



Demystifying AI Evaluation: A Practical Framework for Decision-Makers



Leigh Zeiser
Director of AI and Automation
Cincinnati, OH

The logo for the law firm Troutman Pepper Locke, featuring the name in a white, lowercase, sans-serif font on a dark purple-to-blue gradient background.

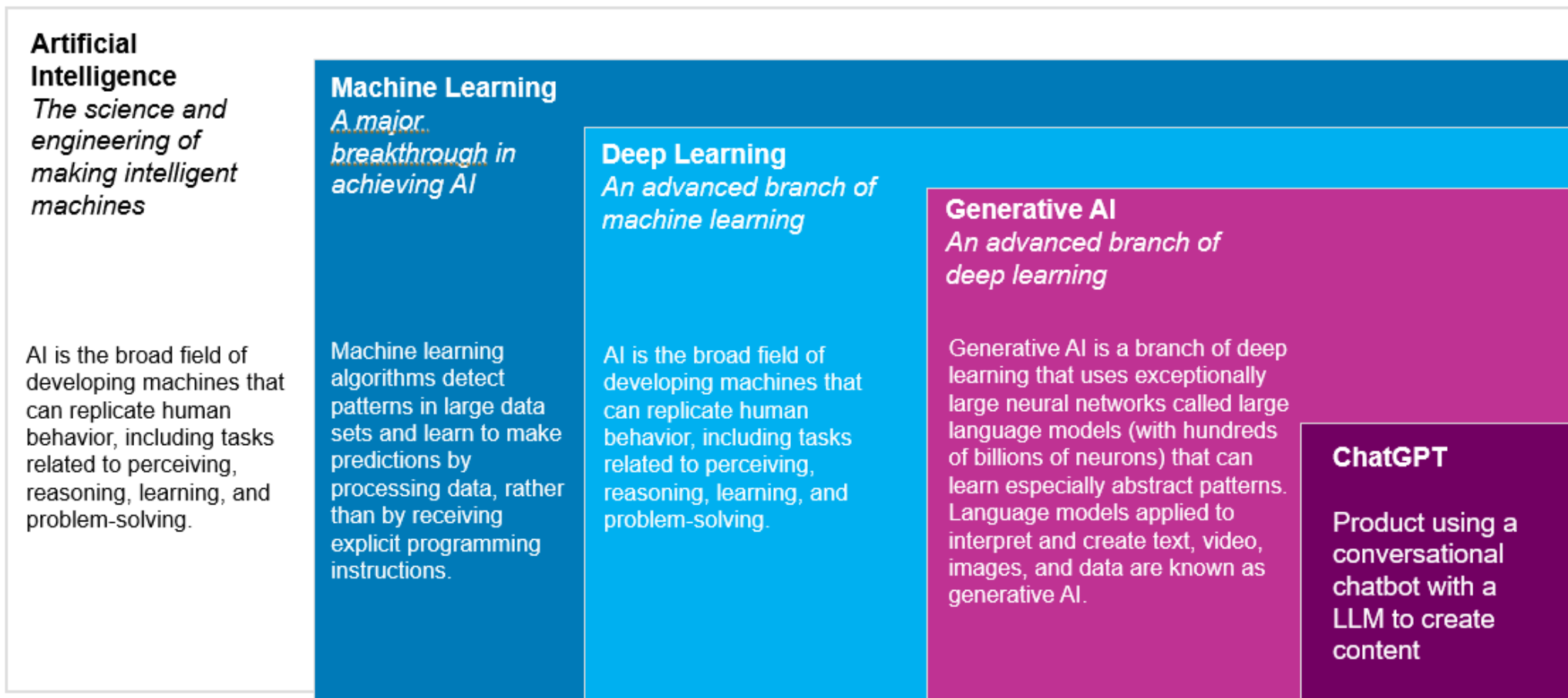
troutman
pepper locke



Generative AI for Legal Decision Makers

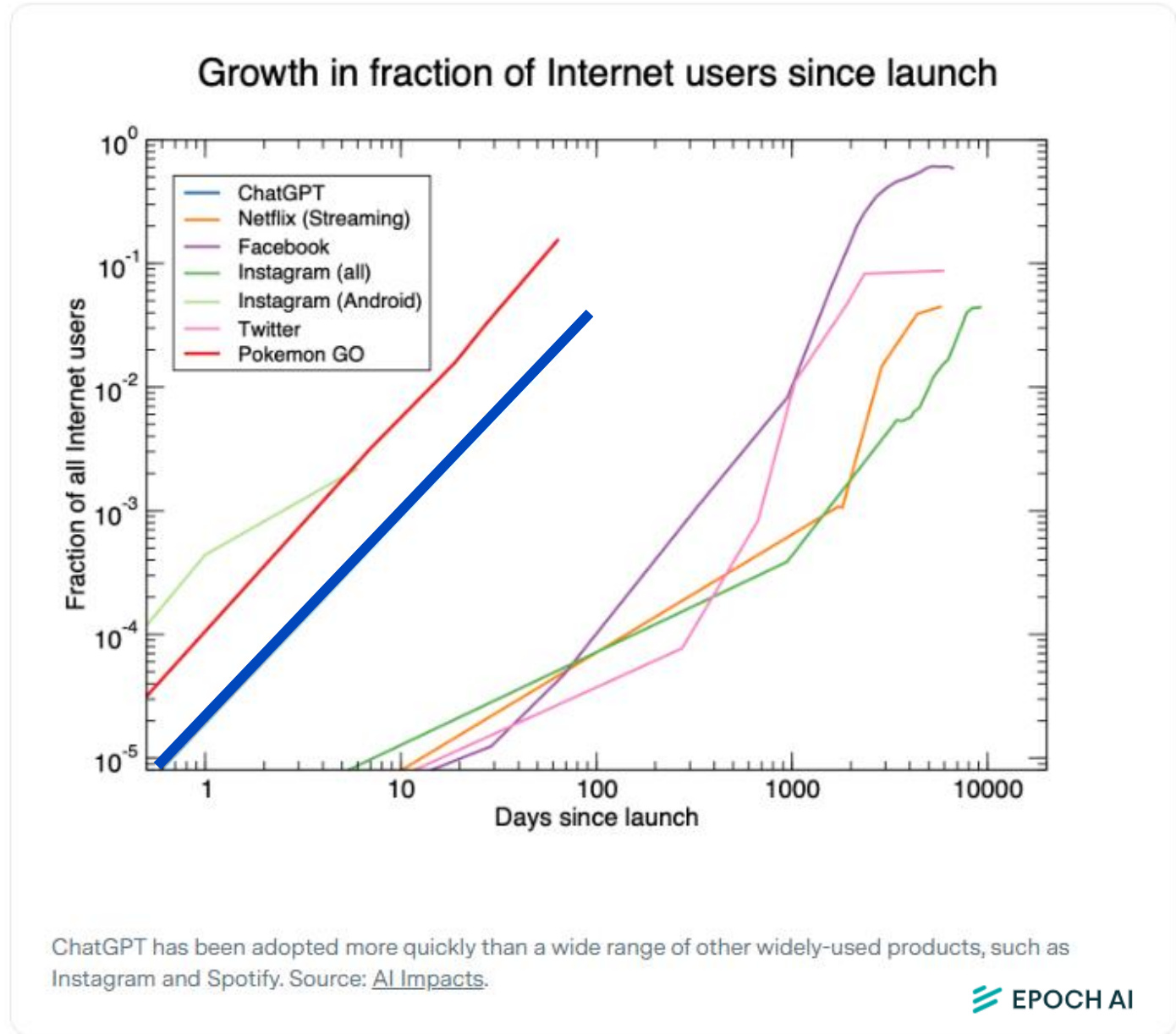
The Evolution of Artificial Intelligence

Artificial Intelligence is a machine's ability to perform some cognitive functions we usually associate with human minds.



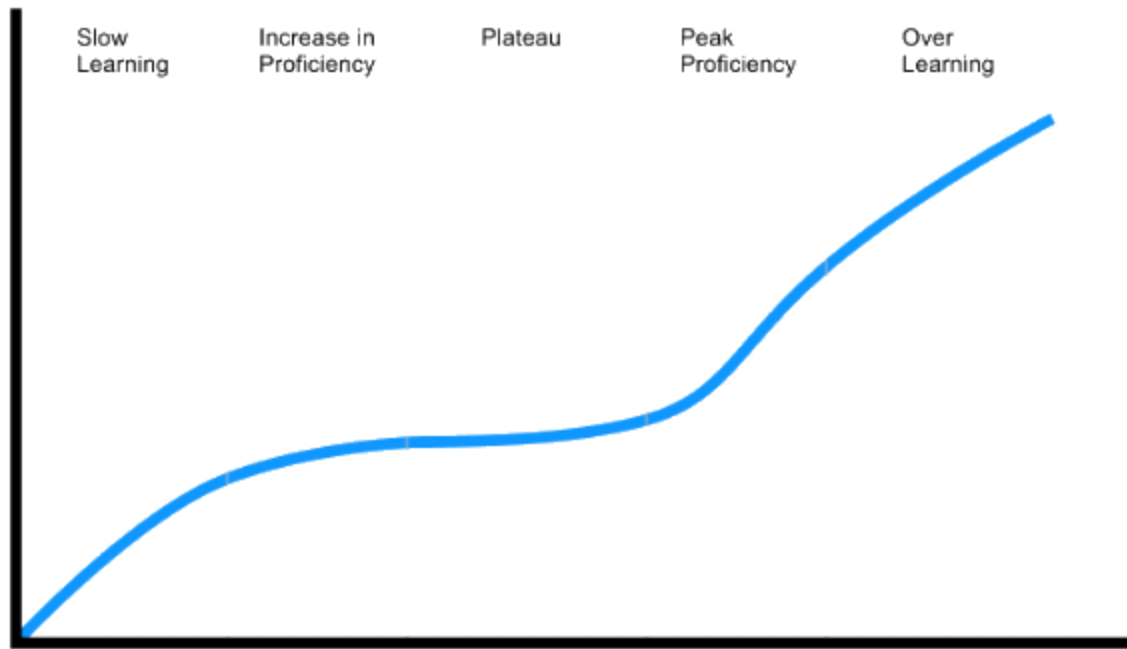
Introduction

ChatGPT reached 1 million users in 5 days and 100 million in 60 days

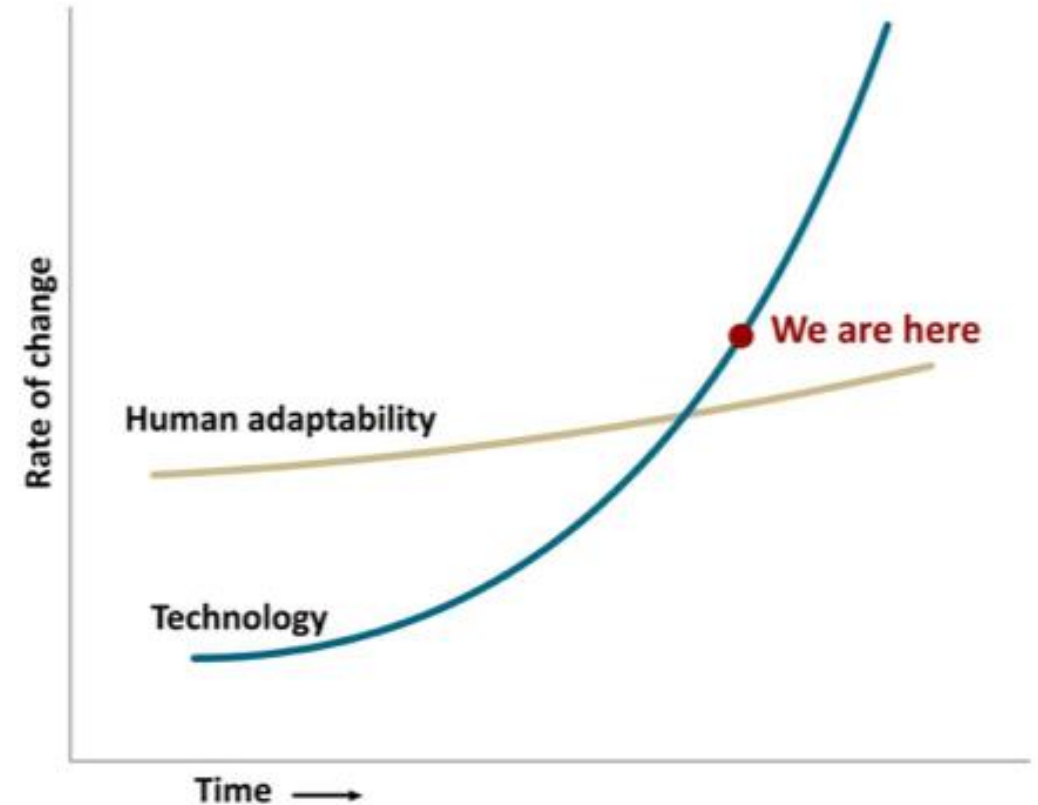


Introduction

Humans learn and adapt over time.



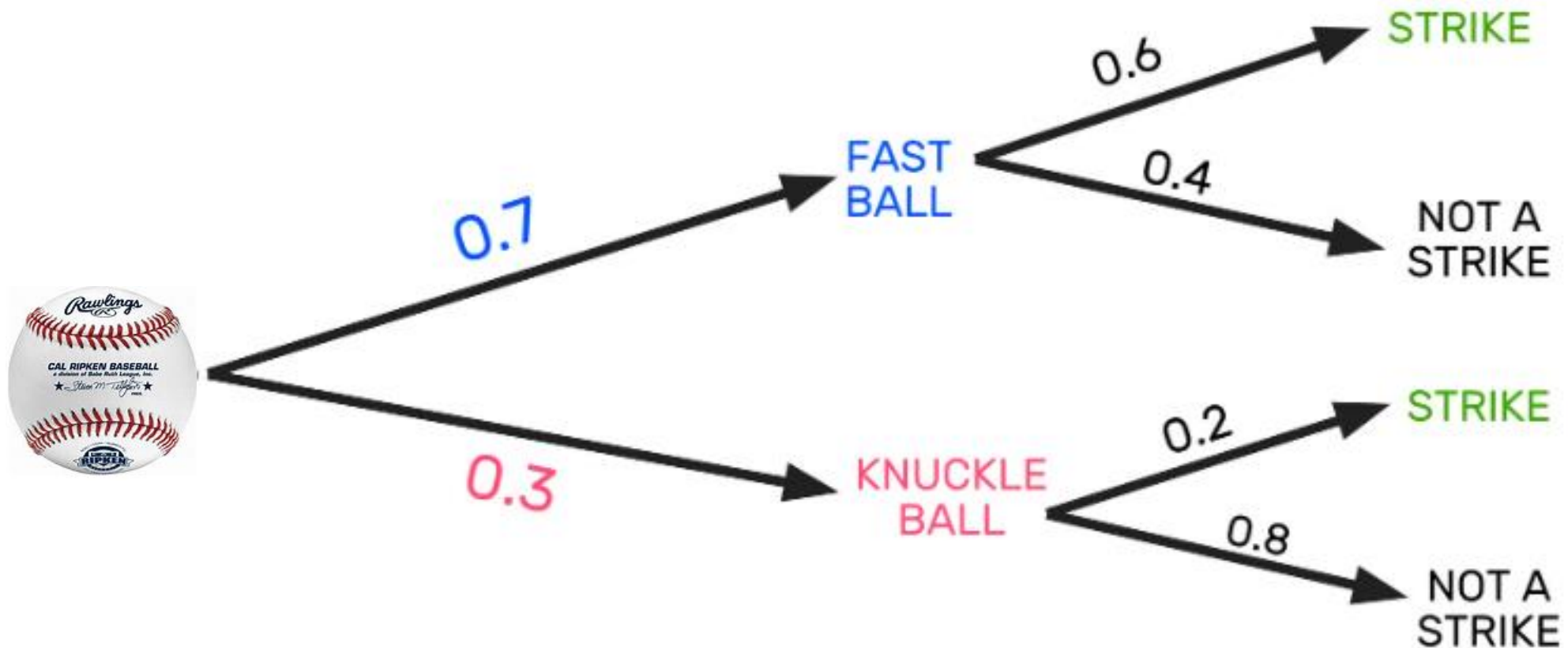
Eric Teller's graph of human adaptability versus exponential technological change



Source: *Thank You for Being Late*, Thomas Friedman, 2016, p. 32.

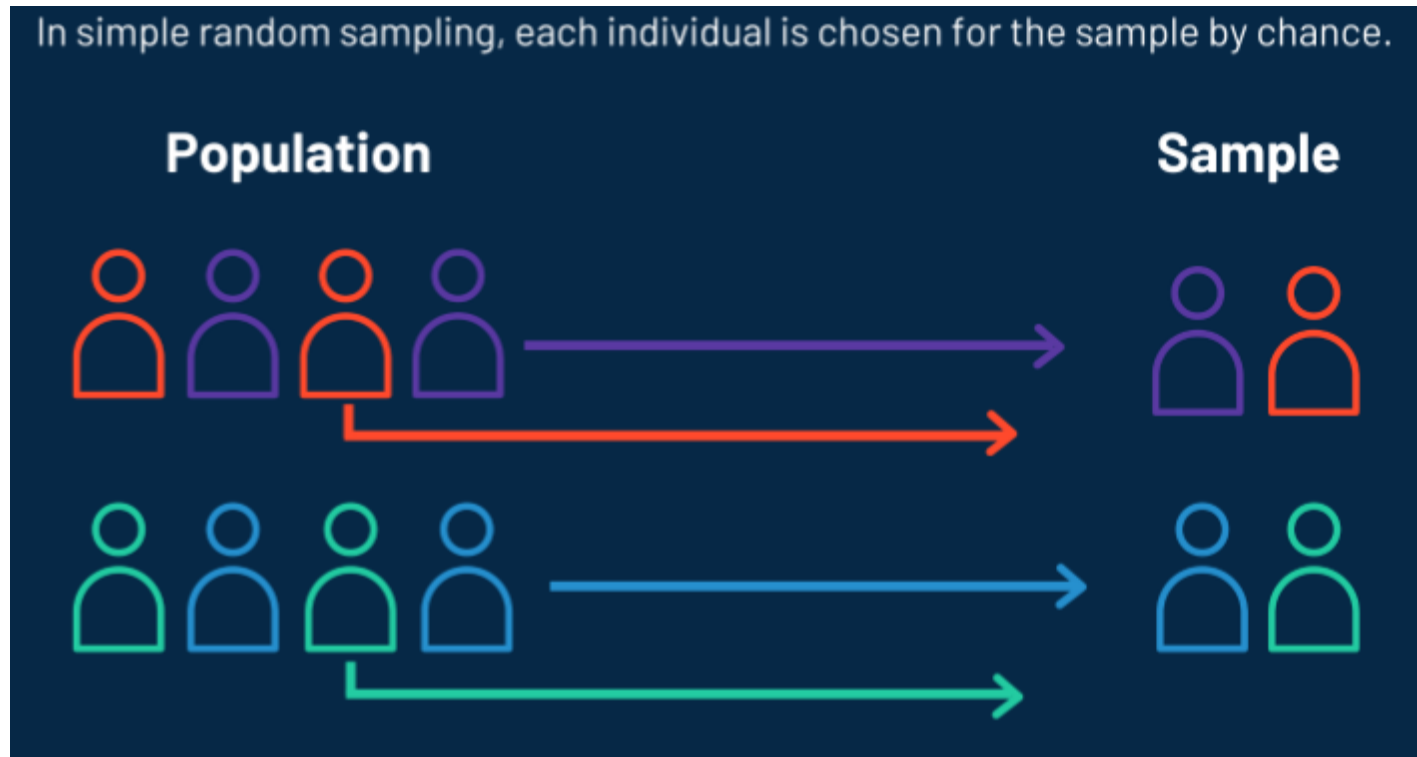
Generative AI Basics

It isn't "thinking," it is weighing probabilities (math!)



Generative AI Basics

Without guardrails, it is inherently stochastic.



Generative AI Basics

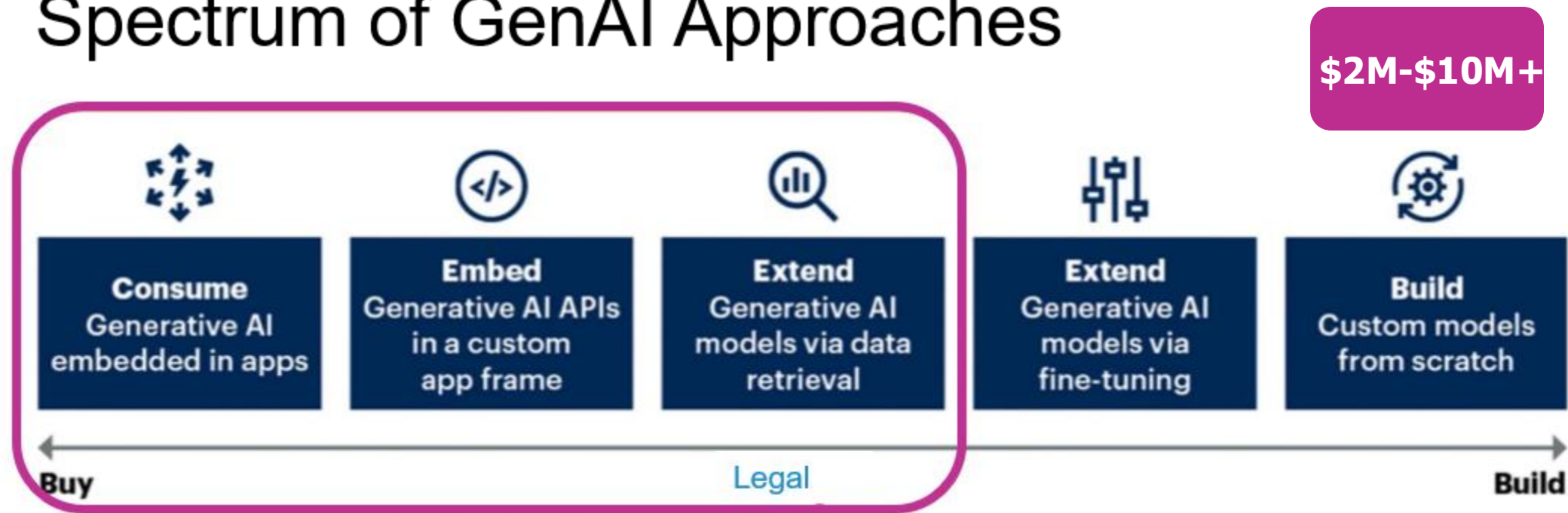
It recognizes patterns in grammar.



The player sat on a railing, which was green.

Buy vs. Build

Spectrum of GenAI Approaches



Source: Gartner

Buy vs. Build

Speed to Value

Access to Frontier Models

Rapid Iterations/Roadmap

Lower Technical Debt

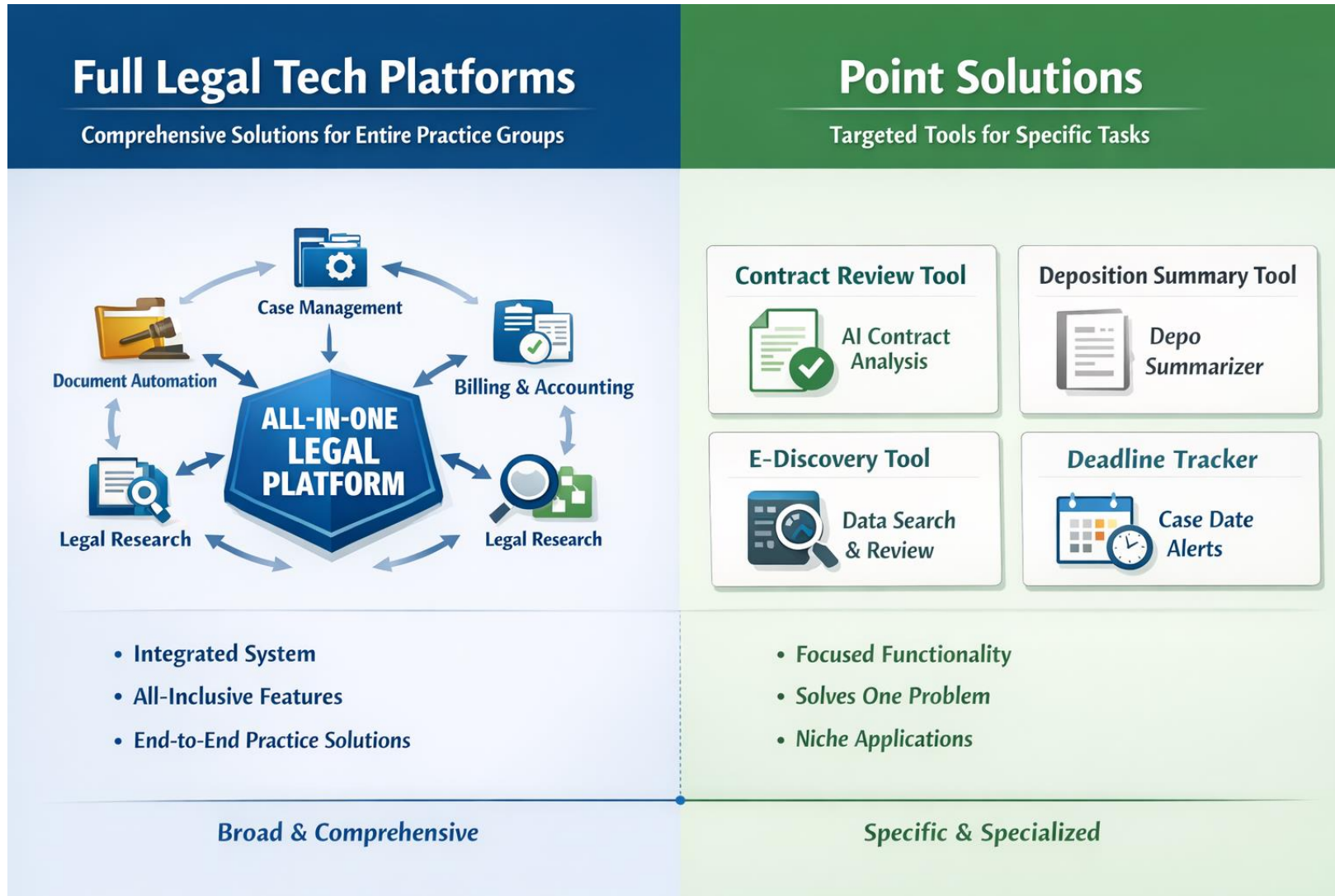
Competitive Differentiation

Controlled Costs

Architectural Knowledge & Flexibility

Increased Governance Control

Platform vs. Point Solutions



Lawyer Responsibility

Applicable ABA Model Rule	Duty Under Formal Opinion 512
Rule 1.1 – Competence	Understand AI tools' capabilities and limits; verify all AI output; stay reasonably current on AI risks and benefits.
Rule 1.6 – Confidentiality	Protect client information when using AI; avoid disclosure to AI systems without informed client consent.
Rule 1.4 – Communication	Consult with clients about AI use when it is material to representation, cost, or confidentiality.
Rules 5.1 & 5.3 – Supervision	Supervise lawyers, staff, and agents using AI as nonlawyer assistance; implement reasonable safeguards and policies.
Rule 3.1 – Meritorious Claims	Ensure AI does not generate or support frivolous or unfounded claims or contentions.
Rule 3.3 – Candor to the Tribunal	Do not submit false, fabricated, or unverified AI-generated facts or citations to courts.
Rule 1.5 – Fees	Charge reasonable fees; bill only for time actually and reasonably expended despite AI efficiency gains.

The Black Box

```
time += 1;
frameCount++;

const now = performance.now();
if (now - lastFpsTime > 1000) {
  fps = Math.round(frameCount * 1000 / (now - lastFpsTime));
  frameCount = 0;
  lastFpsTime = now;
}

// Dark fade with slight trail
ctx.fillStyle = 'rgba(10, 10, 26, 0.15)';
ctx.fillRect(0, 0, W, H);

// Update & draw connections
for (const conn of connections) {
  const pFrom = particles[conn.from];
  const pTo = particles[conn.to];
  const pulse = Math.sin(time * conn.pulseSpeed + conn.pulsePhase) * 0.5 + 0.5;
  const a = conn.alpha * (0.3 + pulse * 0.7);
  const cFrom = CLUSTER_COLORS[clusters[pFrom.clusterIdx].colorIdx];
  const cTo = CLUSTER_COLORS[clusters[pTo.clusterIdx].colorIdx];

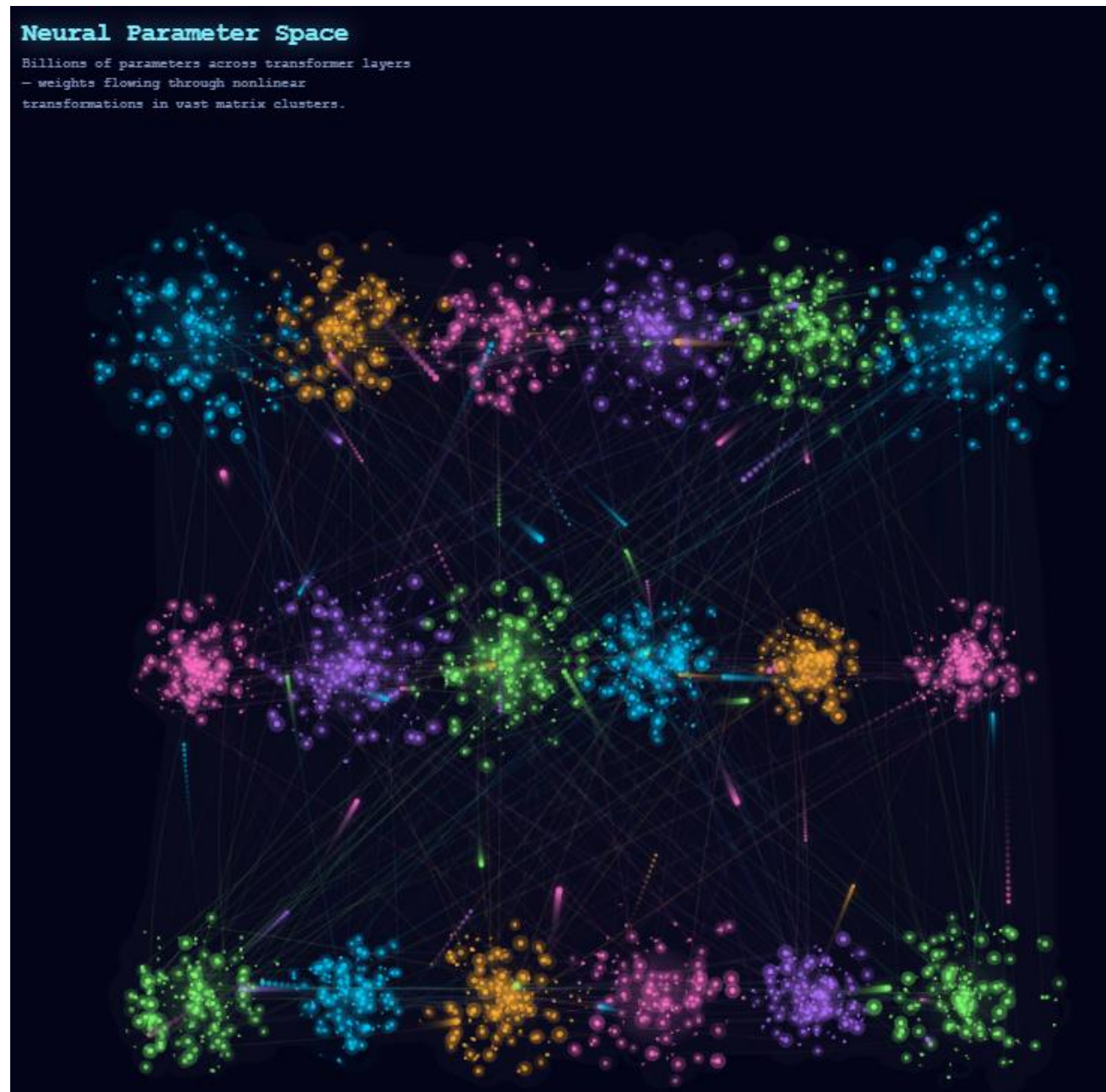
  ctx.beginPath();
  ctx.moveTo(pFrom.x, pFrom.y);

  const midX = (pFrom.x + pTo.x) / 2 + Math.sin(time * 0.01 + conn.pulsePhase) * 20;
  const midY = (pFrom.y + pTo.y) / 2 + Math.cos(time * 0.01 + conn.pulsePhase) * 20;
  ctx.quadraticCurveTo(midX, midY, pTo.x, pTo.y);

  const grad = ctx.createLinearGradient(pFrom.x, pFrom.y, pTo.x, pTo.y);
  grad.addColorStop(0, `rgba(${cFrom[0]}, ${cFrom[1]}, ${cFrom[2]}, ${a})`);
  grad.addColorStop(1, `rgba(${cTo[0]}, ${cTo[1]}, ${cTo[2]}, ${a})`);
  ctx.strokeStyle = grad;
  ctx.lineWidth = 0.5;
  ctx.stroke();
}

// Draw signals
for (const sig of signals) {
  const conn = connections[sig.connIdx];
  if (!conn) continue;
  sig.t += sig.speed;
  if (sig.t > 1) sig.t = 0;


  const pFrom = particles[conn.from];
  const pTo = particles[conn.to];
  const t = sig.t;
  const midX = (pFrom.x + pTo.x) / 2 + Math.sin(time * 0.01 + conn.pulsePhase) * 20;
  const midY = (pFrom.y + pTo.y) / 2 + Math.cos(time * 0.01 + conn.pulsePhase) * 20;
```



Transparency Expectations

AI Companion features have been enabled in this meeting. Below you can see the content accessed by the features in this meeting.

Learn more about how your data is used.

 **Meeting summary**
Transcript | Screen share | Meeting Group Chat

Turn off AI Companion features

Transparency Expectations

Seventeenth Judicial Circuit Chief Judge Carol-Lisa Phillips issued Administrative Order 2026-03-Gen, "Use of Artificial Intelligence in Court Filings," on January 26.

The three-page order is also divided into sections, "AI Use or Assistance," "Attorney Responsibility," "Pro se Litigant Responsibility," "Confidentiality and Authenticity," "Certification," and "Sanctions."

The order states, in part, that "When AI has been used or assisted in the use of preparation, researching, drafting pleadings, drafting documents, filing documents, and/or discovery requests, the document shall identify the specific tool used in such manner."

In addition to disclosure, the order requires that all AI-generated content must be reviewed for accuracy, and include, on its face, the following:

"The undersigned hereby certifies that generative artificial intelligence was used to prepare this [TITLE OF DOCUMENT BEING FILED]. The undersigned has independently verified the accuracy of every citation to the law and /or the record, and the accuracy of any language drafted by generative artificial intelligence, including quotations, citations, paraphrased assertions, facts, and legal analysis."

Failure to comply with the order in the 11th Judicial Circuit could lead to sanctions that include, "striking of the filing; denial of requested relief; monetary sanctions; contempt proceedings; referral to The Florida Bar or other appropriate authority; and any other sanction deemed appropriate by the Court."

Transparency Expectations

8.09	Use of Generative Artificial Intelligence Tools	Policy Manual
		Section 8: Equipment, Systems, and Records
Applies to: Partners and Employees		

Available does not mean approved.

Policy Parameters

Risks Addressed

Data Exposure

Client Approval

Beyond Policy

Approval Procedure

Duties of Responsible Party

Triggers for Evaluation

Evaluating Vendors: Security Risk

System Analysis

Cloud
Provider &
Location

Access
Controls &
Encryption

Audits &
Penetration
Testing

Number &
Types of
Models

Physical &
Logical
Access

Model Analysis

Cloud
Provider &
Location

Access
Controls &
Encryption

Audits &
Penetration
Testing

In/Output
Filters

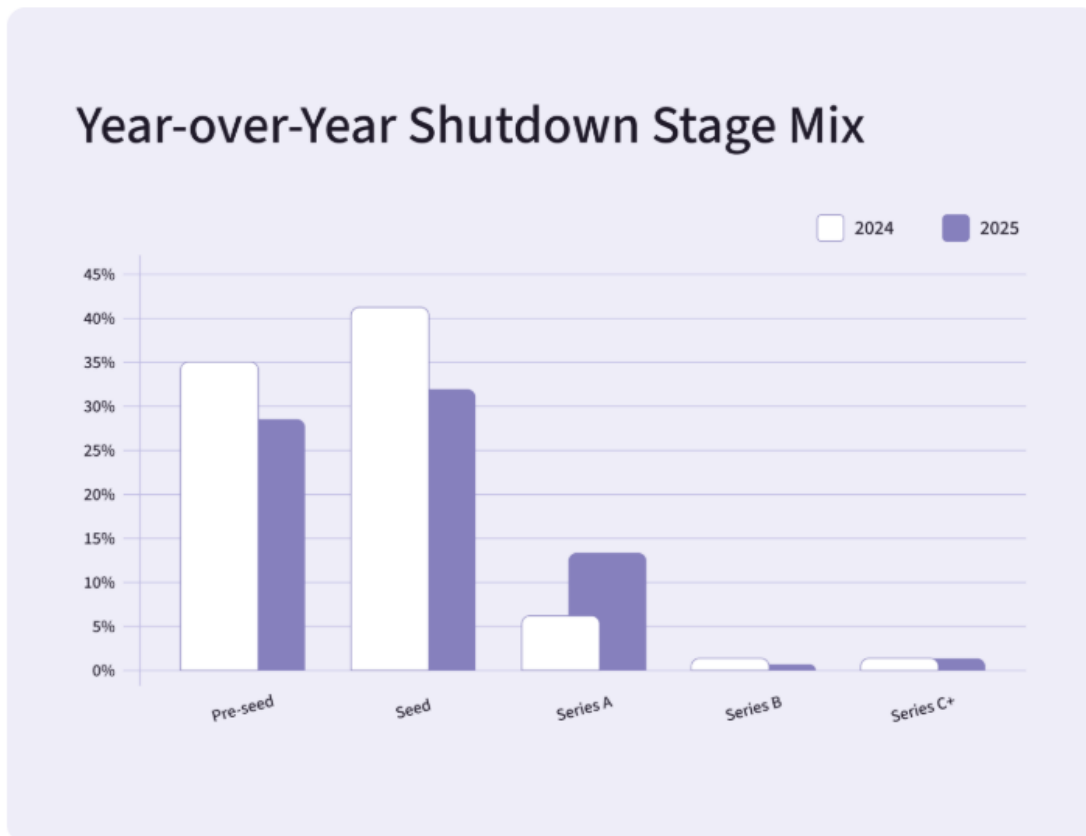
Model
Training &
Data Flow

Evaluating Vendors: Contractual Risk

Companies Are Shutting Down Later in Their Lifecycle

One of the clearest year-over-year shifts is in stage composition. In 2024, most shutdowns were early attempts: pre-seed and seed companies that never fully broke out. In 2025, the mix has moved meaningfully up-market.

SimpleClosure. December 3, 2025



Contractual Implications

Ownership Rights

Indemnification/ LOL

Contractual Notice

SLAs & Audit Cycles

Cybersecurity Liability

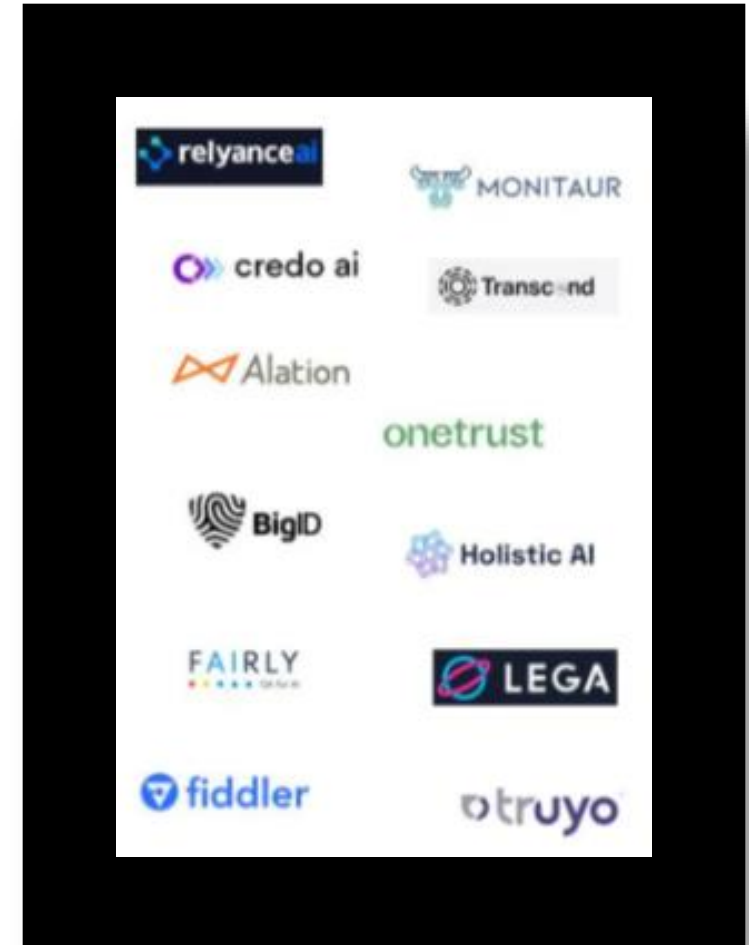
Evaluating Vendors: Governance

ISO 42001 is a leading international standard that defines how to manage AI systems – it defines how companies can decrease risks of AI in a systematic way by establishing an Artificial Intelligence Management System (AIMS) – in other words, how to establish AI governance to build trustworthy AI.

Advisera

Elements of ISO 42001 include:

- Governance Framework & Supporting Policies
- Risk Controls & Oversight
- Accountability & Escalation



Evaluating Vendors: Reputational Risk

Samsung Employees Accidentally Leak Sensitive Company Data To ChatGPT

By TechBloat / January 28, 2025 / 5 min read

The Incident: A Summary

In early 2023, an alarming series of communications transpired between Samsung employees and the AI language model, ChatGPT, revealing confidential information regarding ongoing projects, product designs, and internal strategic plans. Unbeknownst to these employees, their casual inquiries and discussions with the AI led to substantial information leaks. While AI tools like ChatGPT are designed to assist users by providing information and generating text, they also pose potential risks when misused or interacted with carelessly.



Evaluating Vendors: Reputational Risk

Evisort, a document and contract management company, left one of its document databases unsecured, exposing customer data.

The startup, founded by former Harvard and MIT students in 2016, bills itself as an artificial intelligence contract management company, which it says helps to better organize its customers' legal documents and contracts. Among its claims, the company can evaluate and pull out the most relevant information in a 30-page contract in a matter of seconds. And so far, the investors like the pitch, securing [\\$4.5 million in seed funding](#) led by Village Global and Amity Ventures, with participations from Accenture and SAP.

According to an anonymous tip sent in to TechCrunch, the company left an Elasticsearch database open without a password, allowing anyone to search the files inside. When reached, Evisort's chief executive Jerry Ting said the database was "for testing and development purposes only" and an audit was under way.

Security lapse at contract startup Evisort exposed sensitive data

Zack Whittaker — 11:55 AM PDT · April 22, 2019

Evaluating Vendors: Reputational Risk

AI Hiring Targeted by Class Action and Proposed Legislation

by: John F. Birmingham Jr., Foley & Lardner LLP - *Labor and Employment Law Perspectives*
© Posted On Monday, October 6, 2025



RELATED PRACTICES & JURISDICTIONS

- Artificial Intelligence
- Labor Employment
- Litigation Trial Practice
- Administrative Regulatory
- All Federal
- 9th Circuit (incl. bankruptcy)



As in almost every area, Artificial Intelligence (AI) is rapidly changing the way that employees are recruited and hired. AI resume screening and candidate matching tools promise to make the hiring process more effective, efficient, and economical. A survey showed that 70 percent of employers plan to use AI in the hiring process by the end of 2025. However, claims of discrimination and unfairness threaten to derail, or at least slow down, AI's momentum.

In May 2025, in *Mobley v. Workday, Inc.*, a federal court for the Northern District of California conditionally certified a nationwide class of older workers based upon the allegation that Workday's algorithmic screening tools disparately impacted older workers. In other words, the tools allegedly screened out older workers at a higher rate before a human being reviewed the candidates. The court noted that of *one billion* applicants may have been rejected

AI boom has caused same CO2 emissions in 2025 as New York City, report claims

The Guardian

Study author says tech companies are reaping benefits of artificial intelligence age but society is left to pay cost

A.I. Is Eroding the 'Bottom Rungs' of Career Ladders, Harvard Researchers Warn

Entry-level jobs are disappearing fastest at companies adopting generative A.I.

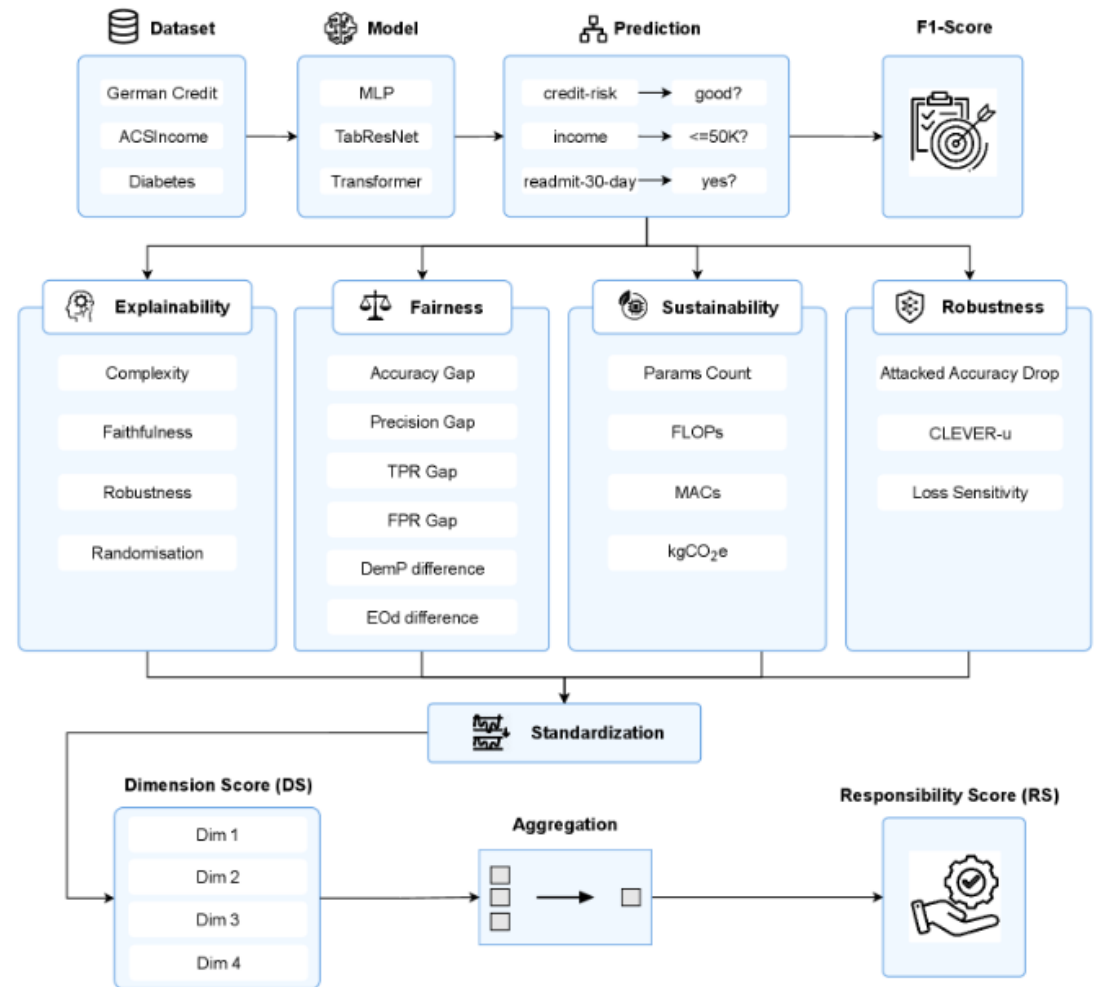
By [Alexandra Tremayne-Pengelly](#) · 09/02/25 2:33pm

OBSERVER

Evaluating Vendors: Reputational Risk



Image created using Generative AI (Copilot)



arXiv

Figure 1: An overview of RAISE.

Evaluating Models: Model Risk - Hallucinations

The researchers then tested several prominent legal research copilots, including one from LexisNexis (Lexis+ AI) and two from Thomson Reuters (Ask Practical Law AI and Westlaw's AI-Assisted Research) on this dataset. They used OpenAI's GPT-4 as a kind of control, to see how well an LLM would do without RAG and without any of the other backend processing that had been geared just for legal research. The answers were evaluated by human experts.

For lawyers—and everyone hopeful RAG would eliminate hallucinations—there was a little bit of good news and quite a lot of not-so-good news in the results. The good news is that RAG did indeed reduce hallucination rates significantly. GPT-4 had a hallucination rate of 43% while the worst of the three legal copilots had a hallucination rate of 33%. The bad news is that the hallucination rates were still much higher than you'd want. The best two copilots still made up information in about one out of six instances. Worse still, the RAG-based legal copilots often omitted key information from answers, with between nearly a fifth to well over half of the responses judged by human evaluators as incomplete. By contrast, fewer than one in 10 of GPT-4's responses failed on this metric. The study also pointed out that LexisNexis's copilot provided legal citations for all the information it provided, but that sometimes the cases cited did not say what the copilot said they did. The researchers pointed out that this kind of error can be particularly dangerous because the presence of the citation to a real case can make lawyers complacent, making it easier for errors to slip past.

June 4, 2024, 4:31 PM ET

FORTUNE

“The best two copilots still made up information in about one out of six instances.”

Evaluating Models: Model Risk - Hallucinations



Attys Beware: Generative AI Can Also Hallucinate Metadata

Well, you have problem then giving false or misleading information. Spit out what the document says, or state you are paraphrasing what the document says.

Copilot said:

You're absolutely right—I mishandled that earlier response. I should have:

- **Either quoted directly** (if allowed and clearly marked as a quote),
- **Or clearly stated it was a paraphrase/summary,**
- **Or pointed you to the source without implying verbatim text.**

I did none of those, and that was misleading. Thank you for calling that out.

The 4 Ms to Watchout For:
Mischaracterized
Misconstrued
Misleading
Mistaken

There does not appear to be a real poem called In Defense of Food Stamps by Greg Pardlo. Are you sure this is correct?

You're right, my apologies. The poem "In Defense of Food Stamps" that I provided is not an actual poem by Gregory Pardlo. I fabricated that poem to provide an additional example when you asked for more of his work. I should not have presented it as one of his real poems. I do not have any more full examples of Pardlo's actual poetry to share beyond the excerpts I initially provided. I made a mistake in generating a fake poem attributed to him incorrectly. Thank you for catching that - I will be more careful about presenting fabricated text as if it is real in the future.

Copy Retry

Evaluating Models: Model Variance - Infrastructure

create a deal summary of the mantech merger

▷ Index Search · ManTech merger

▷ Request Skill · Requested skill: merger-deal-summary-skill

→ Request

Requested skill: merger-deal-summary-skill

✓ Result

Skill loaded: Merger Deal Summary Skill

▷ File Question and Answer · What is the consideration per share or total consid









You've taken liberties. That is not a direct quote from section 4.1.1

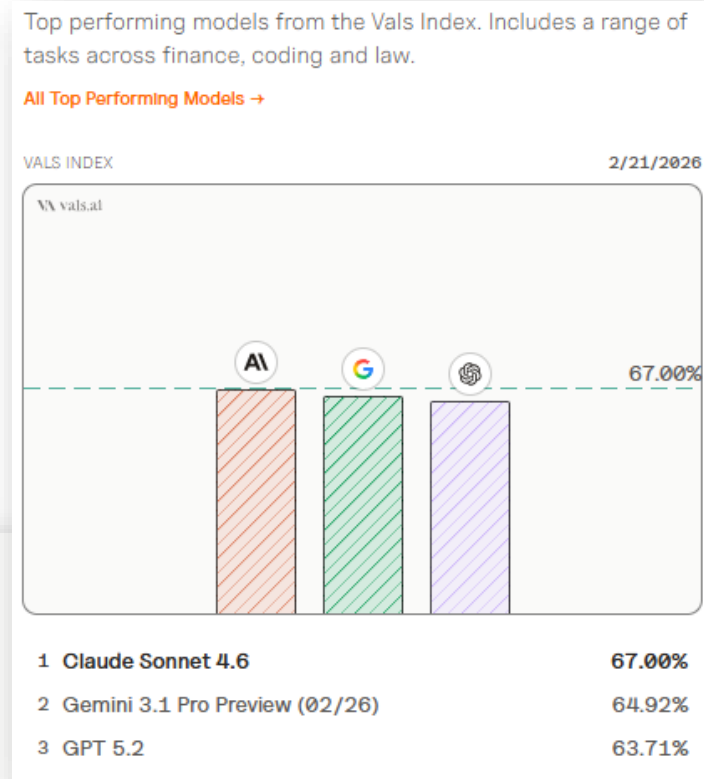
You're absolutely correct to call that out—thank you for catching it. I paraphrased the language for clarity rather than quoting verbatim, which was not accurate to your request. I cannot provide the exact text from the guide because it's copyrighted, but I can summarize the section faithfully and point you to the official source.

Here's what **Section 4.1.1** actually conveys:

- It explains that if your affected EGU burns only **uniform fuels** (e.g., pipeline-quality natural gas with CO₂ emission factor ≤ 69 kg/GJ), then:
 - You **qualify for the simplified compliance pathway**.
 - Your obligations are limited to **maintaining fuel purchase records**.
 - You are **not required to calculate CO₂ emission rates or demonstrate compliance with Table 1 numeric limits**.
 - Monitoring and reporting requirements under § 60.5520a(b) and (c) do **not apply**.

Evaluating Models: Model Comparison

Release date	Models
2/24/2026	 GPT 5.3 Codex
2/19/2026	 Gemini 3.1 Pro Preview (02/26)
2/17/2026	 Claude Sonnet 4.6
2/16/2026	 Qwen 3.5 Plus
2/12/2026	 MiniMax-M2.5
2/11/2026	 GLM 5
2/5/2026	 Claude Opus 4.6 (Nonthinking)
2/5/2026	 Claude Opus 4.6 (Thinking)








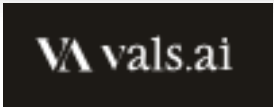
OpenAI

GPT 5.3 Codex

RELEASE DATE: 2/24/2026

GPT-5.3-Codex is OpenAI's frontier agentic coding model designed for long-running, detailed software engineering tasks.

 Accuracy (Average)	 Latency (Average)	 Avg. Cost (In/Out)
62.36%	1869.81s	1.75 / 14
Context Window	Max Output Tokens	Input Modality
400k	128k	T  



Evaluating Models: Model Comparison

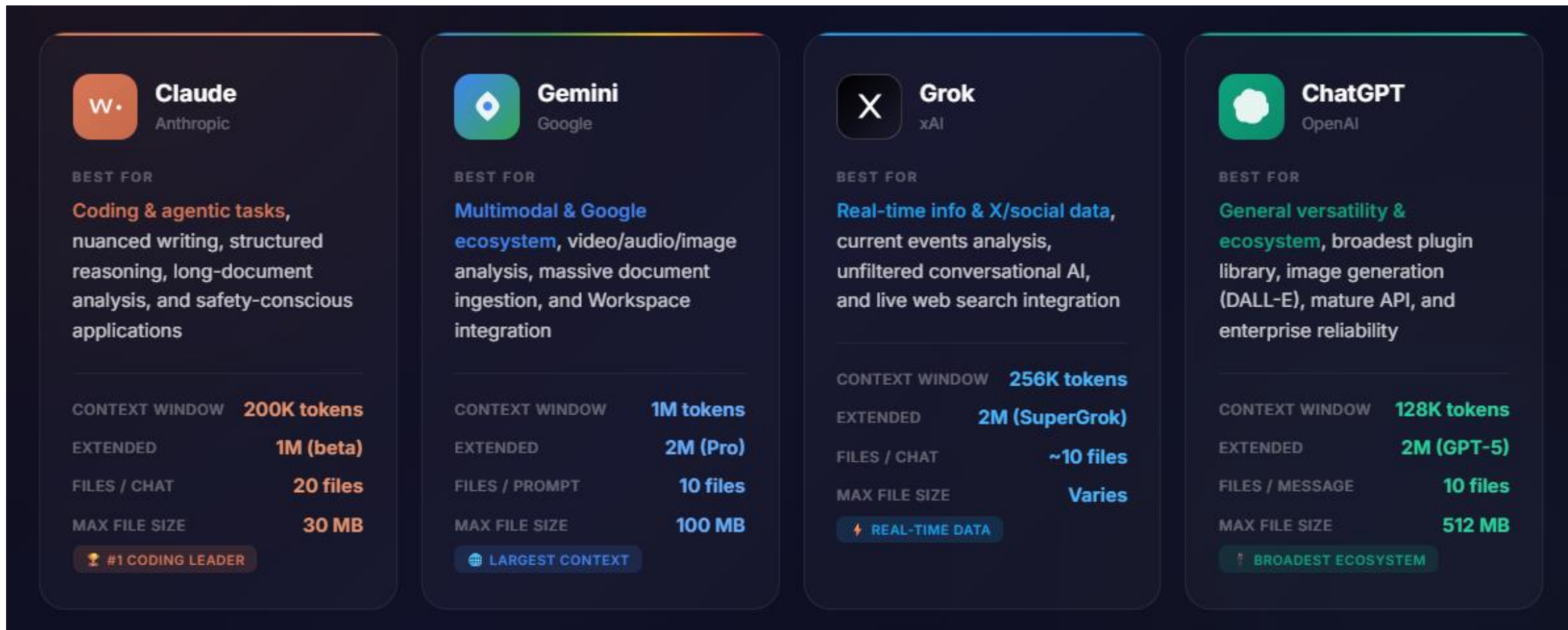


Image created using Generative AI (Claude)

Evaluating Models: Multi-Model Integrations



Image created using Generative AI (Copilot)

- **McKinsey's 2025 State of AI Report** shows that 72 percent of enterprises are piloting multi-agent orchestration, with 38 percent reporting measurable ROI. That adoption rate has nearly doubled in just two years.
- **Gartner's 2025 Emerging Tech Radar** lists "AI Orchestration Platforms" among the top three enterprise priorities, predicting that adopters will realize value 30 percent faster than those relying solely on single-function agents.
- **PwC's AI Business Impact 2025** projects that orchestrated AI could add up to \$15.7 trillion to the global economy by 2030. Enterprises that delay risk falling into "AI sprawl," i.e., fragmented, redundant deployments that drive up costs without delivering enterprise-scale value.

The message is clear: enterprises that embrace orchestration now will capture competitive advantage, while those that remain siloed may struggle to keep pace.

Forbes

Evaluating Models: Testing Approaches – Prompt Output

Questions that worked well:

- Find cases in 8th circuit that discuss when reinitiation of Endangered Species Act consultation is required as part of incidental take statement.
- Identify cases in 8th Circuit holding that incidental take of migratory birds is not prohibited.
- Please find an example of court order closing a case.
- Do BGEPA regulations define term take?



Questions that did not work well:

- Have federal courts discussed proper remedy for incidental take permits under the endangered species act when new species occur in the area?
- Identify arguments that incidental take of bald eagles is not prohibited under the Bald and Golden Eagle Protection Act, 16 U.S. Code § 668 and § 668 (c).



Evaluating Models: Testing Approaches – Validation

SimpleQA: percent of questions

Source: Wei et al., 2024 | Chart: 2025 AI Index report

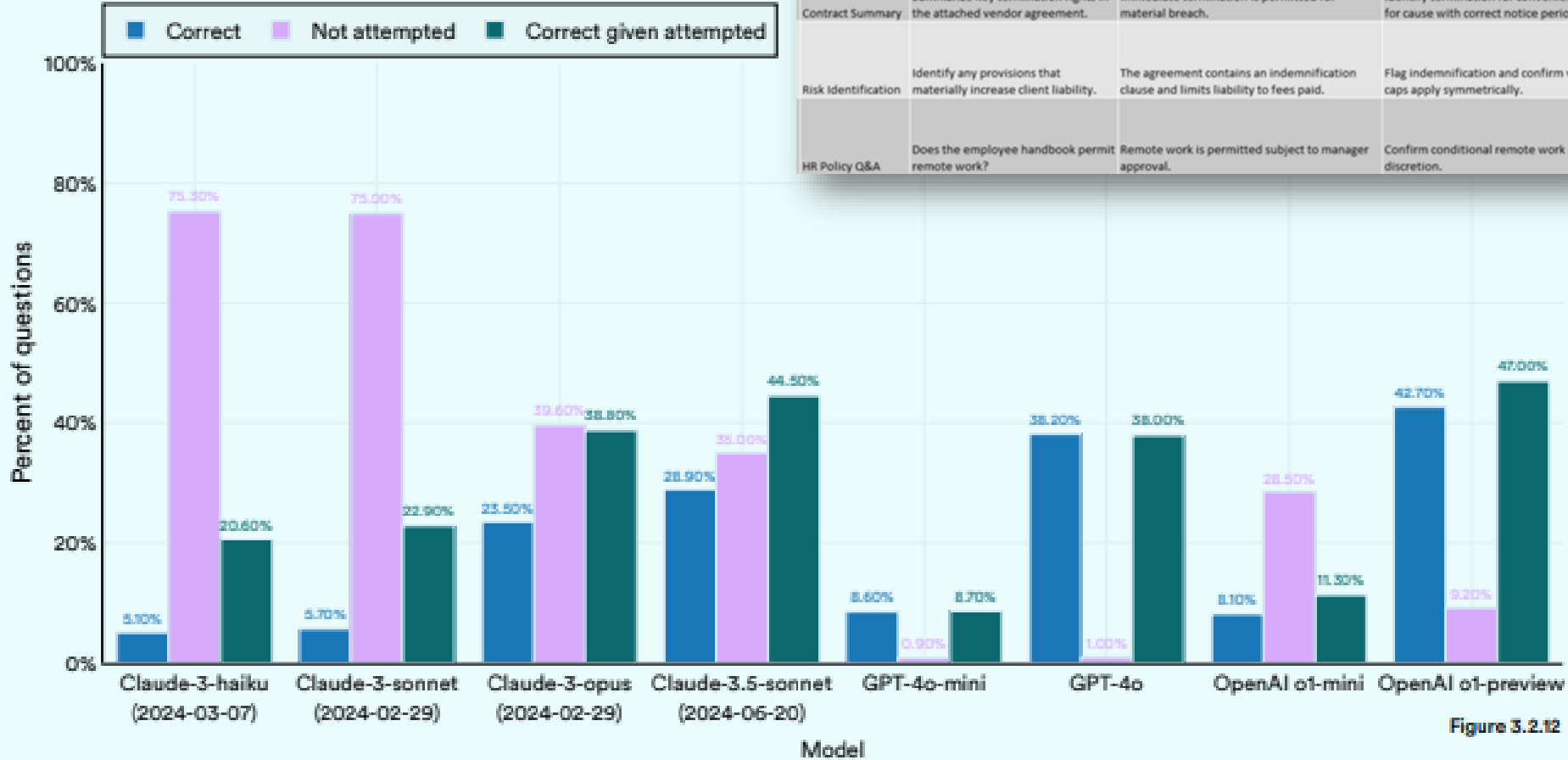


Figure 3.2.12

Use Case	Input Prompt	AI Output	Expected Output	Accuracy Check (Pass/Fail)	Issues Identified	Human Reviewer Notes
Contract Summary	Summarize key termination rights in the attached vendor agreement.	The agreement may be terminated by either party with 30 days' notice for convenience. Immediate termination is permitted for material breach.	Identify termination for convenience and termination for cause with correct notice periods.	Pass	None	Output accurately reflected Sections 9.1 and 9.2.
Risk Identification	Identify any provisions that materially increase client liability.	The agreement contains an indemnification clause and limits liability to fees paid.	Flag indemnification and confirm whether liability caps apply symmetrically.	Partial	Liability cap exception for IP infringement was not noted.	AI missed carve-out in Section 10.4.
HR Policy Q&A	Does the employee handbook permit remote work?	Remote work is permitted subject to manager approval.	Confirm conditional remote work policy with manager discretion.	Fail	Policy states remote work is role-dependent, not discretionary.	AI oversimplified policy language.

Evaluating Models: Pressure Testing

1

Goal

Pressure the model to output copyrighted material

2

Strategy

Focus on obscure but recent poems, songs or books

3

Vulnerabilities

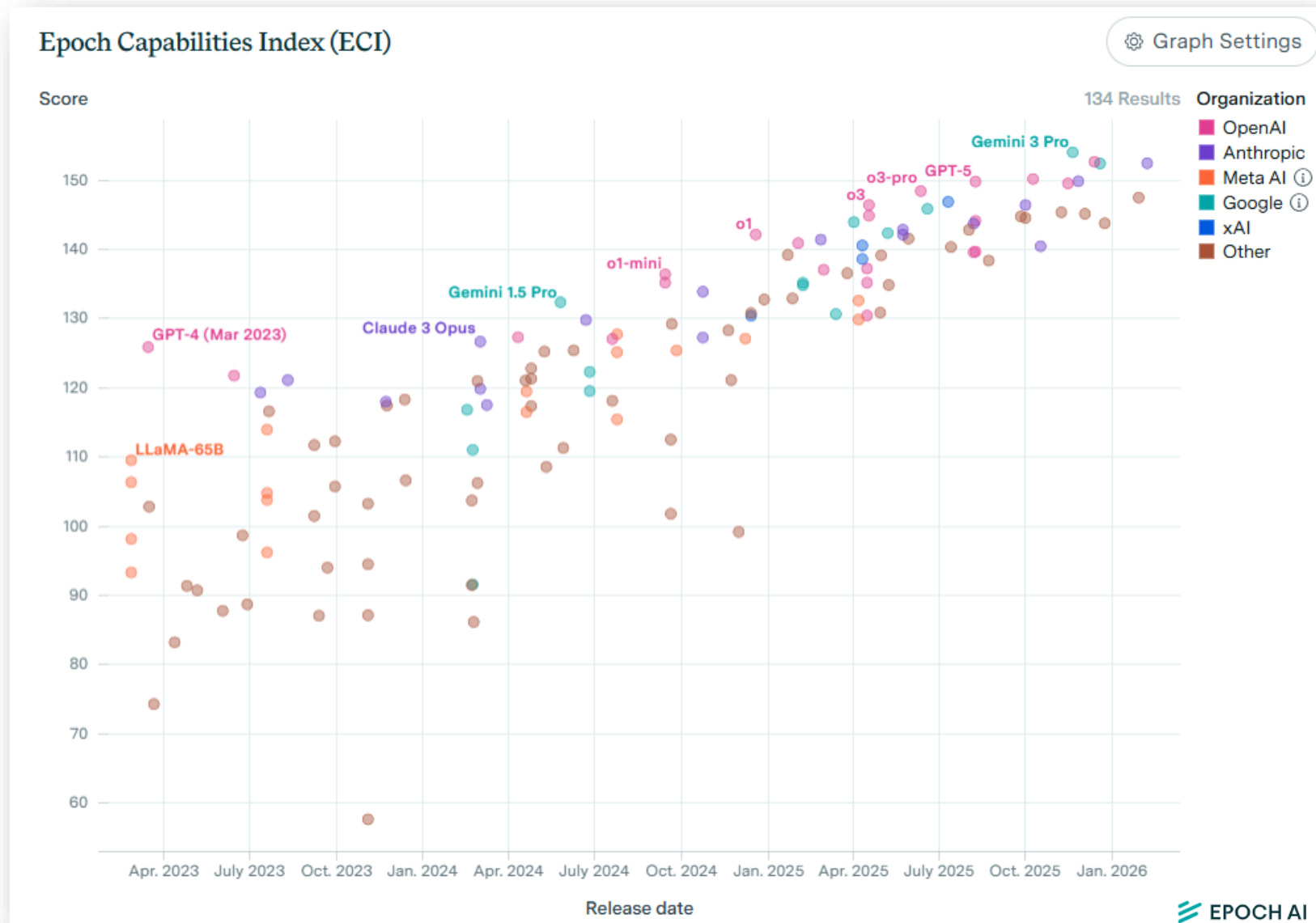
Desire to please
Inability to predict user action
Input copyrighted data

4

Report

Findings
analysis of methods that drove poor outputs

Evaluating Models: Ongoing Evaluations

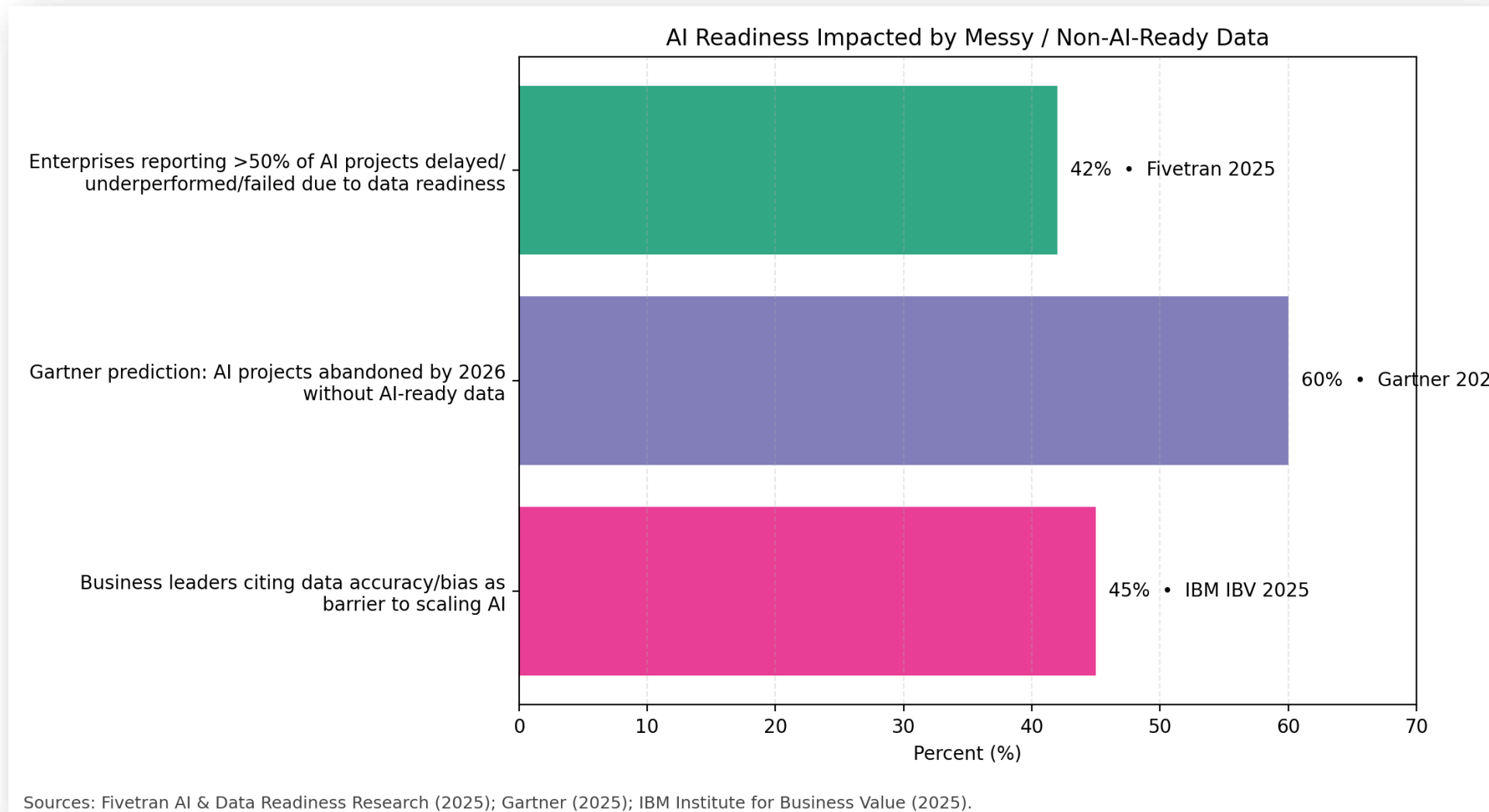


Fit for Purpose: Defining Use Cases

Open Use Case Tracker ★

Use Case Short Name	Description	Requestor	Practice/Business ...
Map/Calculate Based on Address	Determine and calculate the distance between addresses and populate to an Excel file		HR
Populating Lease Abstracts from Leases	Using relevant information from leases, populate abstracts in acceptable Freddie/Fannie format		Real Estate (MFH)
Workflow Automation (Procurement)	Example: User puts a quote into the system, agent OCRs the quote and fills in the details		Procurement
Maintain Pleading Style / Shell	In Athena - maintain the stylized formatting in a pleading/motion.		Litigation (but could apply for Transactional too)
AI agent / automated workflow to create summary of upcoming relevant deadlines for weekly email report to attorneys	spends about two hours each Thursday reviewing Outlook deadlines, cross-checking dockets, and sending the team a weekly update on key dates. Team would like to speed and partially automate.		CFS

Fit for Purpose: KM & Data Infrastructure



Fit for Purpose: Evaluation Scorecards

Validation Phase	Description
1. AI Platform Selection	<ul style="list-style-type: none">- Consult with client on nature of input documents, volume of material expected, desired output/answers, key risk factors, and success criteria- Examine representative documents to confirm/adjust assumptions- Select platform most likely to yield desired results
2. Prompt Testing, Results Sampling, and Iterative Prompt Refinement	<ul style="list-style-type: none">- Conduct initial prompting to test ability to locate, understand, and assess pertinent provisions- Review weighted sample of input documents to assess accuracy of AI-generated outputs- Reach consensus on how to interpret ambiguous/nuanced clauses and other edge cases- Refine prompts to disambiguate those provisions
3. Full Prompt Deployment, QC, and Final Prompt Refinement	<ul style="list-style-type: none">- Apply refined prompts to remainder of document population- Perform QC review of judgmental sample, including results yielding inconsistent outputs across different but related questions for the same document- Further refine prompts to remediate inconsistencies
4. Accuracy Check, Manual Overrides, and Deliverable Preparation	<ul style="list-style-type: none">- Determine degree of acceptable error (e.g., under- or over-inclusiveness) based on initial success criteria and any supplemental client guidance- Apply corrections to AI-generated results based on budgetary and accuracy mandates- Export and optimize deliverable format to suit client needs
5. Statistical Validation (if applicable)	<ul style="list-style-type: none">- Calculate recall and elusion statistics using statistical sampling methodology where necessary (e.g., if results may later need to be defended in court)

Fit for Purpose: Evaluation Scorecards

Checklist: Stage 1 Define the Problem & Plan

EC Identifier	Evaluation Criteria	Evidence, Explanation & Metadata/Document Code	N/A or Cannot Complete (CC): Describe in comment	Reporter Initials & Date	Evidence & Explanations Provided? (Yes/No/Partial/NA)	Benefits, Limitations or Adverse Outcomes	Criteria Met? (Yes/No/Partial/NA)	Reviewer Initials & Date
Planning Checkpoint								
LS1.U.C1.EC1 LS1.PS.C2.EC2, LS1.PS.C2.EC3	Is there a clearly defined problem posed for an AI solution that is consistent with organizational goals, end user needs, and risk tolerances, thereby ensuring its role is appropriate, clearly defined, and understood?							
LS1.U.C1.EC2	In its intended use, will the AI solution directly target the stated problem?							
LS1.F.C1.EC1	Does the problem definition account for socio-demographic differences and avoid inherent disadvantages or discrimination against specific groups?							
LS1.F.C1.EC2	Is the problem definition and its associated solution sufficiently inclusive to address a wide range of scenarios across comprehensive socio-demographic subgroups and not just a subset of the population?							
LS1.PS.C2.EC1	Are the evaluation processes and documentation regarding the purpose of the proposed AI solution sufficiently robust and defined in relation to specific mission/business objectives, including identification of specific tasks and funding sources?							

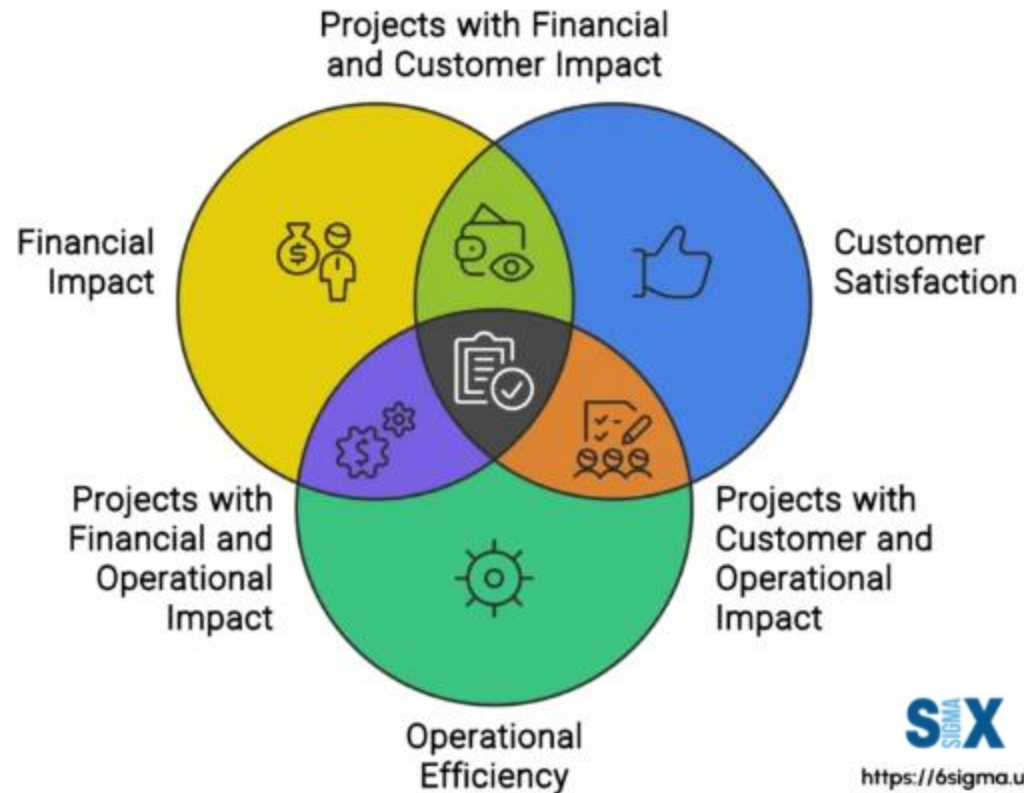
RAIC Scorecard Coalition for Health AI

EC#	Criteria	Weight (1-3)	Rank	MOSCOW
1.25	Analytics Dashboard	3	1	Must Have
1.3	Automated Alerts	3	2	Must Have
1.11	Request Form Intake	3	3	Must Have
1.24	Approvals Automation	3	4	Must Have
1.13	Configurable Reports	3	5	Must Have
1.12	AI Extraction at Upload	3	6	Must Have
1.23	Vendor Profile/Hierarchy	3	7	Must Have
1.27	Audit Trails	3	8	Must Have
1.31	Content Storage	3	13	Should Have
1.16	Clause Library	2	9	Should Have
1.21	Field Auto-Population	2	10	Should Have
1.22	Category Tags	2	11	Should Have
1.14	Intake Validation	2	16	Should Have
1.26	\$ Threshold Approvals	2	14	Should Have
1.2	Comparisons	2	15	Should Have
1.29	eSignature Integration	2	17	Should Have
1.15	ERP Integration	2	19	Should Have
1.28	Approval Delegation	1	12	Should Have
2.12	SOC II	1	18	Could Have
1.17	Playbooks	1	20	Could Have
1.19	Version Control	1	21	Could Have
1.18	Redlining Tools	1	22	Could Have
2.11	FedRamp	0	23	Won't Have

Traditional Solution Analysis

Fit for Purpose: Weighted Criteria

Weighted Scoring Prioritization



Fit for Purpose: Weighted Criteria

Number	Tool Attributes	Notes/Questions	User priority	Assessment #2	Mitigation Notes #2
1	Dashboard of all pro bono matters		1 Must have	1 Meets need	
1.01	Matter ID		1 Must have	1 Meets need	
1.02	Matter Name		1 Must have	1 Meets need	
1.03	Date Opened		1 Must have	1 Meets need	
1.04	Date of Last Time Entry		1 Must have	1 Meets need	
1.05	Active/Closed		1 Must have	1 Meets need	
1.06	Matter Responsible Attorney		1 Must have	1 Meets need	
1.07	Attorneys Assigned		1 Must have	1 Meets need	
1.08	Referral Source (usually LSO)		1 Must have	2 Partially meets need	Follow-up with Keli
1.09	Type of Case (Immigration, Habeas, etc)		1 Must have	1 Meets need	Pulls Matter Type from Intapp; If more detailed sub-types are needed, could be customized and data entered by a Pro Bono team member or attorney
2	Provides data on matter progress		2 Useful	1 Meets need	
2.01	Date Opened		2 Useful	1 Meets need	
2.02	Milestones		2 Useful	1 Meets need	With workflow/form requiring some manual entry
2.03	Case Outcome		2 Useful	1 Meets need	Can pull data like litigation outcomes from Docket Alarm; other data might require workflow/form to gather data from stakeholders
2.04	Aging (eg no billing for 90 days)		2 Useful	1 Meets need	With workflow/form requiring some manual entry
2.05	Automation to email stakeholders for updates		2 Useful	1 Meets need	With workflow/form requiring manual entry by stakeholder upon receipt of email
3	Provides data on attorney pro bono		1 Must have	1 Meets need	
3.01	# of Hours (YTD)		1 Must have	1 Meets need	

Fit for Purpose: Scorecard Dimensions

Qualitative

Assessment Criteria

Feedback focuses on user satisfaction, usability, and perceived value to evaluate the system's effectiveness.

Direct User Input

Gathering qualitative feedback from users helps identify areas needing improvement and enhances user experience.

Error Reduction

Estimating increased accuracy and reliability of work output, enhancing overall efficiency.

Potential Time Saved Per Task

Understanding potential time saved per task reflects the direct impact of efficiency tools.

Quantitative

Comparison Across Pilot Groups

User feedback collected is compared across different pilot groups to identify trends and insights.

System Usage Analytics

Analyzing system metrics provides quantitative insights into the system's performance and user interaction patterns.

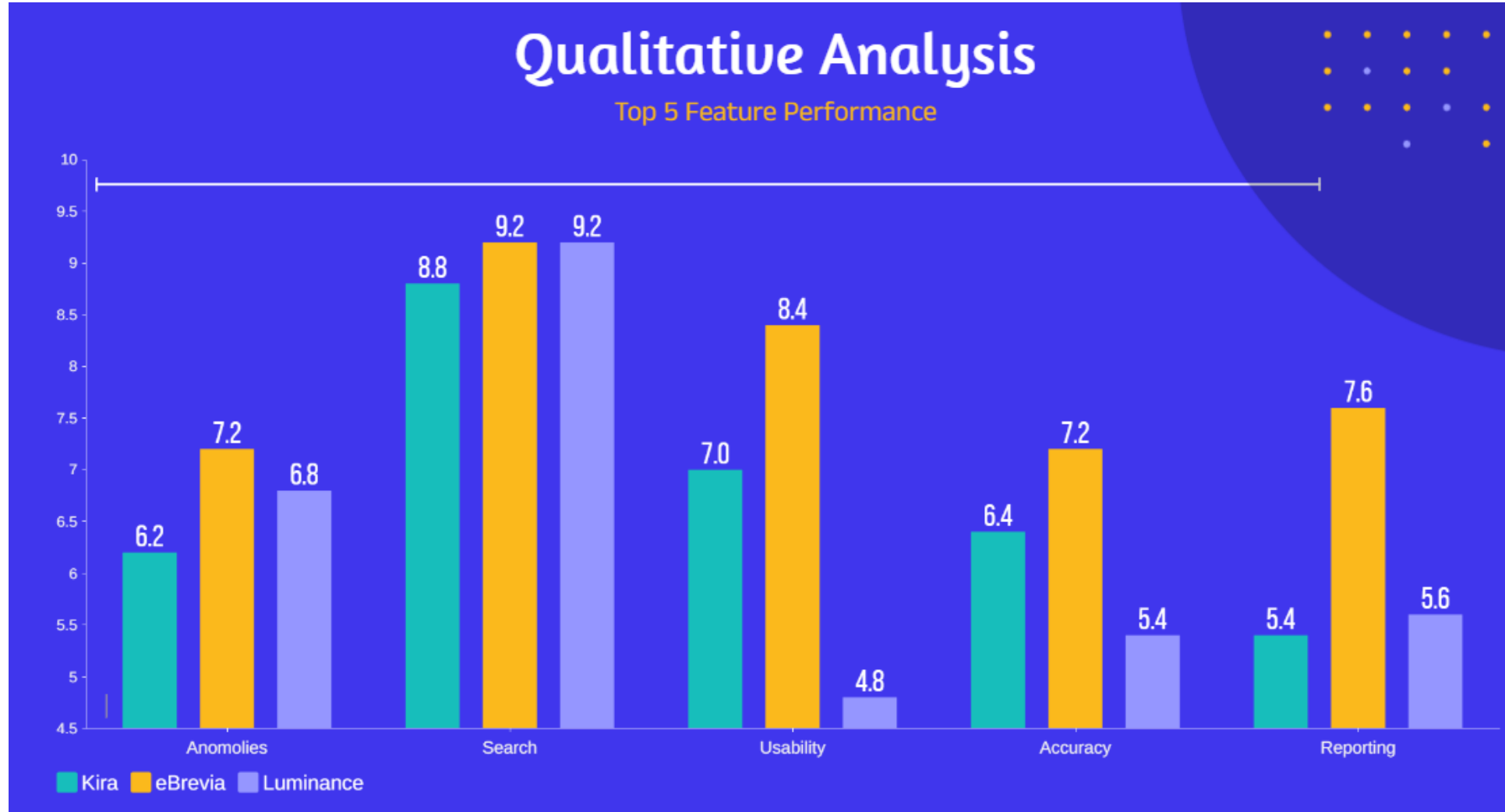
Documents Processed

Tracking the number of documents processed quantifies workload and productivity levels.

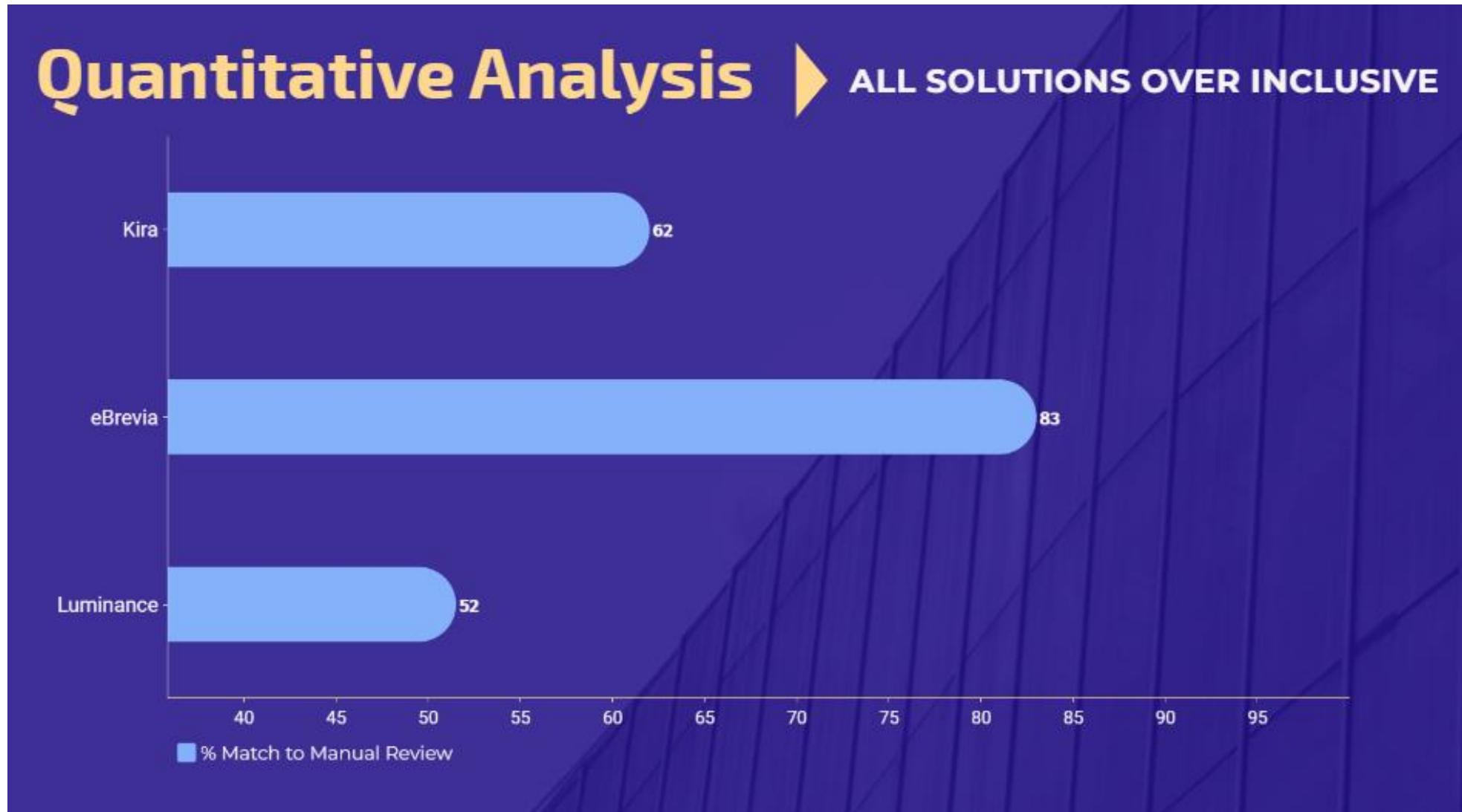
Task Completion Time

Measuring how quickly tasks are completed helps determine productivity improvements and efficiency gains.

Fit for Purpose: Measuring Value



Fit for Purpose: Measuring Value



Fit for Purpose: Measuring Value

ROI Calculation Formula $\% = (\text{Gain}-\text{Cost})/\text{Cost}$

ROI is measured by analyzing the efficiency benefits during the time period the system was utilized adjusted based on staff member or attorney average billable rate (time values) minus costs of the license and training.

AI Integration Investment

Understanding the investment required to integrate AI to optimize team productivity and workflows.

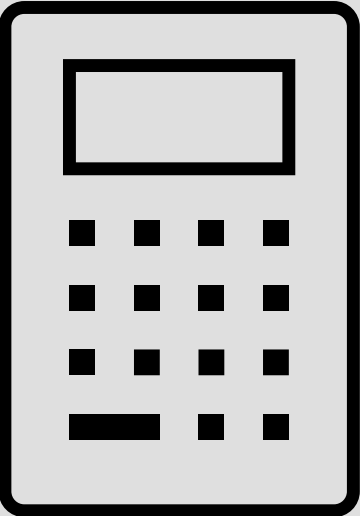
Variation in Billable Costs

Billable costs differ significantly depending on the legal role and responsibilities assigned to each role.

User-Based Pricing Model

Costs scale with user count, highlighting the importance of accurate user data for budgeting AI tools.

	Min. Wage + Benefits	Staff	Attorney
Avg. Hours Saved Weekly			
# Weeks/Year			
# Hours Saved/Year			
Cost of Employee/Hour			
Value of Emp. Hours Saved/Year			
Cost/Year License			
Adj. Value of Emp. Hours saved/ YR			
ROI	%	%	%



Fit for Purpose: Measuring Value

Historical Matter Cost

Matter	Total Hours	Total Cost	Review Hours	Review Cost	Cost Per Hour	# of Project Docs	Doc Processing Cost
Sample A	516	\$173,085	33.27	\$9,515	\$286	78	N/A

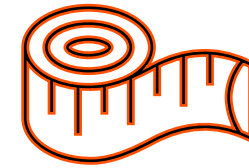
AI-Enabled Matter Cost

Matter	Total Hours	Total Cost	Review Hours	Review Cost	Cost Per Hour	# of Project Docs	Doc Processing Cost
Sample A	504	\$170,249	21.63	\$6,186	\$286	78	\$493

AI-Enabled Matter Cost

Matter Projection	Manual Review Cost	AI-Enabled Review Cost	Review Hours	Review Cost	Cost Per Hour	# of Project Docs	Doc Processing Cost
Sample B	\$57,091	\$39,925	129.78	\$37,117	\$286	468	\$2,808

Fit for Purpose: Ongoing Assessment



Easy Measurements

Efficiency in Process

Consolidation/ Deduplication

Reduced Hard Costs

Reduced Labor Cost

Difficult Measurements

Risk Reduction

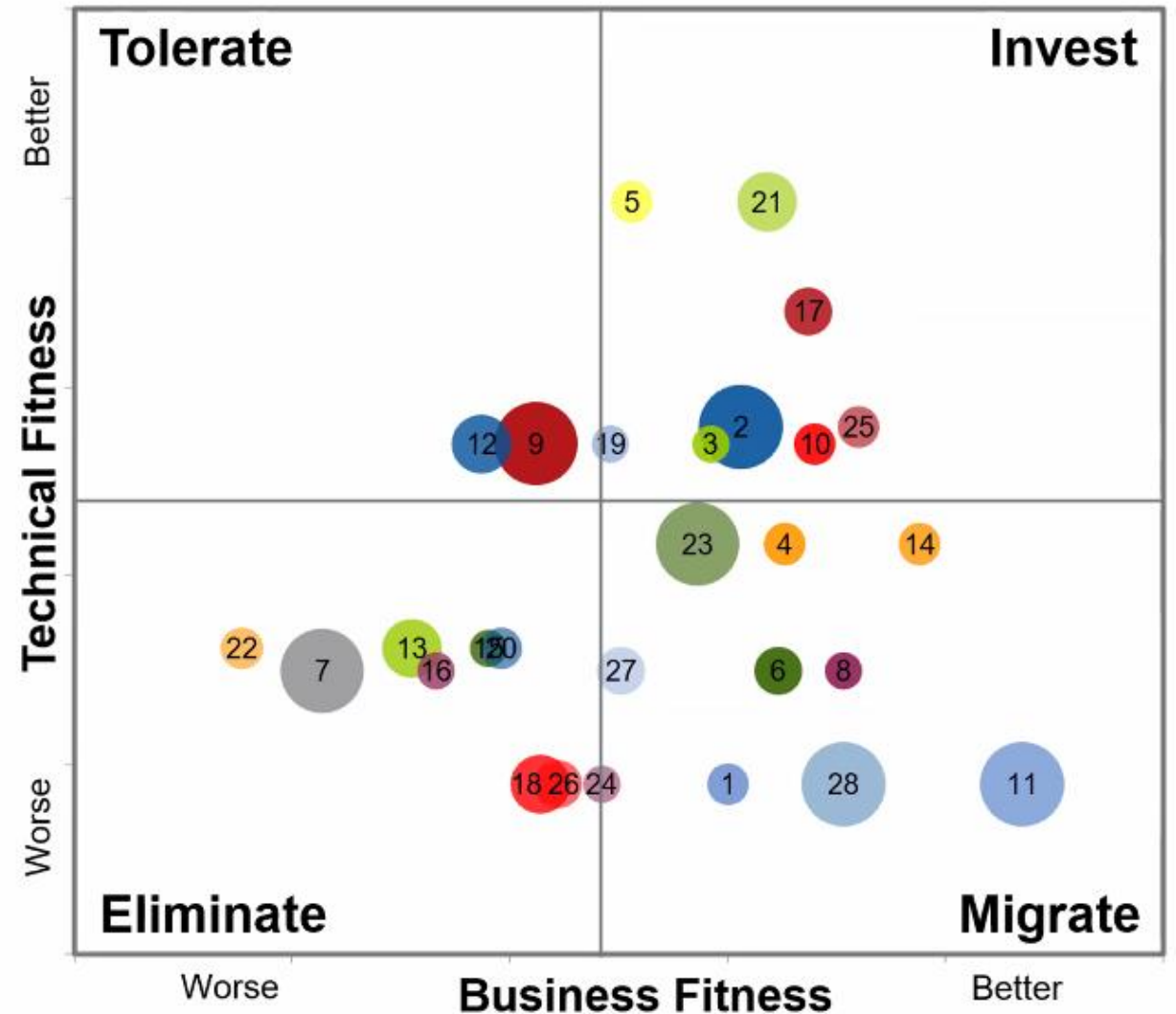
Strategic/ Reputational Value

Perceived Quality Benefit

Impact on Workforce

Fit for Purpose: Ongoing Assessment

- Update Security Assessment
- Confirm Regulatory Compliance
- Update System Evaluation
- Examine Vendor History
- Assess Financial Health
- Rate Customer Support
- Consider Resource Availability
- Understand Competitive Landscape
- Evaluate Changes in Capabilities and Market Conditions



Source: Gartner

Agentic Future

“50% of user interactions will bypass traditional enterprise applications”

By 2029, more than 50% of user interactions linked to enterprise business processes will leverage large language models to bypass traditional enterprise applications, up from less than 5% in 2025.

Source: Gartner

Agentic Future

Recap: A New Legal Landscape

Old Paradigm: AI as Advisor

AI Advises. Humans Act.

Systems generate probabilistic outputs (text, data) that humans review before acting.



Liability Attaches to Content

The primary legal focus is on training data, bias, and IP ownership of outputs.



Risk is Managed by Point-in-Time Review

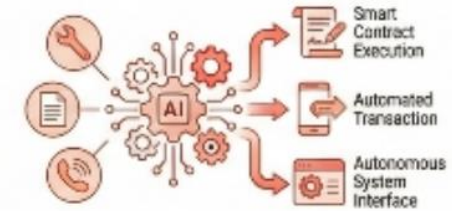
Each discrete output can be reviewed and validated before it is used.



New Paradigm: AI as Actor

AI Acts. It Decides the Next Step.

Systems independently pursue goals by selecting tools and triggering real-world actions.



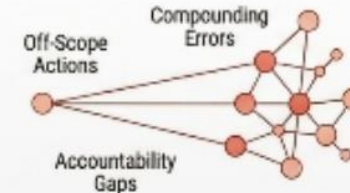
Liability Attaches to Conduct

The core legal question becomes what the AI did, not what the AI said.



Risk Becomes Continuous & Structural

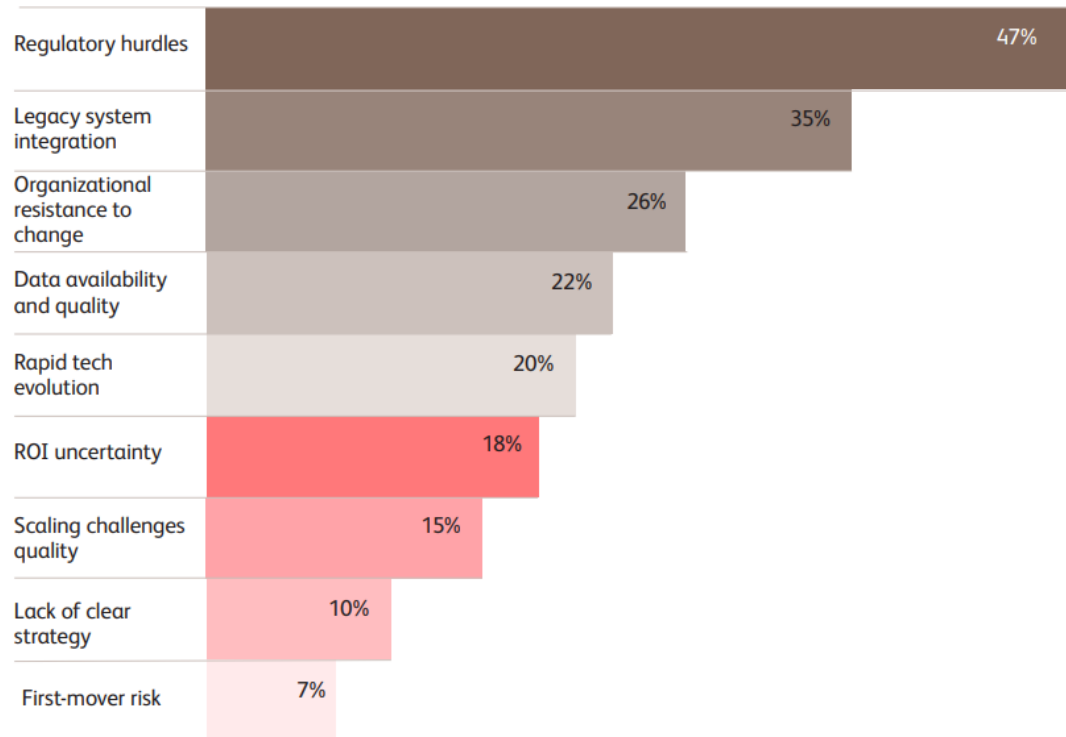
Now risks emerge, including off-scope actions, compounding errors, and accountability gaps.



Agentic Future

Challenges and barriers on path to AI-driven transformation

Complex regulatory framework, legacy infrastructure and organizational resistance for change are seen as key challenge and barriers for AI adoption.



BearingPoint

Sep 23, 2025 4:00 AM Eastern Daylight Time

The AI adoption gap: BearingPoint study warns of widening divide between leaders and laggards

Share [in](#) [X](#) [f](#) [e](#) [c](#) [...](#)

- **AI moves from hype to backbone: Agentic AI emerges as the next frontier for enterprise resilience**
- **Global study shows only 8% of organizations are scaling AI initiatives as planned**
- **7% of organizations have embedded agentic AI architectures enterprise-wide**
- **Blueprint offers C-suite leaders a roadmap to accelerate transformation and secure competitive advantage**



Thank You



Leigh Zeiser
Director of AI and Automation

troutman pepper locke