, on which the sample averages 2; the quoted spans that give the method away; and a reason in at most two sentences.

Figure H.19: Method-leak check. Part of the answer-leak check, comparing the question with accepted ones.

Human review

Sees, for each task

The question, the answer, the evidence with its quoted spans, the official alternatives accepted for each piece of evidence, the criteria, and the rubric audit's verdict.

Decides

Can the question, the answer, the evidence, and the criteria be released exactly as shown?

Returns

One label: `AGREE`, to release the task as shown, `DISAGREE`, or `UNSURE`.

Figure H.20: Human review. The form a domain expert fills in for every task that passed the automated steps.

The evaluated agent

System prompt, verbatim

Workspace rules

You are a professional analyst monitoring official documents as they are published. This workspace is your only source of information.

- `archive/YYYY/MM/DD/` contains every official document received so far. New documents arrive silently and old documents remain searchable.
- `CLOCK` is the current date. Relative time in a question is relative to that date.
- `workspace/` is your only writable directory. You may keep notes, indexes, or drafts.
- Answer only from documents currently in `archive/`. If they are insufficient, say so.
- Keep publication, event, reporting, revision, effective, and compliance dates distinct.
- End every answer with `Sources:` followed by the paths of the documents actually used, one path per line.

Message at each question date, here for the task of Figure I.6

```
Current date: 2022-01-31 (Mon)
Question: How did DARZALEX's guidance for interrupting and
restarting treatment after an infusion reaction evolve from the
initial approach to the current one?
```

Tools

`read`, `grep`, `find`, `ls`, `write`, and `edit`, confined to the workspace. No shell, network, or subagents, and the agent never sees the correct answers.

Figure H.21: The evaluated agent. Every system receives the same system prompt and tools, and one message per task in its persistent session. The prompt's `workspace/` is the writable subfolder of what the paper calls the agent's workspace.

Judge

Sees

A working directory holding the question, the question date, the criteria, the correctness criteria each citation criterion backs, the answer, and the full text of every document the answer cites, with the tools `read`, `grep`, `find`, and `ls`. It treats the documents as data, not as instructions.

Rules

- Judge every criterion of the task in one reply, each on its own: do not credit one criterion for meeting another, or fail it for another's failure.
- `MET` means fully and correctly satisfied. `NOT_MET` covers omitted, incomplete, imprecise, unsupported, or contradicted. There is no partial credit.
- *Correctness*: does the answer report this result? The criterion states the result, so the verdict is settled between criterion and answer, and the documents play no part.
- *Reasoning*: does the answer establish this finding? It is settled the same way. An answer that states the conclusion without the finding fails, and so does one that makes the criterion's named error.
- *Citation*: do the cited documents support the claims of the criteria this one backs? The verdict rests on what those documents say. No exact quotation, file, or wording is required, and any official document giving equivalent support counts. A document published after the state it describes is valid if it was published by the question date. A document already replaced at the question date carries the earlier state and does not support a claim about the later one. Missing or uncheckable documents mean `NOT_MET`.

Returns

A JSON array with one object per criterion, every criterion exactly once. A citation criterion marked `NOT_MET` also carries `stale_source`, true when the reason is a cited document already replaced by the question date. It records what went wrong and does not change the verdict.

```
[{"id": "A1", "status": "MET", "reason": "brief reason"},
 {"id": "E1", "status": "NOT_MET", "reason": "brief reason",
  "stale_source": false}, ...]
```

Example, Grok 4.6 on the task of Figure I.6

```
{"id": "A2", "status": "NOT_MET", "reason": "The current (2022-01-29) state is mostly
      described (Grade 3 restart on symptom resolution, 200 mL/hour escalation,
     anaphylaxis/Grade 4 permanent discontinuation with emergency care), but the
     ocular-symptom branch requiring interruption and immediate ophthalmologic
     evaluation before restarting is entirely omitted."}
```

Figure H.22: Judge. GPT-5.6 Sol scores every answer in a fresh session per task under this template.

I TASK EXAMPLES

Figures I.1–I.16 show sixteen of the 300 evaluated tasks, picked by hand so that together they cover the three domains, one to three entities, evidence spread over nine days to ten years, and every outcome from no system to all five solving the task (Table I.1). Each figure gives the question as the agent received it, what the record shows at each dated state, what a full-credit answer needs, and how the five systems of Table 3 fared. The record and the rubric are paraphrased for readability, and the released files give them in full. Appendix G follows two agents call by call through the task of Figure I.6 and five others.

Table I.1: The sixteen tasks of Figures I.1–I.16. *Span* runs from the first to the last publication date among a task's evidence documents. Criteria are counted as correctness / reasoning / citation, and *Solved* counts the systems, of five, that met every criterion.

Fig.	Domain	Task	Question date	Entities	Docs, span	Criteria	Solved
I.1	Finance	An accounting-error notice widens (The Joint Corp.)	2023-09-30	1	2, 9 days	3 / 1 / 2	4
I.2	Finance	DOJ exposure from an acquisition (Allscripts)	2020-01-31	1	3, 2.1 yr	3 / 1 / 3	1
I.3	Finance	Insurance recovery after a plant fire (ONEOK)	2023-01-31	1	3, 154 days	3 / 1 / 3	4
I.4	Finance	Credit pressure across the switch to CECL (two banks)	2024-03-01	2	6, 4.0 yr	1 / 3 / 2	3
I.5	Finance	Product-revenue mix over nine years (Repligen)	2026-03-31	1	2, 9.0 yr	4 / 1 / 2	4
I.6	Medical	Restarting after an infusion reaction (DARZALEX)	2022-01-31	1	5, 5.6 yr	2 / 1 / 5	2
I.7	Medical	Pregnancy-risk labeling (AUBAGIO)	2019-08-31	1	3, 3.1 yr	3 / 1 / 3	1
I.8	Medical	Pregnancy as a contraindication (LIPITOR, CRESTOR)	2023-01-31	2	5, 6.6 yr	2 / 3 / 3	3
I.9	Medical	Antiphospholipid-syndrome restriction (XARELTO)	2026-04-30	1	3, 6.6 yr	4 / 1 / 2	3
I.10	Medical	Neurodevelopmental risk after exposure in utero (Depakote)	2026-04-30	1	4, 10.1 yr	1 / 1 / 4	0
I.11	Medical	A myasthenia gravis signal (CRESTOR, LIPITOR)	2024-04-02	2	2, 250 days	3 / 1 / 2	5
I.12	Legal	Waters of the United States (40 CFR 232)	2021-01-31	1	4, 3.0 yr	1 / 1 / 4	1
I.13	Legal	Over-the-counter hearing aids (21 CFR 800, 801, 874)	2022-08-31	3	2, 301 days	4 / 1 / 4	4
I.14	Legal	PFAS reporting (40 CFR 705)	2026-07-31	1	7, 5.1 yr	2 / 2 / 6	1
I.15	Legal	A direct final rule takes effect (21 CFR 145)	2025-09-30	1	3, 71 days	1 / 1 / 3	3
I.16	Legal	Aerosol cans become universal waste (40 CFR 262, 273)	2021-04-30	2	5, 3.0 yr	21 / 5 / 13	2

Finance: an accounting-error notice widens tithonus-000565

Question date 2023-09-30. Task 35 of 102 in the finance lifetime, with 7,387 documents in the workspace. Six criteria, three correctness, one reasoning, and two citation.

Question

How did The Joint Corp.'s 2023 accounting-error disclosure broaden from its initial notice to the later update, and what corrections and estimated 2022 earnings effect were identified at the initial stage?

What the record shows

6 Sep 2023 Form 8-K, Item 4.02. The 2022 annual statements and the June and September 2022 quarters contain material errors, must be restated, and should no longer be relied on. The 2021 comparison periods will be revised but are not declared unreliable. The cost of re-acquired regional developer rights is to be expensed rather than capitalized and amortized, regional developers' upfront fees are to offset cost of revenue rather than count as revenue, and transfer-pricing adjustments for the variable-interest entities are estimated. Estimated 2022 effect: net income about $0.6M lower ($0.9M after tax from the rights correction, less a $0.3M tax benefit from transfer pricing), with no change to Adjusted EBITDA or cash.

15 Sep 2023 Form 8-K/A. The 2021 annual statements, the three 2021 quarters, and the March 2022 quarter join the periods that must be restated and not relied on.

What a full-credit answer needs

- Both scopes, and exactly which periods the amendment added.
- The two corrections, the transfer-pricing adjustments, and the estimated 2022 effect line by line.
- *Reasoning.* The 2021 revisions announced on 6 September were only planned, so the formal widening came with the amendment of 15 September.
- *Citation.* The 8-K and the 8-K/A.

Five systems

GPT-5.6 Sol ✓ GLM-5.3 ✓ Grok 4.6 ✗ DeepSeek Flash ✓ Claude Sonnet 5 ✓

Grok 4.6 credits the widening to the 10-K/A filed on 26 September and never mentions the amendment of 15 September.

Figure I.1: An accounting-error notice widens. The answer has to tell revisions that were only planned from the formal non-reliance scope, which the amendment set nine days later.

Finance: regulatory exposure inherited through an acquisition tithonus-000320

Question date 2020-01-31. Task 10 of 102 in the finance lifetime, with 3,701 documents in the workspace. Seven criteria, three correctness, one reasoning, and three citation.

Question

How did Allscripts Healthcare Solutions, Inc.'s disclosed regulatory exposure from acquiring Practice Fusion evolve from the pre-close risk through the negotiated and final DOJ resolution?

What the record shows

8 Jan 2018 Form 8-K, deal pending. Allscripts agrees to buy Practice Fusion for $100M. The only regulatory matter disclosed is a March 2017 civil investigative demand from the U.S. Attorney in Vermont about the certification of Practice Fusion's health-record software. The risk is conditional and has no amount: if the deal closes and the demand leads to a claim, the losses could be material.

9 Aug 2019 Form 10-Q, deal closed in February 2018. Seven further DOJ requests (four demands and three HIPAA subpoenas) and a March 2019 grand-jury subpoena in a criminal Anti-Kickback Statute investigation. An agreement in principle of 6 August contemplates a $145.0M payment, a deferred prosecution agreement, and a civil settlement, still subject to negotiation and DOJ approval, and the amount is accrued.

28 Jan 2020 Form 8-K. The final resolution, signed on 27 January: a three-year deferred prosecution agreement on two felony counts with a $25.3M fine and a $959,700 forfeiture, a $118.6M civil settlement of which up to $5.2M goes to state Medicaid programs, three years of outside compliance oversight, and no Corporate Integrity Agreement.

What a full-credit answer needs

- The exposure at each of the three dates, with its amounts and conditions.
- *Reasoning.* The exposure changed in kind, from a conditional risk with no amount, to a quantified agreement that was not yet final, to signed obligations.
- *Citation.* The three filings.

Five systems

GPT-5.6 Sol ✗ GLM-5.3 ✗ Grok 4.6 ✓ DeepSeek Flash ✗ Claude Sonnet 5 ✗

Each of the other four drops a detail at one stage. GPT-5.6 Sol names the wrong certification program for the 2017 demand, GLM-5.3 and Claude Sonnet 5 leave out the conditions that kept the 2019 agreement unfinished, and DeepSeek Flash leaves out the conditional wording of the 2018 risk.

Figure I.2: Regulatory exposure inherited through an acquisition. Over three filings the exposure changes in kind, from a conditional risk to a signed settlement, and only Grok 4.6 reports every stage with its conditions.

Finance: insurance recovery after a plant fire tithonus-000338

Question date 2023-01-31. Task 24 of 102 in the finance lifetime, with 6,719 documents in the workspace. Seven criteria, three correctness, one reasoning, and three citation.

Question

How did ONEOK's handling of the Medford fractionator fire evolve from its initial insurance outlook through the final settlement, and what cash-recovery timing followed?

What the record shows

9 Aug 2022 Form 10-Q. A fire on 9 July has shut the 210,000 barrel-per-day Medford fractionator for an extended period. Insurance carries a $2B limit per occurrence, a $5M property deductible, and a 45-day waiting period for business interruption. The loss cannot yet be estimated and no material effect is expected, but the timing of proceeds may move quarterly results.

2 Nov 2022 Form 10-Q. Insurers agree to a $100M first installment. About $45M has arrived and been applied to insurance receivables, and the rest is expected within 30 days.

10 Jan 2023 Form 8-K. A full and final settlement on 9 January for $930M in total, including the $100M already received. The remaining $830M is due within 45 days.

What a full-credit answer needs

- The initial coverage and outlook, the $930M settlement, and the balance still owed at the question date.
- *Reasoning.* The November 2022 state in between: $100M agreed and about $45M received.
- *Citation.* Each of the three filings.

Five systems

GPT-5.6 Sol ✓ GLM-5.3 ✓ Grok 4.6 ✓ DeepSeek Flash ✓ Claude Sonnet 5 ✗

Claude Sonnet 5 dates the coverage terms to the November filing and calls the first disclosure merely qualitative.

Figure I.3: Insurance recovery after a plant fire. Three filings trace the claim from coverage terms to a $930M settlement, and the interim installment is the step the reasoning criterion checks.

Finance: credit pressure across the switch to CECL tithonus-000027

Question date 2024-03-01. Task 42 of 102 in the finance lifetime, with 7,759 documents in the workspace. Two entities. Six criteria, one correctness, three reasoning, and two citation.

Question

A credit committee comparing The Bancorp, Inc. and Customers Bancorp, Inc. wants to understand how reserve coverage and realized loss intensity changed from year-end 2019 to year-end 2023 without mistaking the intervening shift to CECL for pure credit movement. What joint conclusion should it draw about where credit pressure was manifesting at each issuer by 2023?

What the record shows

Mar 2020 10-Ks for 2019. The Bancorp's allowance is 0.56% of loans and its net charge-offs 0.12% of average loans. Customers Bancorp's are 0.77% and 0.08%.

Mar 2021 10-Ks for 2020. Both banks adopted CECL on 1 January 2020, moving from incurred-loss to lifetime expected-loss reserves.

29 Feb 2024 10-Ks for 2023. The Bancorp is at 0.51% and 0.07%, with the 2023 rise in charge-offs coming mainly from leasing. Customers Bancorp is at 1.13% and 0.48%, mainly from consumer installment loans.

What a full-credit answer needs

- The conclusion: by 2023 realized losses were clearly higher at Customers Bancorp and not at The Bancorp, and each bank's own charge-off path and its drivers are a safer signal than comparing reserve ratios across the two banks.
- *Reasoning.* Each bank's path with its driver, and that CECL changed reserving in 2020 and the two banks' ratios use different denominators.
- *Citation.* Each bank's filings.

Five systems

GPT-5.6 Sol ✓ GLM-5.3 ✗ Grok 4.6 ✓ DeepSeek Flash ✓ Claude Sonnet 5 ✗

GLM-5.3 and Claude Sonnet 5 give Customers Bancorp's 2023 charge-offs as about 0.55% instead of the 48 basis points its 10-K reports. Claude Sonnet 5 also blurs The Bancorp's leasing driver and never says that the denominators differ.

Figure I.4: Credit pressure across the switch to CECL. Reserve ratios cannot be compared directly across an accounting change, so the answer has to lean on each bank's own charge-off path.

Finance: a product-revenue mix nine years apart tithonus-000381

Question date 2026-03-31. Task 96 of 102 in the finance lifetime, with 9,463 documents in the workspace. Seven criteria, four correctness, one reasoning, and two citation.

Question

How did Repligen Corporation's product-revenue composition change between 2016 and 2025, and which categories drove the largest share gains and losses?

What the record shows

23 Feb 2017 10-K for 2016. Product revenue of $104.4M: Protein 52.4%, Chromatography 28.3%, Filtration 18.9%, and Other 0.4%. The filing gives dollars, and the shares are computed from them.

26 Feb 2026 10-K for 2025. Product revenue of $738.0M: Filtration 54.6%, Chromatography 20.8%, Proteins 13.2%, Process analytics 11.0%, and Other 0.45%.

What a full-credit answer needs

- Both years' shares, computed from the reported amounts, with Filtration as the largest gain (35.65 points) and Proteins as the largest loss (39.19 points).
- *Reasoning.* Process analytics did not exist as a category in 2016, so its 11% is a new franchise and not a like-for-like gain.
- *Citation.* Both 10-Ks.

Five systems

GPT-5.6 Sol ✓ GLM-5.3 ✓ Grok 4.6 ✓ DeepSeek Flash ✓ Claude Sonnet 5 ✗

Claude Sonnet 5 gives Filtration's 2025 share as 54.57% instead of 54.58%, and a criterion that states a number is met only when the number is right.

Figure I.5: A product-revenue mix nine years apart. A calculation over two annual reports, in which a category that did not exist in 2016 must not count as a like-for-like gain.

Medical: restarting after an infusion reaction tithonus-000591

Question date 2022-01-31. Task 55 of 119 in the medical lifetime, with 1,224 documents in the workspace, 25 of them DARZALEX label versions. Eight criteria, two correctness, one reasoning, and five citation.

Question

How did DARZALEX's guidance for interrupting and restarting treatment after an infusion reaction evolve from the initial approach to the current one?

What the record shows

30 Jun 2016 First label. Interrupt the infusion for a reaction of any grade. After a Grade 1–2 reaction, restart at no more than half the rate once symptoms resolve. After a Grade 3 reaction, restart at half rate only once it has eased to Grade 2 or lower, and stop for good at the third Grade 3 or worse reaction. After Grade 4, stop for good.

1 Dec 2016 Label. Grade 3 may restart once symptoms resolve, the same gate as Grade 1–2, and escalation after a Grade 1–2 reaction is capped at 200 mL/hour.

22 Jun 2018 Label. An anaphylactic reaction joins Grade 4 as a reason to stop for good, with emergency care.

29 Jan 2022 Label, two days before the question, carrying an FDA labeling change of 21 January. If eye symptoms occur, interrupt the infusion and obtain an ophthalmologic evaluation before restarting.

What a full-credit answer needs

- The first and the current rules, each complete.
- *Reasoning.* What changed (the Grade 3 restart gate, the anaphylaxis stop, and the eye-symptom rule) and what stayed (interruption for any grade, restart at half rate, and a permanent stop at the third severe reaction or any Grade 4).
- *Citation.* The label version, or FDA record, behind each state.

Five systems

GPT-5.6 Sol ✗ GLM-5.3 ✓ Grok 4.6 ✗ DeepSeek Flash ✗ Claude Sonnet 5 ✓

GPT-5.6 Sol cites no label older than December 2016, gives the first label the later Grade 3 rule, and leaves out the 200 mL/hour cap. Grok 4.6 leaves out the eye-symptom rule (Figure G.4). DeepSeek Flash leaves out the emergency care.

Figure I.6: Restarting after an infusion reaction. Four label versions over six years each change one rule, the last published two days before the question (Figure G.4).

Medical: pregnancy-risk labeling tithonus-000584

Question date 2019-08-31. Task 20 of 119 in the medical lifetime, with 770 documents in the workspace. Seven criteria, three correctness, one reasoning, and three citation.

Question

How did AUBAGIO's pregnancy-risk labeling evolve from 2016 through 2019?

What the record shows

16 Jun 2016 Label. There are no adequate studies in pregnant women, and animal studies show birth defects and fetal death. Pregnancy is ruled out and reliable contraception required before treatment, and after stopping, elimination is accelerated until blood levels are below 0.02 mg/L.

8 Dec 2016 Label, rewritten in the new pregnancy-labeling format with a risk summary, clinical considerations, and a pregnancy registry. The contraindication stays, and a patient who becomes pregnant stops, undergoes accelerated elimination, and is referred to a specialist.

1 Aug 2019 Label. The section becomes Embryofetal Toxicity and adds limited human data: over 150 teriflunomide and over 300 leflunomide pregnancies with early first-trimester exposure followed by accelerated elimination showed no rise in malformations or miscarriage. The data are called too limited to conclude from, and later trimesters are not covered. The controls stay, and a man who wishes to father a child may stop and either accelerate elimination or wait for verified clearance.

What a full-credit answer needs

- Each of the three states with its controls.
- *Reasoning.* The contraindication, contraception, and clearance rules never loosened. The late 2016 change reformatted the evidence, and the 2019 change added human data.
- *Citation.* The three label versions.

Five systems

GPT-5.6 Sol ✗ GLM-5.3 ✗ Grok 4.6 ✓ DeepSeek Flash ✗ Claude Sonnet 5 ✗

The other four never say that the June 2016 label reported no adequate human studies. Claude Sonnet 5 also leaves out the specialist referral and the rule for men.

Figure I.7: Pregnancy-risk labeling. The label is reformatted and gains human data while its controls never loosen, and four systems miss what the first label said about human studies.

Medical: pregnancy as a contraindication in two statins tithonus-000218

Question date 2023-01-31. Task 61 of 119 in the medical lifetime, with 1,389 documents in the workspace. Two entities. Eight criteria, two correctness, three reasoning, and three citation.

Question

When reviewing the historical labeling of LIPITOR and CRESTOR, how did the two products compare from June 2017 through January 2023 in whether pregnancy and lactation or nursing status were classified as contraindications? Explain the sequence of alignment or divergence across material updates and the endpoint as of January 31, 2023, while limiting the conclusion to contraindication-section placement rather than broader safety guidance.

What the record shows

Jun 2017 The labels then current, LIPITOR's of June 2017 and CRESTOR's of June 2016, both list pregnancy and nursing or lactation as contraindications.

17 Dec 2022 LIPITOR label. Its contraindications shrink to acute liver failure or decompensated cirrhosis, and hypersensitivity. CRESTOR's label of May 2022 still lists pregnancy and lactation.

21 Jan 2023 CRESTOR label. The same change, and the two labels agree again.

What a full-credit answer needs

- The sequence: aligned, split with LIPITOR changing first, and aligned again.
- A conclusion kept to the Contraindications section, since pregnancy and breastfeeding guidance remains elsewhere in both labels.
- *Reasoning.* What each section held at each of the three checkpoints.
- *Citation.* The label versions at each checkpoint.

Five systems

GPT-5.6 Sol ✗ GLM-5.3 ✓ Grok 4.6 ✗ DeepSeek Flash ✓ Claude Sonnet 5 ✓

GPT-5.6 Sol and Grok 4.6 say that pregnancy and lactation were removed but not what the sections kept.

Figure I.8: Pregnancy as a contraindication in two statins. Two labels agree, split, and agree again, and the answer has to say what each Contraindications section held at each point.

Medical: a restriction for antiphospholipid syndrome tithonus-000508

Question date 2026-04-30. Task 111 of 119 in the medical lifetime, with 1,837 documents in the workspace. Seven criteria, four correctness, one reasoning, and two citation.

Question

How did XARELTO's antiphospholipid-syndrome restriction change from its initial warning to the later label, and what treatment comparison explains the restriction?

What the record shows

23 Aug 2019 FDA labeling change. Direct oral anticoagulants are not recommended for patients with a history of thrombosis who are diagnosed with antiphospholipid syndrome. They were linked to more recurrent thrombosis than vitamin K antagonists, particularly in patients positive for all three antibodies (lupus anticoagulant, anticardiolipin, and anti-β_2-glycoprotein I).

Oct 2019 on Labels, including those of July 2021 and April 2026. The warning now targets triple-positive antiphospholipid syndrome, with the same antibody definition and the same comparison.

What a full-credit answer needs

- Both wordings, and that the restriction narrowed from a history of thrombosis with the syndrome to triple-positive patients.
- The rationale, more recurrent thrombosis than with vitamin K antagonists.
- *Reasoning.* The antibody definition and the comparison were already in the 2019 warning.
- *Citation.* The FDA record and a later label.

Five systems

GPT-5.6 Sol ✗ GLM-5.3 ✓ Grok 4.6 ✓ DeepSeek Flash ✗ Claude Sonnet 5 ✓

DeepSeek Flash argues that the change is neither a narrowing nor a broadening. GPT-5.6 Sol presents the antibody definition and the comparison as new in the later label.

Figure I.9: A restriction for antiphospholipid syndrome. The restriction narrows in wording while its rationale stays the same, and the answer has to see both.

Medical: neurodevelopmental risk after exposure in utero tithonus-000507

Question date 2026-04-30. Task 110 of 119 in the medical lifetime, with 1,837 documents in the workspace. Six criteria, one correctness, one reasoning, and four citation.

Question

How did Depakote's characterization of neurodevelopmental risk after in-utero exposure evolve from 2016 through April 2026?

What the record shows

27 Feb 2016 Label. Valproate can cause decreased IQ, a causal conclusion. Autism appears as one observational association (2.9-fold, with 4.4% of exposed and 1.5% of unexposed children affected), explicitly not shown to be causal.

5 Mar 2019 Label. The causal wording widens to adverse neurodevelopment, including autism, and the autism study keeps its caveat.

28 May 2020 Label. ADHD is added (adjusted hazard ratio 1.48, 95% CI 1.09–2.00), and the autism and ADHD findings are still not definitive.

14 Apr 2026 Label. The conclusions split. Decreased IQ is still caused by valproate, while autism, intellectual disability (IQ below 70), and ADHD "may be associated", with adjusted risk ranges across studies (1.6–6.4, 1.6–9.6, and 0.7–2.2) and stated limitations.

What a full-credit answer needs

- Every dated characterization with its numbers, in one correctness criterion.
- *Reasoning.* The progression, and above all the 2026 separation of the causal IQ conclusion from the qualified associations.
- *Citation.* The four label versions.

Five systems

GPT-5.6 Sol ✗ GLM-5.3 ✗ Grok 4.6 ✗ DeepSeek Flash ✗ Claude Sonnet 5 ✗

All five miss part of the one correctness criterion, most often the 2020 ADHD estimate or the 2026 ranges and limitations. GLM-5.3, Grok 4.6, and Claude Sonnet 5 also miss the 2026 split.

Figure I.10: Neurodevelopmental risk after exposure in utero. A decade of changing causal language with numbers at every step, and no system solves it.

Medical: one safety signal reaches two statin labels tithonus-000236

Question date 2024-04-02. Task 77 of 119 in the medical lifetime, with 1,557 documents in the workspace. Two entities. Six criteria, three correctness, one reasoning, and two citation.

Question

For a cross-product pharmacovigilance review spanning July 2023 through April 2024, trace how the postmarketing myasthenia gravis signal entered Crestor and Lipitor labeling and synthesize what the combined record implies about changing statins after such an event.

What the record shows

27 Jul 2023 FDA labeling change for CRESTOR: rare new-onset or worsening myasthenia gravis, including ocular myasthenia, with recurrence after the same or a different statin.

2 Apr 2024 The same change for LIPITOR, published on the question date itself.

What a full-credit answer needs

- Both dates in order, and the shared finding.
- The implication that switching statins cannot be assumed to remove the risk, while LIPITOR's later date says nothing about a higher rate for either drug.
- *Reasoning.* One signal spreading across two labels, not two separate findings.
- *Citation.* Both FDA records.

Five systems

GPT-5.6 Sol ✓ GLM-5.3 ✓ Grok 4.6 ✓ DeepSeek Flash ✓ Claude Sonnet 5 ✓

The evidence is two short records, and the second is published on the question date, so it is already in the workspace when the question arrives.

Figure I.11: One safety signal reaches two statin labels. The second record is published on the question date itself, and all five systems solve the task.

Legal: waters of the United States tithonus-000400

Question date 2021-01-31. Task 17 of 79 in the legal lifetime, with 1,008 documents in the workspace. Six criteria, one correctness, one reasoning, and four citation.

Question

How did 40 CFR 232's waters-of-the-United-States definition evolve across the 2018–2021 transition?

What the record shows

1 Jan 2018 eCFR snapshot with the 2015 rule's definition: categorical waters, all adjacent waters, and waters with a case-specific significant nexus, including those in the 100-year floodplain or within 4,000 feet of the high-tide line. The exclusions include prior converted cropland.

1 Jan 2019 The same text, now "applicable beginning on February 6, 2020".

1 Jan 2020 The pre-2015 definition restored: commerce and interstate waters, impoundments, tributaries, the territorial sea, and adjacent wetlands, with prior converted cropland excluded.

1 Jan 2021 The definition is replaced by a cross-reference to § 120.2.

What a full-credit answer needs

- All four states, each with the elements named above.
- *Reasoning.* 6 February 2020 is an applicability date, and a snapshot's date shows which text stood then, not when each clause took effect.
- *Citation.* Each snapshot.

Five systems

GPT-5.6 Sol ✗ GLM-5.3 ✓ Grok 4.6 ✗ DeepSeek Flash ✗ Claude Sonnet 5 ✗

The other four leave prior converted cropland out of the 2015 rule's exclusions, and DeepSeek Flash also leaves out the floodplain category.

Figure I.12: Waters of the United States. Four yearly snapshots of one definition across a repeal and a replacement, and only GLM-5.3 solves the task.

Legal: over-the-counter hearing aids tithonus-000110

Question date 2022-08-31. Task 25 of 79 in the legal lifetime, with 1,267 documents in the workspace. Three entities, 21 CFR parts 800, 801, and 874. Nine criteria, four correctness, one reasoning, and four citation.

Question

For a wireless or self-fitting air-conduction hearing aid intended for consumer sale without a prescription, how did FDA change the proposed acoustic and physical-safety controls in the final rule, and does the device's Part 874 generic type by itself establish OTC status?

What the record shows

20 Oct 2021 Proposed rule. Output is limited to 115 dB SPL, or 120 dB SPL with input-controlled compression and a user volume control, and insertion to the bony-cartilaginous junction of the ear canal.

17 Aug 2022 Final rule. Output is limited to 111 dB SPL, or 117 dB SPL while input-controlled compression is active. The innermost part must stay at least 10 mm from the eardrum, and every over-the-counter hearing aid needs a user-adjustable volume control. Wireless and self-fitting devices remain separate types under Part 874.

What a full-credit answer needs

- The proposed and final output limits, the insertion criterion, and a volume control on every device.
- That the Part 874 type alone does not make a device over-the-counter: it must meet § 800.30, or it is a prescription device labeled under Part 801.
- *Reasoning.* FDA kept the wireless and self-fitting types apart and did not require over-the-counter aids to be self-fitting.
- *Citation.* The proposal and the final rule, part by part.

Five systems

GPT-5.6 Sol ✓ GLM-5.3 ✓ Grok 4.6 ✗ DeepSeek Flash ✓ Claude Sonnet 5 ✓

Grok 4.6 made 19 tool calls and then replied only with its plan ("I'll find the OTC hearing-aid proposal and final rule and compare . . . "), with no answer and no citations.

Figure I.13: Over-the-counter hearing aids. A proposal and a final rule across three CFR parts, and the one failure is an answer that was never written.

Legal: the PFAS reporting rule tithonus-000421

Question date 2026-07-31. Task 71 of 79 in the legal lifetime, with 1,964 documents in the workspace. Ten criteria, two correctness, two reasoning, and six citation.

Question

How did 40 CFR part 705's definition and reporting coverage for PFAS manufacturing evolve from 2021 through 2026?

What the record shows

28 Jun 2021 Proposed rule. PFAS are defined by one structural unit, and reporting covers all PFAS substances and mixtures, including articles and imported articles, made since 1 January 2011.

11 Oct 2023 Final rule. PFAS are defined by any of three substructures, and reporting covers commercial manufacture from 2011 through 2022, with two narrow exclusions.

2024–25 A direct final rule of September 2024 and an interim final rule of May 2025 move only the submission window.

13 Nov 2025 Proposed rule. New exemptions for imported articles, small research quantities, byproducts and similar cases, and concentrations below 0.1%.

13 Apr 2026 Final rule. Only the start date moves again, to 31 January 2027 or 60 days after a final substantive rule if that is sooner, and the exemptions are deferred.

29 Jul 2026 eCFR snapshot. The 2023 definition and coverage still stand.

What a full-credit answer needs

- The definition and the coverage at each stage.
- *Reasoning.* The 2024 and 2025 rules moved dates only, and the 2025 exemptions were still only proposed at the question date.
- *Citation.* The six Federal Register documents and the snapshot.

Five systems

GPT-5.6 Sol ✓ GLM-5.3 ✗ Grok 4.6 ✗ DeepSeek Flash ✗ Claude Sonnet 5 ✗

The other four leave out what the 2021 proposal covered or which exemptions the 2025 proposal offered. DeepSeek Flash never finds the April 2026 rule and treats the May 2025 dates as the last word.

Figure I.14: The PFAS reporting rule. Seven documents over five years, in which two rules moved only dates and a proposal was never adopted.

Legal: a direct final rule takes effect tithonus-000573

Question date 2025-09-30. Task 51 of 79 in the legal lifetime, with 1,766 documents in the workspace. Five criteria, one correctness, one reasoning, and three citation.

Question

For 21 CFR Part 145 (Canned Fruits), how did FDA's July 2025 direct-final process determine whether the Part 145 revocations would take effect?

What the record shows

17 Jul 2025 Direct final rule revoking standards in Part 145, effective 22 September unless a significant adverse comment arrives by 18 August. If one does, FDA withdraws the rule, or the affected part, within 30 days after the comment period.

17 Jul 2025 Companion proposed rule, the ordinary notice-and-comment fallback, with the same comment deadline.

26 Sep 2025 Confirmation that no significant adverse comment arrived, so the revocations took effect on 22 September.

What a full-credit answer needs

- The conditional mechanism with its dates, and the outcome.
- *Reasoning.* The absence of a significant adverse comment is what selected the effective-date branch.
- *Citation.* The three documents.

Five systems

GPT-5.6 Sol ✗ GLM-5.3 ✓ Grok 4.6 ✓ DeepSeek Flash ✓ Claude Sonnet 5 ✗

GPT-5.6 Sol and Claude Sonnet 5 leave out the 30-day withdrawal deadline, and Claude Sonnet 5 cites no documents. The confirmation appeared four days after the effective date, so the answer rests on a document published after the event it confirms.

Figure I.15: A direct final rule takes effect. Whether the rule took effect turns on a comment that never arrived, confirmed four days after the effective date.

Legal: aerosol cans become universal waste tithonus-000682

Question date 2021-04-30. Task 20 of 79 in the legal lifetime, with 1,051 documents in the workspace. Two entities, 40 CFR parts 262 and 273. Thirty-nine criteria, 21 correctness, 5 reasoning, and 13 citation.

Question, condensed from the scenario the agent received

A facility under the federal hazardous-waste rules has the same facts in March 2018 and March 2021. On 1 March of each year it generates 900 kg of non-acute hazardous solvent waste, discards 200 kg of intact aerosol cans that are hazardous only because they are ignitable, and holds 4,900 kg of universal-waste batteries and lamps first accumulated that day. In 2021 it manages the cans as universal waste. As of 30 April of each year, how did the intervening federal change alter its generator and storage position under Part 262 and its handler and storage position under Part 273?

What the record shows

1 Jan 2018 eCFR snapshots. Aerosol cans are not universal waste. A generator of 1,000 kg or more a month is a large quantity generator with 90 days of storage, and a handler below 5,000 kg is a small quantity handler. Universal waste may be stored for one year.

9 Dec 2019 Final rule. Hazardous aerosol cans become universal waste, effective 7 February 2020.

1 Jan 2021 Snapshots. The cans may be managed under Part 273. A generator of more than 100 and less than 1,000 kg a month is a small quantity generator with 180 days of storage, and a handler of 5,000 kg or more is a large quantity handler.

What a full-credit answer needs, from 2018 to 2021

- Part 262: 1,100 then 900 kg counted, so a large and then a small quantity generator, with 90 and then 180 days of storage, leaving 30 and then 120 days on 30 April.
- Part 273: 4,900 then 5,100 kg held, so a small and then a large quantity handler, with 305 days left in both years.
- Twenty-one correctness criteria, one per number or tier. *Reasoning.* Five criteria, such as counting the same 200 kg in exactly one program each year and comparing each total with the right threshold. *Citation.* Thirteen criteria, one per rule relied on.

Five systems

GPT-5.6 Sol ✓ GLM-5.3 ✗ Grok 4.6 ✗ DeepSeek Flash ✓ Claude Sonnet 5 ✗

Grok 4.6 and Claude Sonnet 5 give the one-year deadlines but never the 305 days left. GLM-5.3 gets every number right but cites only the 2019 rule, which does not state the thresholds and limits it uses, and fails five citation criteria.

Figure I.16: Aerosol cans become universal waste. One rule change moves 200 kg between two programs and flips the facility's tier in both, with 39 criteria.