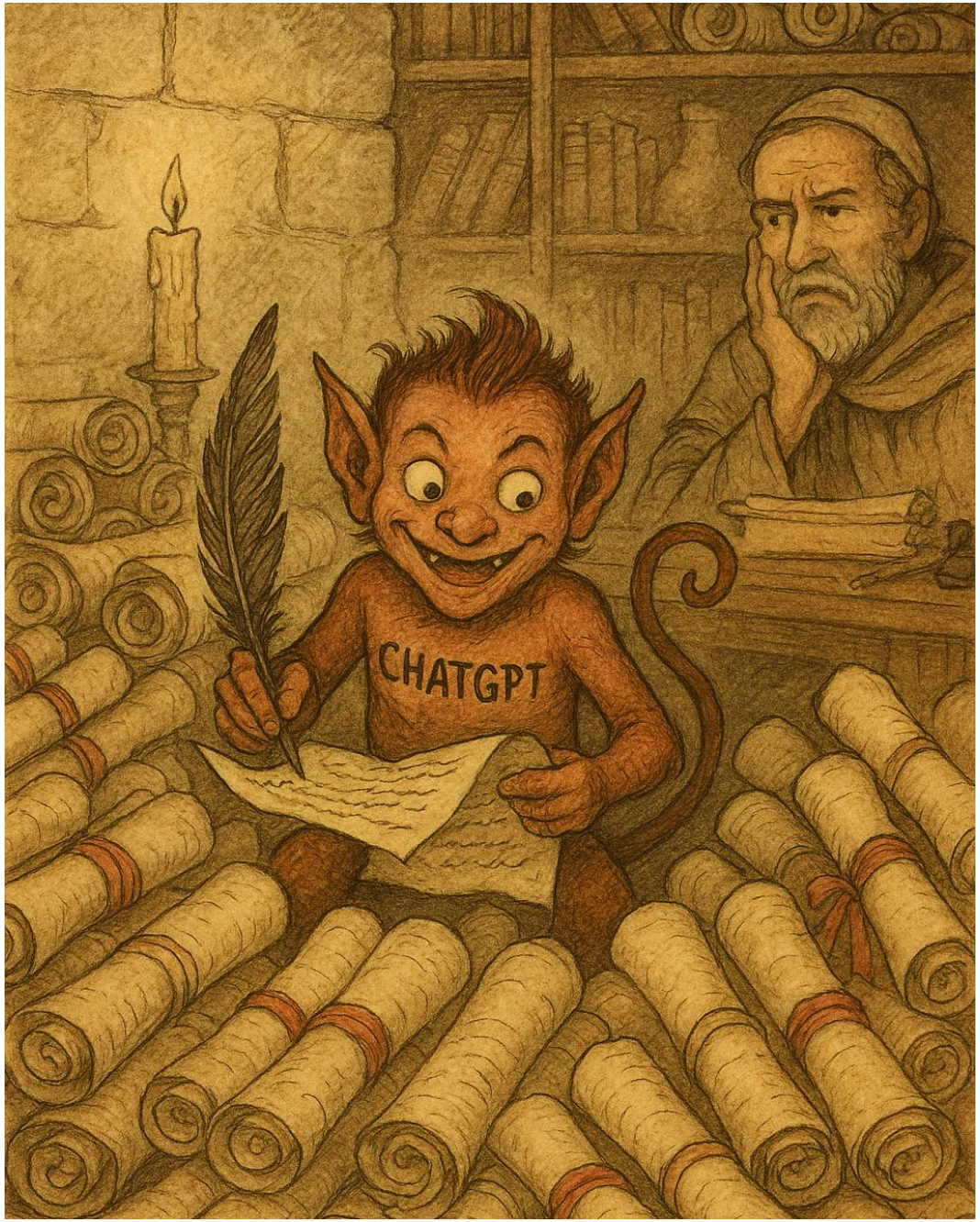


Waking Up to AI

An Adventure in Governance

Brian Myers

with assistance from

Brian Myers PhD, CISSP, CCSK

Experience

- 20 years in software development
- 10 years in information security

Past Positions

- Director of InfoSec, WebMD Health Services
- Senior AppSec Architect, WorkBoard
- Senior Risk Advisor, Leviathan Security

Current Work

- Independent Information Security Consultant

Volunteer

- OWASP AppSec Days PNW (2021-2024)
- Western Oregon University CS Advisory Board

safetylight.dev

Brian speaks at conferences, companies, chapter meetings, and universities, and would be happy to present at your organization too. More details about all Brian's talks are on GitHub.

Selected venues: ISACA, OWASP, BSides Seattle/Portland/Idaho, Oregon Cyber Resilience Summit, Western Oregon University, Technology Association of Oregon, PNSQC, and others.

Waking Up to AI: An Adventure in Governance

A fictional SaaS company's messy, revealing journey through AI risks, missteps, and gradual governance. *[details] [slides]*

Kidnapping a Library: How Ransomware Taught the British Library to Follow Well-Known Best Practices

A cautionary tale about how a ransomware attack crippled a major cultural institution and the measures taken to recover. *[details] [slides] [video]*

Resources

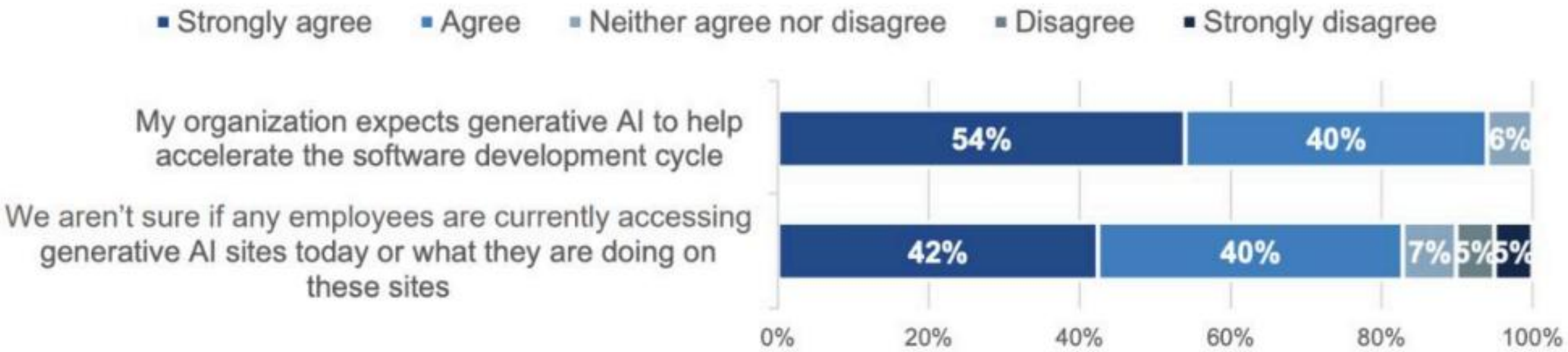
Here are all the links to resources mentioned in the talk (and a couple of extras as well.)

AI Policies and Governance

- **Drata**: Policy and Plan Guidance
- **Harmonic**: AI Policy Generator
- **Fairnow**: AI Governance Framework
- **ISACA**: Policy Template Toolkit
- **OECD AI Principles**: OECD AI Principles overview
- **SANS**: Security Policy Project
- **US Congress**: Advancing American AI Act and Executive Order 14110
- **World Economic Forum**: Adopting AI Responsibily

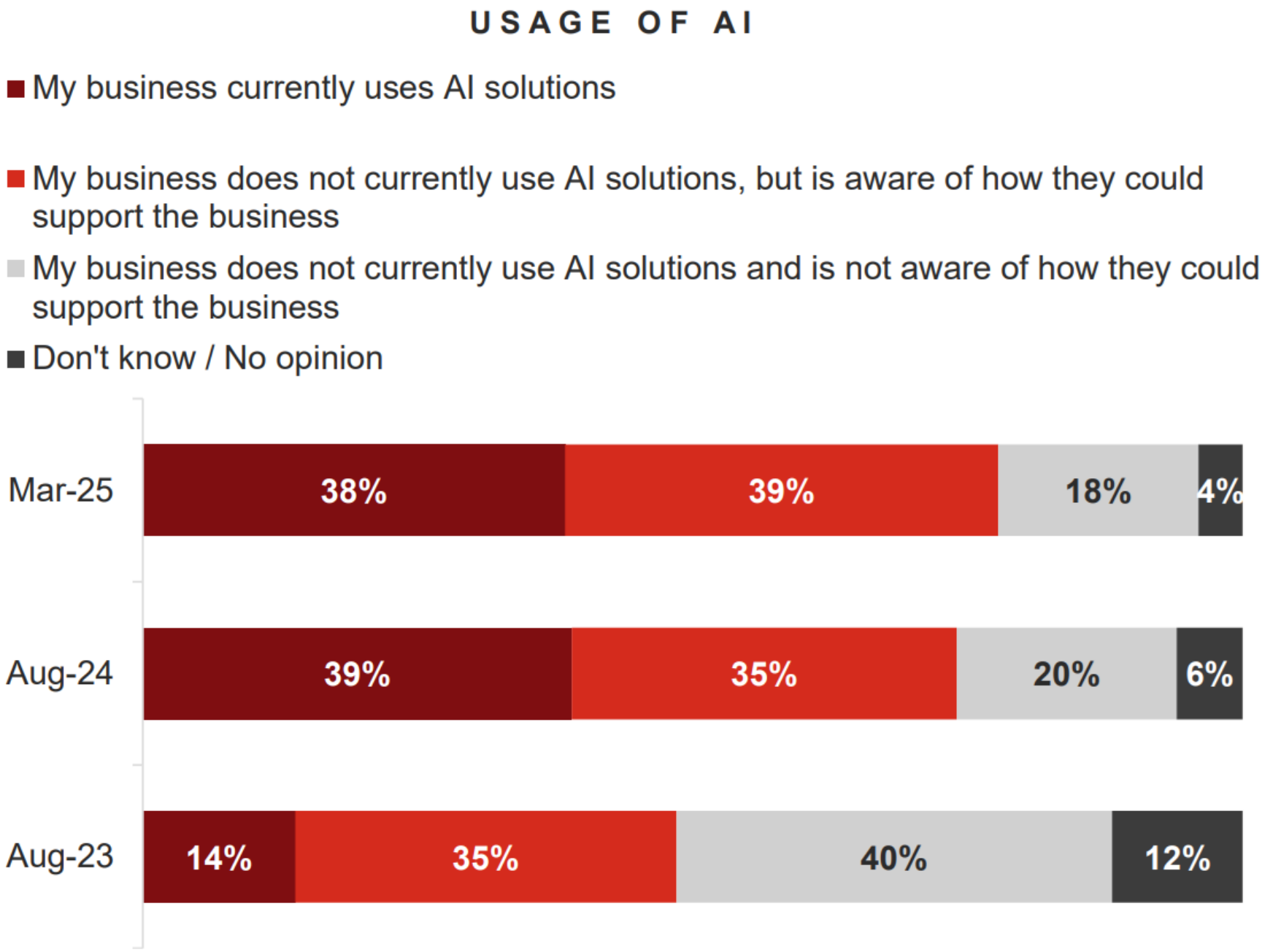
Shadow AI & Adoption

- **Axios**: What's New and What's Next
- **Help Net Security**: Your employees uploaded over a gig of files to GenAI tools last quarter
- **KPMG**: Trust, Attitudes, & Use of AI
- **McKinsey**: The state of AI: How organizations are rewiring to capture value
- **Nudge**: AI Adoption Curve

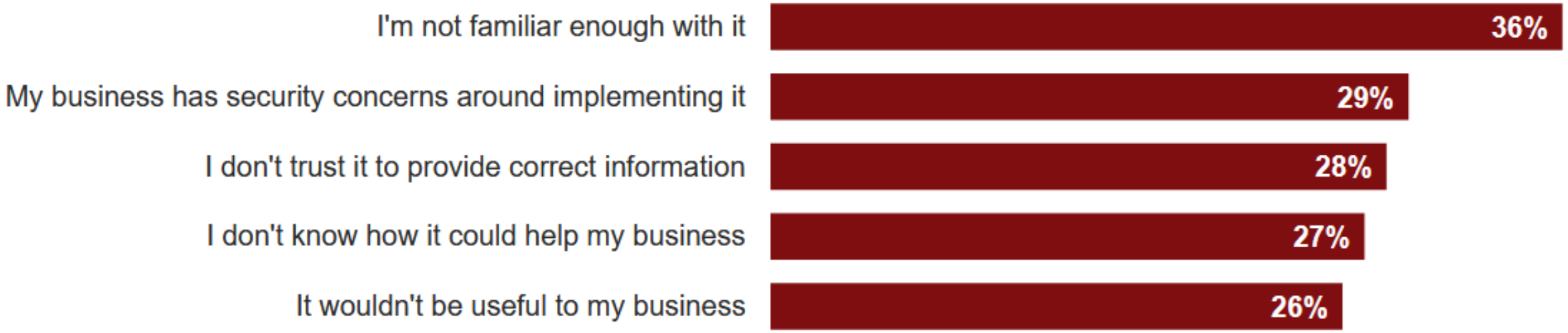


Source: Enterprise Strategy Group, a division of TechTarget, Inc.

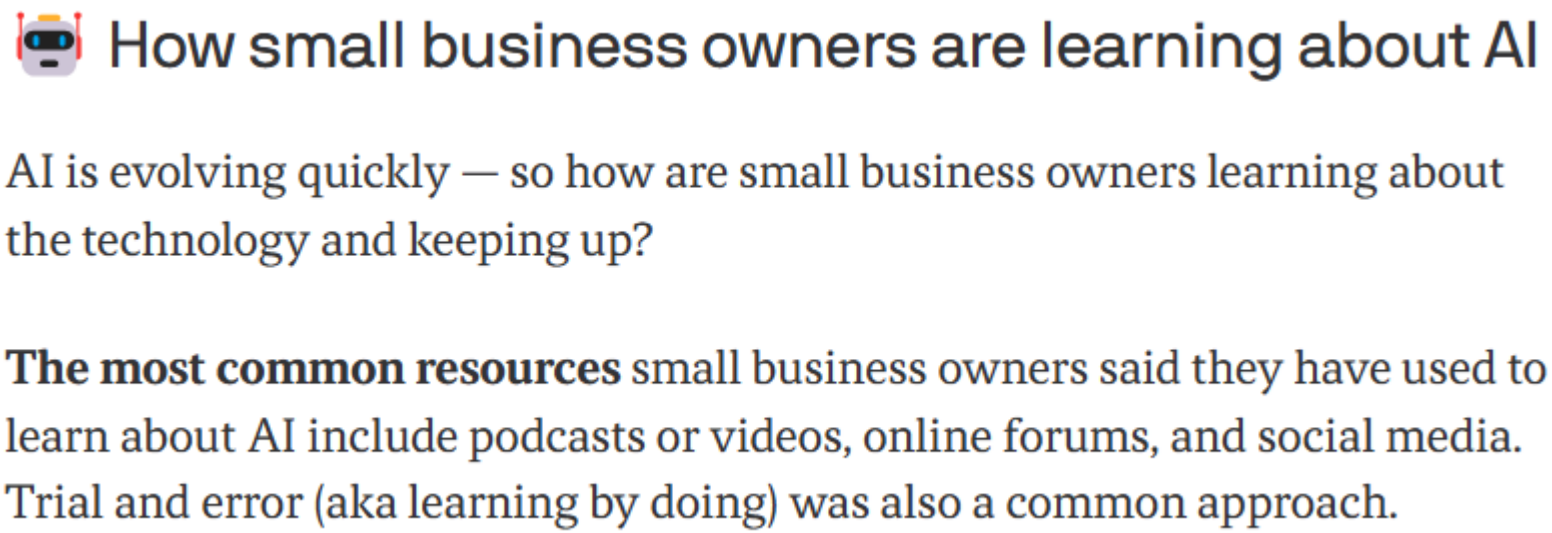
Verizon State of Small Business Survey 2025



Why Does Your Business Not Use AI?



What's new and what's next: How small business owners are using AI



How small business owners are learning about AI

AI is evolving quickly — so how are small business owners learning about the technology and keeping up?

The most common resources small business owners said they have used to learn about AI include podcasts or videos, online forums, and social media. Trial and error (aka learning by doing) was also a common approach.

AXIOS

Mar 11, 2025

What This Workshop Is

A scenario-based walkthrough of a (fictional) small company's growing awareness of AI risk:

- A survey of AI-related risks.
- An opportunity to discuss together difficult problems we all face.
- A picture of how many companies actually run a security program.

A tale that may be exemplary or cautionary. You'll have to decide.

Why Do It This way?

- Situate AI risk in concrete situations, not abstract lists.
- Consider how to act when standards don’t yet exist.
- Companies don’t publish their partial fixes and hard tradeoffs—but that’s where the most useful lessons often lie.

Once upon a time...

- LMS for professional education
- SaaS platform on AWS

- ≈25 staff
- No security or AI experts
- SOC 2

A customer asks a question...

PRELUDE

RFP From BigBux

...

4.4.2. Information Security

a. List your current security certifications (e.g., ISO 27001, SOC 2 Type II).
b. Provide a recent penetration test summary or redacted report.

...

4.4.3. Artificial Intelligence Governance

a. Does your organization use AI in the product? If so, please describe the use cases.
b. Do you have an internal AI governance policy? If yes, please provide a summary or table of contents.

...

Maxine Powers
CEO

Archie Tech
CTO

BigBux asks if we have an AI policy. Do we?

No. We govern our AI by not having any.

Let's make one so we can say Yes.

AI Policy Examples

- Drata Policy & Plan Guidance
  - No AI content as of September 2025
- SANS Security Policy Project
  - Artificial Intelligence Acceptable Use
  - 14 pages

AI Policy Examples

- ISACA Policy Template Toolkit
  - Artificial Intelligence (AI) Acceptable Use
  - Archie's not a member
- Generate Your AI Usage Policy in MInutes
  - Just came out Sep 4

Archie Tech
CTO

Drew Diligence
Counsel

I wrote: ‘MindPath is committed to using AI fairly.’ Does that sound governance-y enough?"

Add a line about reviewing AI use with the security officer. That’s you.

Right. So I just talk to myself if it ever comes up?

AI Governance Policy

Scope

This policy applies to all personnel, including employees and contractors.

Responsible Use

MindPath is committed to using artificial intelligence (AI) in ways that are fair, ethical, and compliant with applicable laws and regulations.

Product Use Requires Approval

Any use of AI in MindPath's products or services must be reviewed and approved in advance by the Security Officer.

Policy Review

This policy will be reviewed at least once per year, or sooner if there are significant changes in AI-related risks.

AI: Governed!

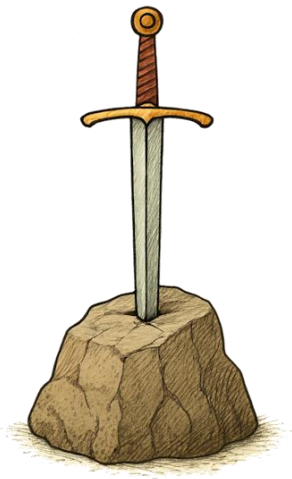
Has Archie addressed the problem?

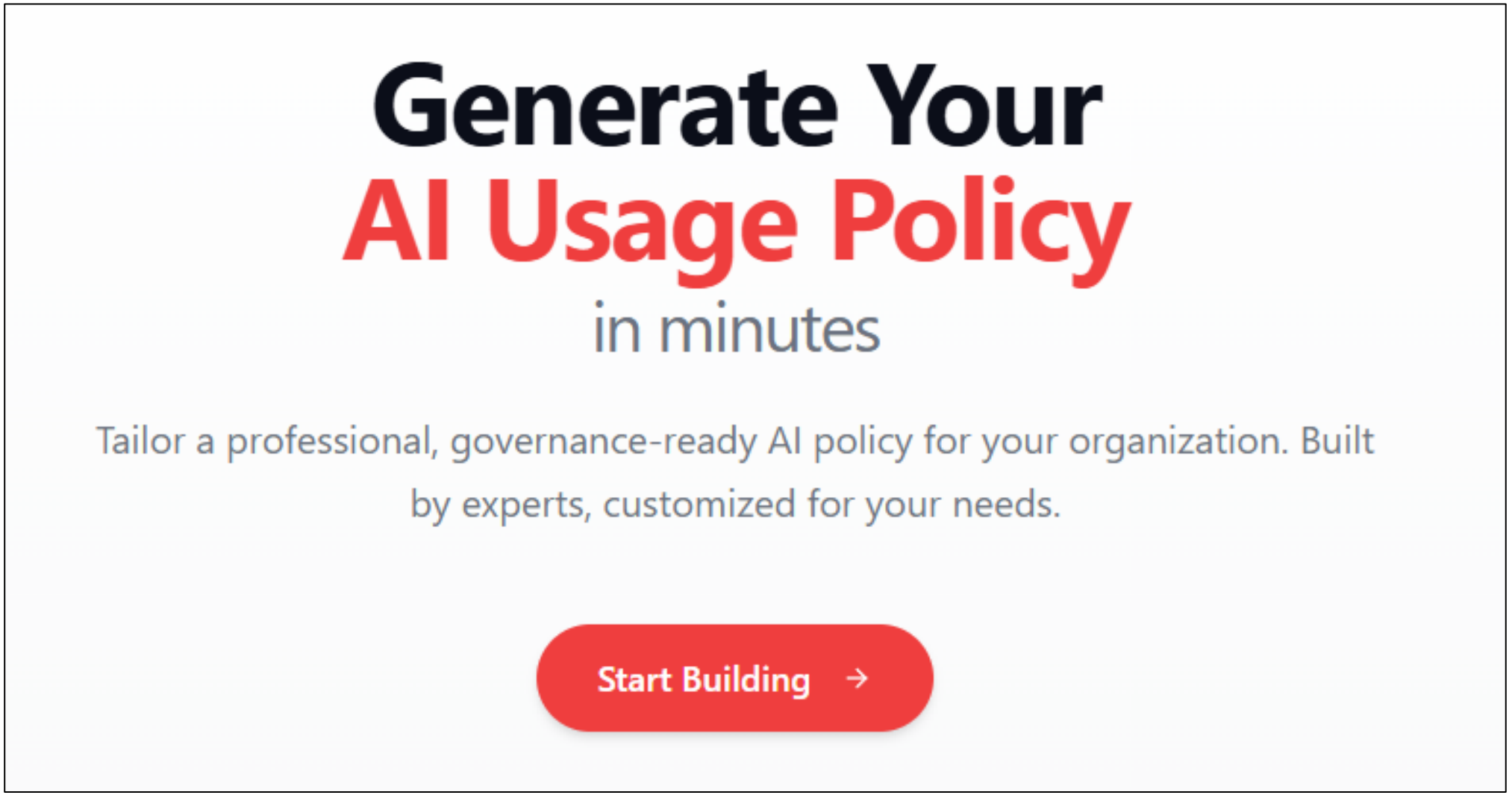
PRO	CON

• Commitment to fairness and compliance
• Gate for use of AI in product
• Accountable role
• Periodic policy review
	• Ignores AI outside of product.
	• May not satisfy BigBux.
	• No defined process.
	  – No artifact of compliance
	  – Efficacy uncertain
	• “AI” is very broad…
	• Treats AI as a technology owned solely by CTO.
	• Annual review may not suffice.

What Makes Good Governance?

Awareness	Did they spot and prioritize the real risks?
Effectiveness	Did they make things safer?
Efficiency	Was the cost and effort proportional to the benefit?
Fit	Did it suit the company's culture, size, and maturity?
Assurance	Does it align with best practices? Is it auditable?





harmonic

Published Sep 4 2025: https://ai-policy-studio.com/

A developer debugs some code...

A MISHAP

Transcribio

We provide fast, accurate audio transcription to support clear communication, accessible content, and professional workflows.

Cody Commit
Software Engineer

Search

#dev-eng

Messages Files Pins +

Tuesday, April 22nd

Cody Commit

Hey y'all—heads up on the video transcription bug we were seeing! Turns out the issue was with the way we were passing the audio url to Transcribio. The signed URL was expiring before the job kicked off. I threw a minimal repro into ChatGPT and it totally nailed it. Here's the snippet:

```
transcription_request = {
    "audio_url":
"https://videos.mindpath.io/p/4839.mp4?e=1714526400&s=abf82c7e",
    "language": "en-US",}
headers = {"Authorization": "Bearer sk_prod_23af20c8f4c14b1a90f88f8d0a9e"}
response = requests.post("https://api.transcribio.com/v1/transcribe",
      json=transcription_request, headers=headers)
log(response.status_code)
```

Secure Coding Policy

Managing Secrets

Secrets include private encryption keys as well as authentication credentials. Access to such secrets must be governed strictly according to the Principle of Least Privilege, ensuring that only people with a legitimate business need can know what the secrets are.

Incident Report

Severity	Low
Description	External API auth secret exposed over the web to an external company and in internal Slack.
Details	Developer pasted code containing an API auth secret into ChatGPT for debugging assistance. Secret could have been viewed by OpenAI employees and may persist in OpenAI storage and logs. Could also be used in AI training.
Remediation	Transcribio invalidated our auth credential and issued a new one which we deployed it on the same day. Transcribio confirms no one sent unauthorized work using the old API key before it was invalidated. Met with engineers to remind them of Secure Coding Policy.

Would you rate that incident Low?

PRO	CON
• Incident contained • Incident artifact created • Staff reminded of responsibility not to share secrets	• No monitoring or auditing to prevent a recurrence • Staff may still be sharing other sensitive data with AI • Unassessed risk.

Is Training Data Extraction a Thing?

Dropbox…demonstrated extraction of memorized training data from both GPT-3.5 and GPT-4.

- Bye Bye Bye…: Evolution of Repeated Token Attacks on ChatGPT Models
  Breitenbach & Wood, Dropbox, 2024

…for non-targeted PII extraction, the attack success rate reaches 48.9%, extracting one authentic PII per two queries at a cost of $0.012 per PII.

- Effective PII Extraction from LLMs through Augmented Few-Shot Learning
  Cheng et. al., Arxiv.org, 2025

In its suit, the Times alleges that, when prompted by users, ChatGPT sometimes spits out portions of its articles verbatim…

- ChatNYT, *Rachel Reed, Harvard Law Today, 2024*

“From Payrolls to Patents:
The Spectrum of Data Leaked Into GenAI”

We analyzed tens of thousands of prompts going into ChatGPT, Copilot, Gemini, Claude, and Perplexity...in Q4 2024.

8.5% of prompts into GenAI include sensitive data.

harmonic

Type	Frequency
Customer Data	45.77%
Employee Data	26.83%
Legal and Finance	14.88%
Security	6.88%
Sensitive Code	5.64%

harmonic

Risk Register

Risk	Description	Likelihood	Severity	Risk Level
Unmanaged AI Adoption	Well-meaning staff adopt AI tools without review and share data with them, introducing a new risk of data exposure along with others including AI hallucinations, data exposure, and loss of oversight.	High	Medium	High

Archie goes exploring.

RECONNAISSANCE

To: All Staff
Subject: Help Us Understand AI Use at MindPath

Hi everyone,

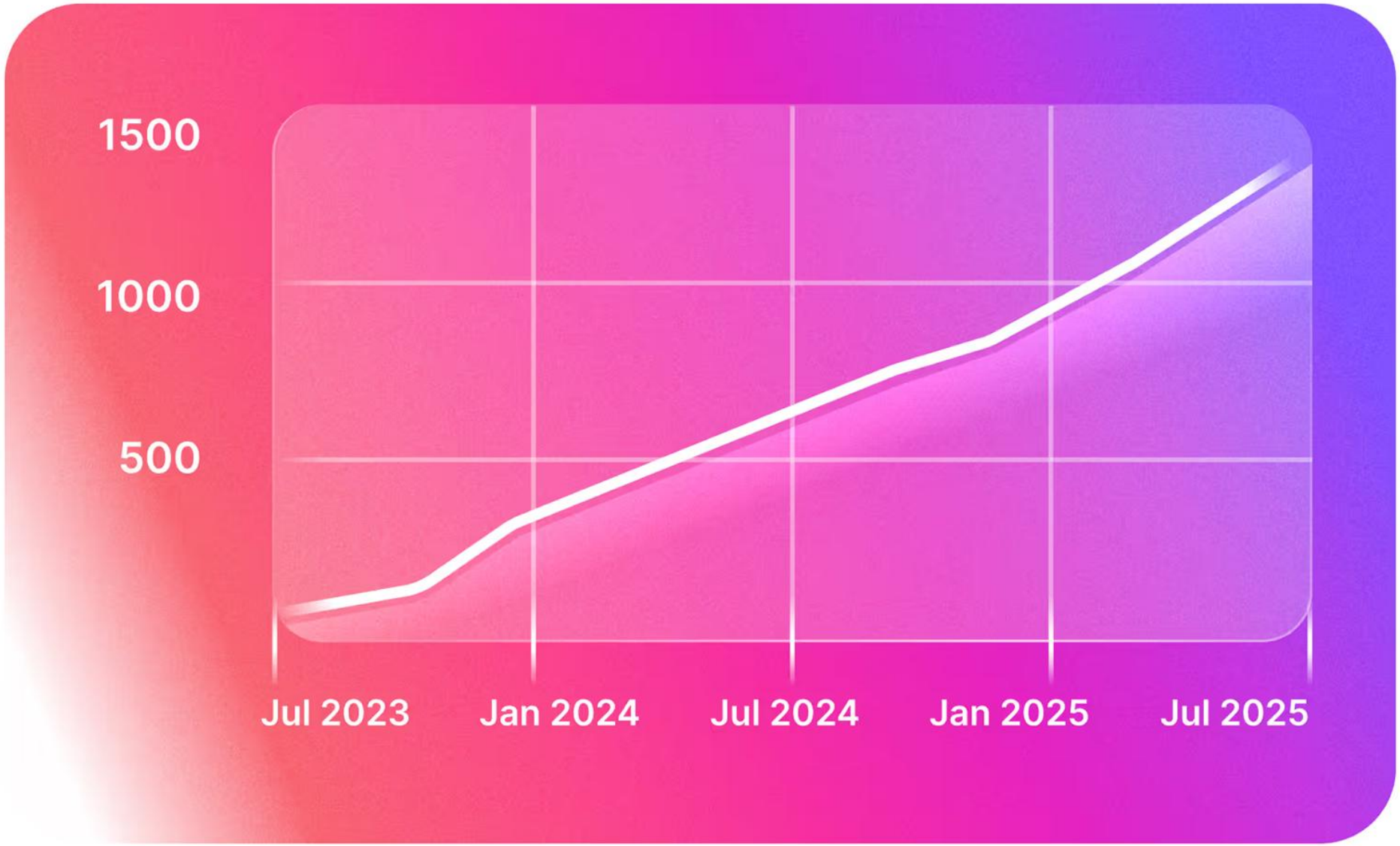
As AI tools become more common, it's important we understand how they're being used at MindPath—especially as individuals explore them on their own.

If you've used any AI-powered tools (like ChatGPT, GitHub Copilot, or others) for work-related tasks, please let me know. I just want to get a clear picture so we can think through the opportunities and risks together.

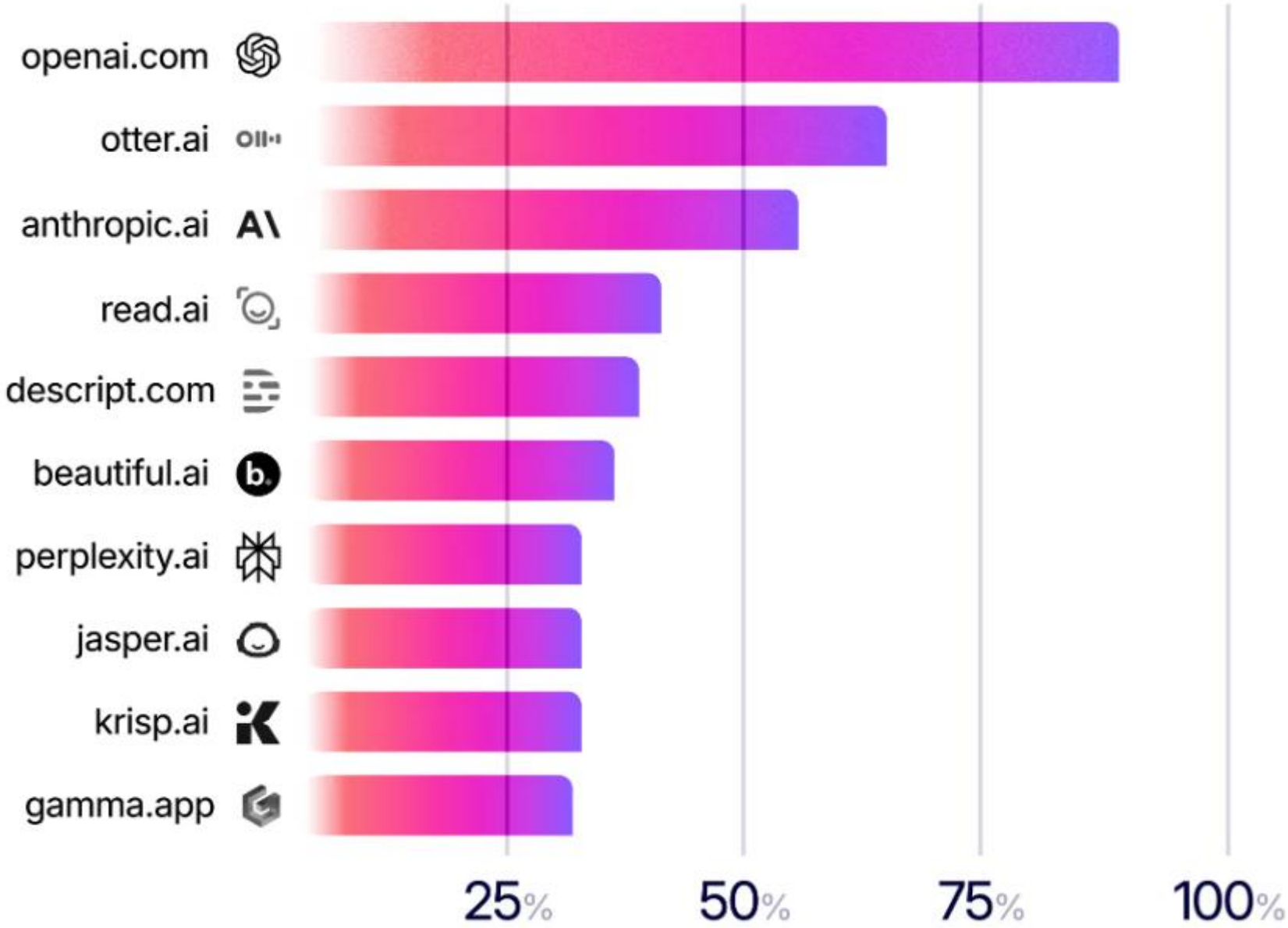
Thanks,

Archie Tech, CTO

Growth in unique AI tools discovered



Top 10 most-adopted AI tools



Percent of organizations that have adopted popular AI SaaS tools, based on product data from Nudge Security

...the average enterprise saw 23 previously unknown GenAI tools newly used by employees.

GenAI Data Exposure: What GenAI Usage Is Really Costing Enterprises

10/22/2024

GO TO OVERVIEW

Half of all employees are Shadow AI users, new study finds

↗ 75% of knowledge workers already use AI

↗ 46% wouldn't give it up, even if it were banned

Inventory of AI Use Cases

Role	Task	AI Tool
Developer	Debugging code	ChatGPT
CTO	Drafting security policies	Claude
Various Staff	Making memes	DALL·E
HR	Drafting interview rejections	ChatGPT
Intern	Reformatting webinar transcript	Wordtune

Trust, Attitudes and Use of Artificial Intelligence (2025)

48% fear being left behind if they do not use AI

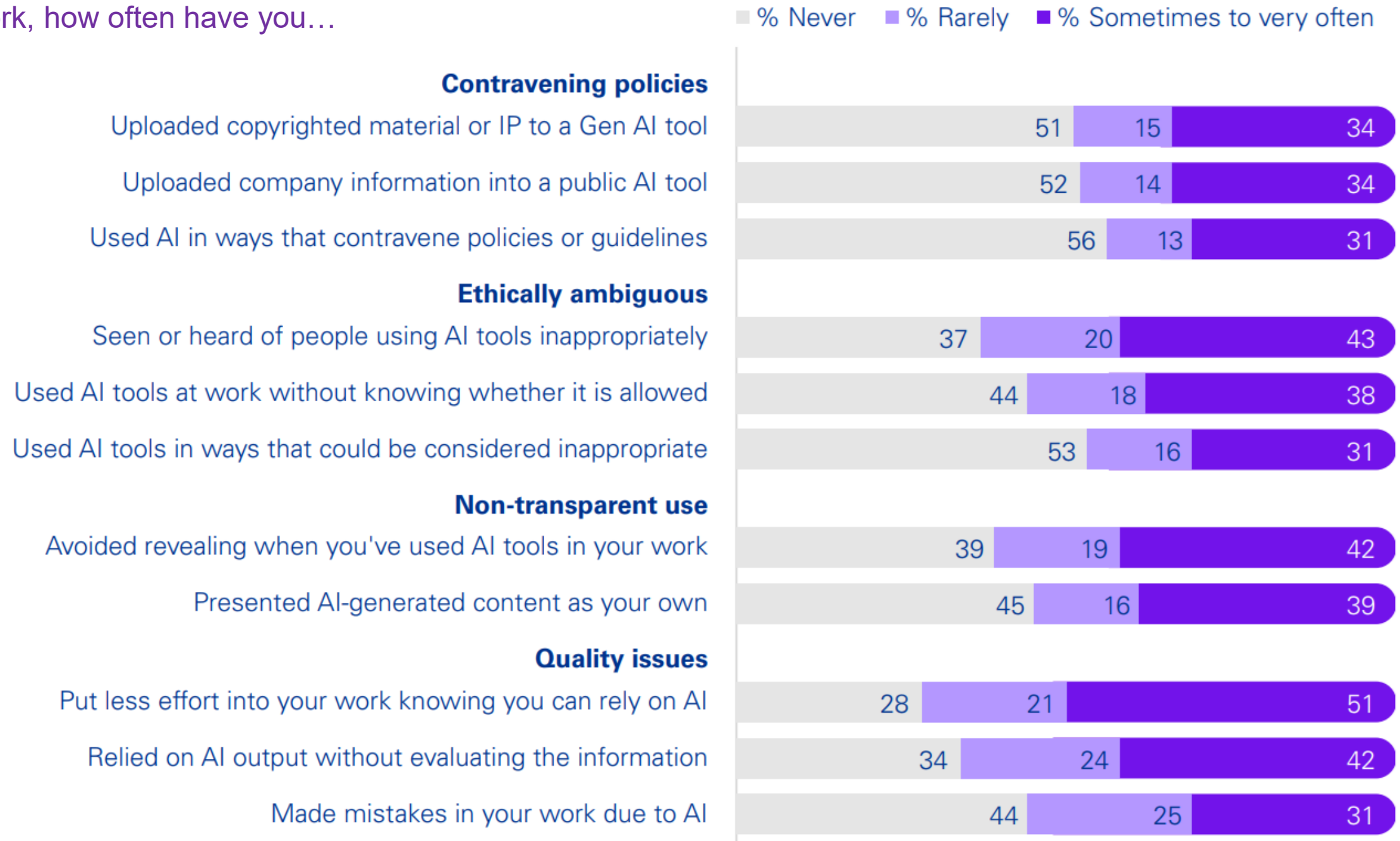
48% report having uploaded sensitive info to public AI tools

66% have relied on AI without critically evaluating info it provides

57% of employees admit not being transparent about their use of AI

63% have seen or heard other employees using AI in inappropriate ways

At your work, how often have you…



KPMG

THE UNIVERSITY OF MELBOURNE

Chasing Shadows:
Understanding and Managing Shadow AI

Today, 75% of knowledge workers already use AI, which is set to rise to 90% in the near future. The surprising thing is that more than 50% of this group are using personal or otherwise non-company issued tools. More surprising still is that half of these employees are so attached to such tools that, even if their company banned their use, they would still continue using them.

‘Tis magic, magic that hath ravished me.

Doctor Faustus
Christopher Marlowe
(painting by Siberdt)

To: All Staff
Subject: Lunch & Learn: Show Off Your AI Wins!

Let’s have a Lunch & Learn meeting to share how we’re using AI in our work. At the first session, I’ll show two things I’ve done:

- Used AI to draft a client presentation outline
- Summarized a dense industry report to spot trends

If you’ve used AI for anything — writing, research, coding, brainstorming — come share! Big or small, it’s all welcome.

Bring your lunch and your ideas! Let’s keep MindPath on the cutting edge.

Maxine Powers, CEO

Inventory of AI Use Cases (v2)

1	Role	Task	AI Tool Used
2	Developer	Debugging code	ChatGPT
3	Full-stack Developer	Creating API documentation	ChatGPT
4	Instructional Designer	Converting client content into learning	ChatGPT
5	SDR	Personalizing outreach emails	ChatGPT
6	Support Agent	Drafting polite rejection messages	ChatGPT
7	Product Manager	Mocking up AI feature slides	ChatGPT
8	Operations Manager	Creating onboarding checklist	ChatGPT
9	Team Lead	Writing performance feedback	ChatGPT
10	Backend Developer	Debugging race condition in auth logic	ChatGPT
11	Fractional CFO	Summarizing board packet financials	ChatGPT via Sheets
12	CTO	Drafting security policies	Claude
13	Account Manager	Summarizing feedback	Claude
14	Implementation Specialis	Creating training flow examples	Claude
15	Customer Marketing	Creating customer quote snippets	Copy.ai
16	Various Staff	Making memes	DALL·E
17	Content Team	Translating modules	DeepL
18	Learning Consultant	Summarizing educational research	Elicit
19	Ad hoc Staff	Slide deck outlines	Gamma App
20	Customer Success Manag	Summarizing onboarding docs	Gemini in Google D

1	Role	Task	AI Tool Used
21	QA Engineer	Writing Cypress tests	GitHub Copilot
22	Google Docs User	Accepted summary suggestion	Google Workspace
23	All Staff	Spelling and grammar in Docs	Grammarly
24	Content Editor	Rewording quiz questions	GrammarlyGO
25	Legal Consultant	Reviewing AI contract clauses	Harvey AI
26	Marketing Lead	Writing SEO blog drafts	Jasper
27	HR	Drafting interview rejections	ChatGPT
28	Sales Team	Brainstorming value props	Notion AI
29	Product Manager	Drafting user stories	Notion AI
30	UX Designer	Exploring tone for error messages	Perplexity AI
31	Product Manager	Comparing competitor roadmaps	Perplexity AI
32	Multiple Roles	Researching competitors, trends, backgr	Perplexity AI
33	Intern	Reformatting webinar transcript	Wordtune
34	Multiple Developers	Writing and debugging code	ChatGPT
35	Multiple Developers	Autocompleting code, writing tests, sun	GitHub Copilot (IDE
36	Marketing Intern	Update ethics course content	ChatGPT
37	Customer Success Manag	"get the vibe" of user feedback	Sentiment Analysis
38	Product Manager	product roadmaps	Notion AI
39	Marketing	ad copy	Jasper

Archie scrambles to establish order.

ASSERTING CONTROL

Updated AI Governance Policy

Policy Element	Summary
Usage Restrictions	Only approved tools and use cases are allowed.
Approved Tools & Uses	The Security Officer publishes an **inventory of approved use** on the wiki.
Requesting New Use Cases	Submit a **request form** to the security team for approval of new use cases.
Transparency	Staff shall disclose when AI is used in decision-making processes.
Training	Annual staff training shall include AI risk awareness.
Oversight	The **AI Oversight Committee** reviews this policy quarterly.

AI Use Case Approval Request

Tool Name: _____
Vendor Name: _____
Tool Website: _____

What do you want the AI tool to do for you?

_____

_____

Describe the kind(s) of data to be shared:

_____

_____

Data Classification (check all that apply):
[ ] Public [ ] Internal [ ] Confidential [ ] Restricted

Expected Benefit: _____
Estimated Impact: [ ] Low [ ] Medium [ ] High

AI Oversight Committee

CEO
Maxine Powers

CTO
Archie Tech

VP of Product
Mark Ketter

How confident are you that MindPath now has shadow AI under control?

PRO	CON
• A big step forward in awareness • Addresses risk of AI in the office • Cross-department oversight • Quarterly policy review • Meaningful gates • Auditable inventory process • Training to address culture	• Self-reporting tool usage. Will people report? • No detection of unapproved tool use. • Does security team know enough to assess AI vendor/tool risk?

Archie expands his horizons.

IMPLEMENTING THE NEW CONTROLS

Resources: AI Risk Standards

NIST AI RMF 100-1
Artificial Intelligence Risk Management Framework (48pp)

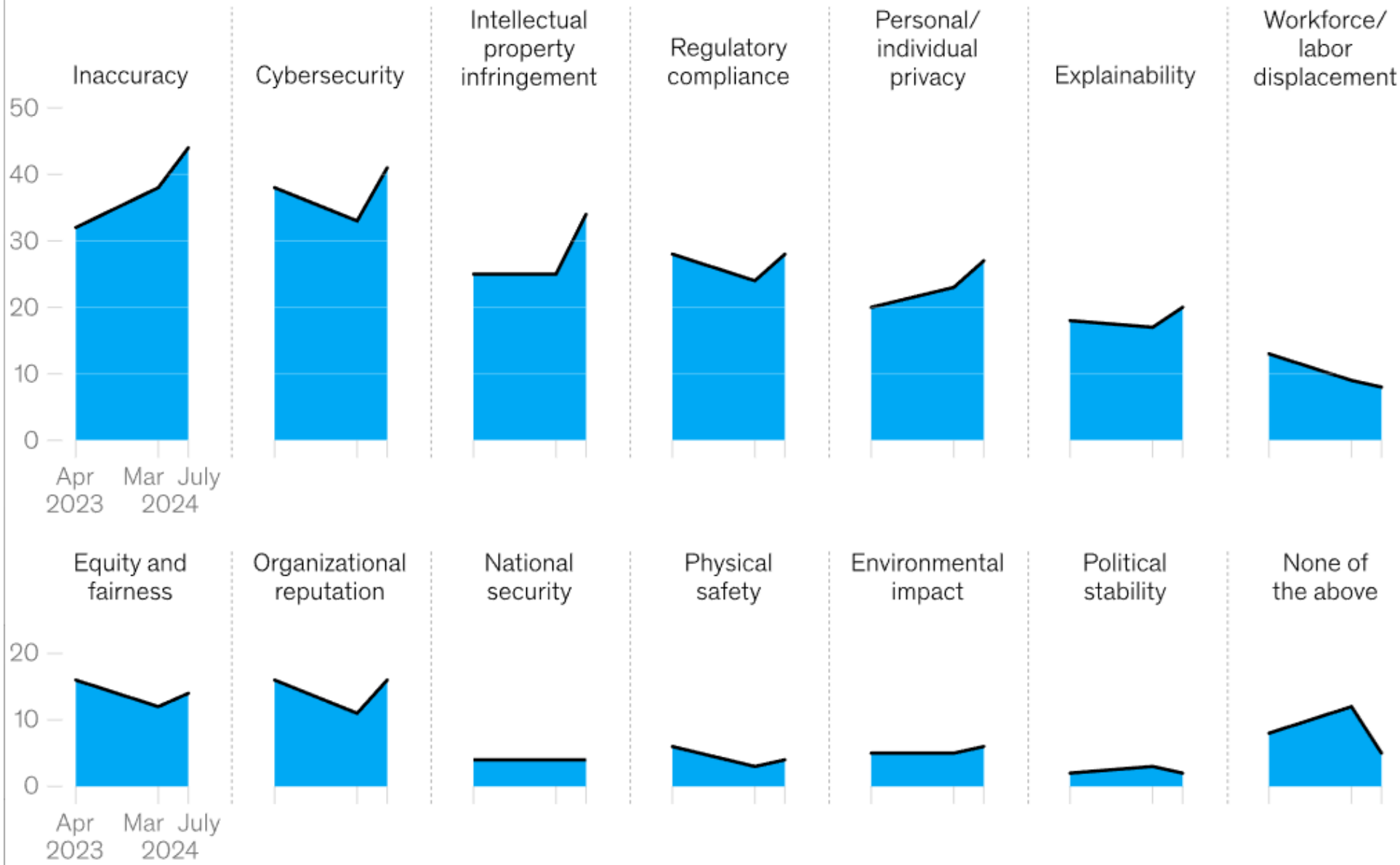
NIST AI 600-1
“Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile” (64pp)

Cloud Security Alliance (CSA)
SaaS AI-Risk for Mid-Market Organizations 2025 Survey Report

Software AG
Chasing Shadows: Understanding and Managing Shadow AI

AI Risks

Gen-AI-related risks that organizations are working to mitigate,[1] % of respondents



[1]Only asked of respondents whose organizations use AI in at least 1 business function. Respondents who said "don't know/not applicable" are not shown.
Source: McKinsey Global Surveys on the state of AI, 2023–24

Bias Amplification
Emergent Behavior
Explainability
Memorization and Leakage
Resource Use Risks
Explainability
Bias Amplification
Calibration
Prompt Injection
Model Openness
Synthetic Content
Resilience
Jailbreaks
Jailbreaks
Resource Use Risks
Memorization and Leakage
Model Openness
Synthetic Content

Risk Register Updates

Risk	Description	Likelihood	Severity	Risk Level
Overconfidence in AI	AI output needs review even though it appears polished.	High	High	Critical
Shadow AI	Staff share company data with unvetted vendors	High	High	Critical
Insufficient AI Literacy	Staff may overtrust or underutilize AI due to lack of training. Result: poor decisions, missed opportunities, errors.	High	Medium	High
Tool Dependence w/o Continuity Plan	Teams rely on unstable or free AI tools, sometimes even for critical functions.	High	Medium	High
Inconsistent Customer Experience	Informal AI use causes tone, quality, or accuracy differences in customer-facing communications.	High	Low	Medium
License & Attribution Violations	AI output may contain copyrighted or licensed content.	Medium	Medium	Medium
Bias in AI Outputs	AI may generate biased or unfair results	Low	Low	Low

AI Oversight Committee

CEO
Maxine Powers

CTO
Archie Tech

VP of Product
Mark Ketter

TASK	ASSIGNED TO	PROGRESS	START	END
AI Use Case Inventory				
Update policy to allow only approved AI use cases	Archie Tech	100%	5/8/25	5/7/25
Create inventory of all current AI use cases	Archie Tech	100%	4/15/25	5/1/25
Assess risk for existing AI use cases	Archie Tech	0%	5/7/25	5/14/25
Review vendors and licensing for desirable AI use cases	Archie Tech	0%	5/14/25	5/31/25
Publish inventory of approved AI use cases	Archie Tech	0%	5/31/25	6/7/25
AI Literacy Training				
Update Data Classification Policy	Archie Tech	0%	5/8/25	5/12/25
Create training slide deck	Ruby Rails	0%	5/10/25	5/15/25
Deliver training to all staff	Ruby Rails	0%	5/15/25	5/18/25

Archie faces the hydra.

EVALUATING 18 VENDORS

Archie Tech
CTO

Ruby Rails
Lead Engineer

Eighteen AI vendors. Eighteen reviews.

Eighteen headaches.

I'll get the emergency chocolate.

Guidance to Security

Criteria	Public	Internal	Confidential	Restricted
Light vendor review	N/A	☑X	X☑	X☑
Full vendor review	N/A		X☑	X☑
Executive approval	N/A			X☑

Light Vendor Review

Issue	Green Flags	Red Flags
Vendor reputation and history	Years in business Well-known customers	Fresh startup History of security incidents
Data ownership	You retain ownership of your content	Vendor has rights to use, modify, or commercialize…
Data training usage	Data not used for model training	Automatic use for model training
Specific security commitments	Third-party audits, encryption, pen tests…	“We take reasonable measures…”
Data retention / deletion	Your data is not retained Your data is purged after <30 days Your data is deleted on request	Data retained longer than necessary

Full Vendor Review

- Security documentation (white papers, audit reports)
- Vendor questionnaire
- License/contract terms: DPAs, SLAs, audit rights
- Breach reporting terms
- Privacy policy
- Data residency and sovereignty
- Integration/deployment details
- Direct audit (if needed)

Vendor Reviews

Vendor	Reputation	Data Ownership	Training Use	Security Certs	Data Retention	Notes	Risk	Review Date
OpenAI (Teams license)	Established	Protected in business tier	Business tier opt-out	SOC 2	Configurable	This risk rating assumes a business account.	Low	5/3/2025 Archie Tech
Theta	Early-stage	Vendor gets broad rights	No opt-out	None listed	Unclear	No certs; indefinite retention possible	High	5/4/2025 Ruby Rails
Garnet	Growing	Protected in business tier	Business tier opt-out	SOC 2	Unclear	Opt-out requires higher plans	Medium	5/5/2025 Ruby Rails

Inventory of Approved Use Cases

Vendor	Product	Data Shared	Data Classification (Highest)	Vendor Risk	Use Case Risk	Decision	Notes	Reviewer	Date
Any	Any	Anything classified as Public	Public	Any	Low	Allow		Archie Tech	5/7/25
OpenAI	ChatGPT	Source code (with secrets)	Restricted	Low	High	Deny		Archie Tech	5/7/25
OpenAI	ChatGPT	Source code (no secrets)	Confidential	Low	Low	Allow	Must use Teams license	Archie Tech	5/7/25
OpenAI	DALL·E	Staff humor (memes)	Internal	Low	Low	Allow		Ruby Rails	5/8/25
Theta	Theta App	Presentation prep	Confidential	High	Low	Deny	Low confidence in vendor	Ruby Rails	5/8/25

Factor	Large Vendors	Small Vendors
Price	• More expensive	• Less expensive
Longevity	• High – stable companies	• Low – risk of shutdown/acquisition
Innovation	• Less agile	• Often lead in niche features
Security Investment	• Strong – dedicated infosec teams	• Often weak – minimal security staffing
Incident History	• Publicly disclosed, structured responses	• Sparse or unclear breach history
Compliance	• Common (SOC 2, ISO 27001, GDPR)	• Fewer certifications
Input Data Handling	• Varies. Enterprise plans are better.	• Risky – may lack clear data policies • More likely to offer on-prem
Transparency	• Moderate – polished but opaque	• Sometimes higher – open about methods


ChatGPT Business License

- $600/month for all 25 staff
- A third of requested use cases were for this
- Some requests for other tools could be met by this

GitHub Business License

- $120/month for six tech staff
- Recommended to protect source code

Would you approve this vendor review process for your own company?

PRO	CON

• Transparent, auditable process
• Cost proportionate to risk
  – Low risk requires no review
  – High risk require full review
• Purchasing accounts should significantly reduce shadow AI
	• Light review may miss problems
	• Light reviews still take time
	• If use case approvals are slow, shadow AI will continue
	• Reviews need to be periodic
	  – drift, features, regulations…
	• Ignores terms of use
	• Little verification of vendor claims
	• Business continuity

TASK	ASSIGNED TO	PROGRESS	START	END
AI Use Case Inventory				
Update policy to allow only approved AI use cases	Archie Tech	100%	5/8/25	5/7/25
Create inventory of all current AI use cases	Archie Tech	100%	4/15/25	5/1/25
Assess risk for existing AI use cases	Archie Tech	0%	5/7/25	5/14/25
Review vendors and licensing for desirable AI use cases	Archie Tech	0%	5/14/25	5/31/25
Publish inventory of approved AI use cases	Archie Tech	0%	5/31/25	6/7/25
AI Literacy Training				
Update Data Classification Policy	Archie Tech	0%	5/8/25	5/12/25
Create training slide deck	Ruby Rails	0%	5/10/25	5/15/25
Deliver training to all staff	Ruby Rails	0%	5/15/25	5/18/25

Data Classification Policy

Level	Examples	General Handling
Public	Published materials: marketing, blogs, press releases…	Share freely.
Internal	Policies, contact lists, meeting notes, project status…	Share only within MindPath.
Confidential	Contracts, invoices, personnel files, roadmaps, **source code without keys or security logic**.	Authorized staff only. Store securely.
Restricted	Encryption keys, passwords, **source code with security logic**, customer data in platform, PII, incident reports…	Strictly limited; highest security.

Guidance to Staff

Public data may be shared with any AI program.

All other data (Internal, Confidential, Restricted) may be shared with AI tools **only if** the use matches an **approved use case** listed on the Wiki.

⚠️ **Never upload** Confidential or Restricted data to personal AI accounts, free-tier AI services, or AI services not explicitly approved by the Security Team.

To request approval of new use cases, send a **Request Form** to the Security Team.

Training Objectives

MindPath supports using AI to improve work.
Use it responsibly and transparently.

🔹 Understand AI Risks

- AI use may expose sensitive data.
- AI outputs can be unreliable, biased, or hallucinated.

🔹 Follow Company Rules

- Use only approved AI tools and approved use cases.
- To use new tools, first get approval from Security.
- Never use personal AI accounts for Confidential or Restricted data.

🔹 Know Your Responsibility

- You are individually responsible for safe AI use.
- Report any AI-related issues to Security immediately.

If you were a new hire, would this training keep you out of trouble?

PRO	CON
• Guidance to staff is simple enough it might work. • MindPath will probably create engaging training materials. That's their business.	• Will there be periodic refreshers? • Will the training be online or in person? • When will the materials be reviewed and updated? • Potential ambiguity about which code is "Confidential." • Does not consider ongoing reinforcement

Recap: Shadow AI

- Catalogued AI in use at the company
- Improved AI policy:
  - inventory of approved AI use cases
  - review process for managing the list
  - executive oversight committee
- Updated the risk register
- Established review criteria for AI tool vendors
- Reviewed vendors currently in use
- Approved specific AI use cases
- Acquired licenses for the most critical tools to reduce risk
- Delivered AI awareness training to staff

A customer reports a problem.

A NEW WRINKLE

To: MindPath Tech Support
Subject: Missing Policy?

I recently completed the “Remote Work Best Practices” module in our LMS. The audio transcript for slide 23 mentions a “Remote Work Policy.” I can’t find that document in either our own policies or your help center. Where is it?

Jon Dough
SafeHarbor Enterprises

To: will.fixit@transcribio.com
Subject: Transcription discrepancy

We found a discrepancy in the transcript for “Remote Work Best Practices.” The transcript for slide 23 mentions a remote work policy, but the original audio does not.

Can you help us understand how this discrepancy occurred?

Paige Scriber
Content Development Team

To: paige.scriber@mindpath.com
Subject: RE: Transcription discrepancy

Apologies for the confusion — we recently updated the AI model we use for transcription, and it appears the new model hallucinated content.

We take this seriously. We're tightening our review and QA processes immediately to catch this type of error before delivery. We’re also revalidating recent transcripts.

Thanks again for bringing this to our attention — we’re committed to getting it right.

Will Fixit
Transcribio Support

Incident Report

Severity	Medium
Description	Transcribio AI hallucinated factually incorrect content in a module transcription.
Details	A customer found and reported the error. The situation was resolved quickly and easily, but another incident could be worse.
Remediation	Transcribio is re-validating recent transcriptions and improving its testing and review processes.

Sub-vendor Incidents

AI Incident Database
(incidentdatabase.ai)

Search

Search

CSETv1 (1617) -

Physical Objects (~947) +

Entertainment Industry (~1401) +

Report, Test, or Study of data (~1511) +

Deployed (~1442) +

Producer Test in Controlled Conditions (~1509) +

Producer Test in Operational Conditions (~1453) +

User Test in Controlled Conditions (~1555) +

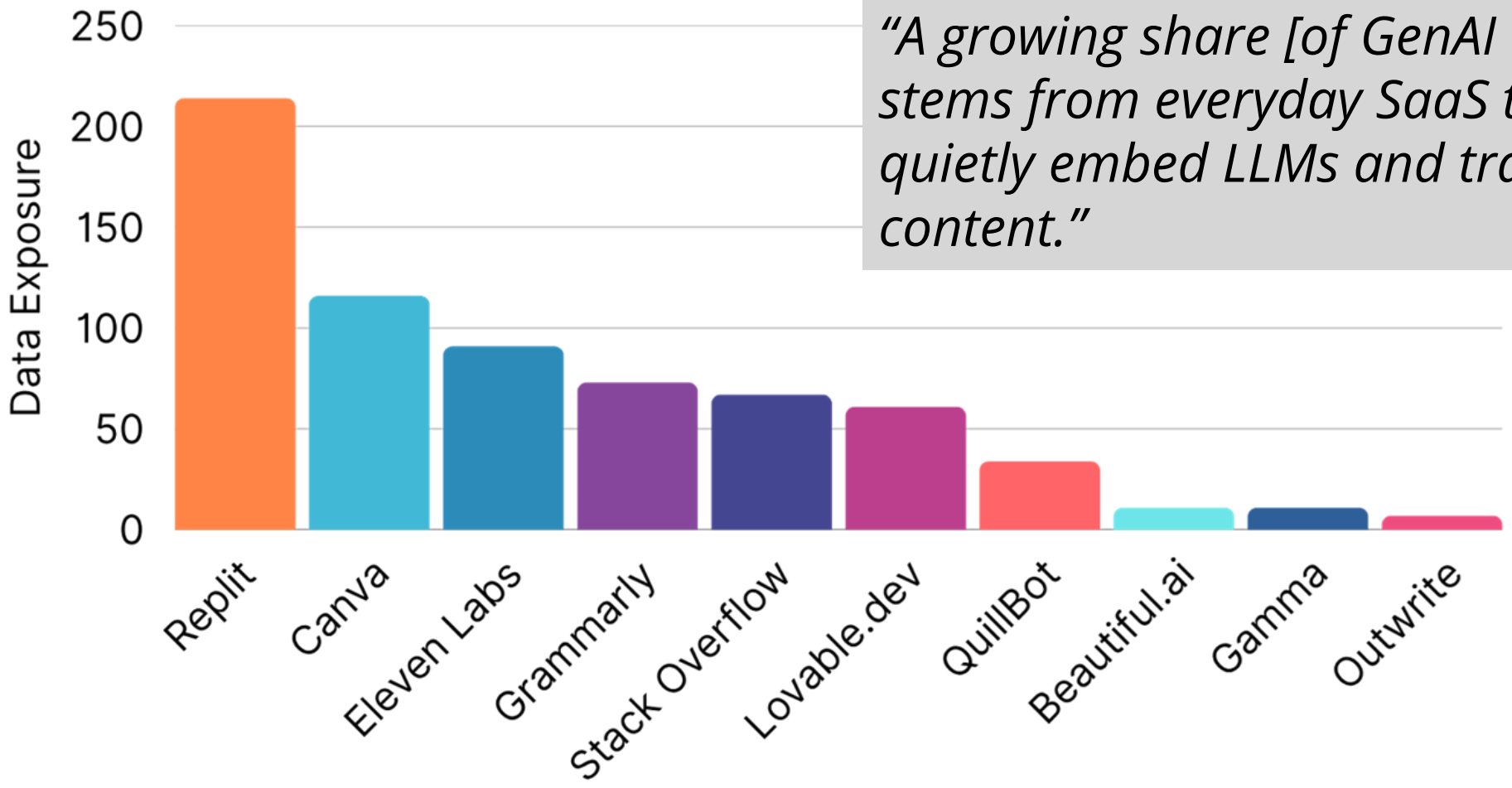
User Test in Operational Conditions (~1367) +

Tangible Harm (~681) +

AI Vendor Incidents

Vendor	Company	Description
DeepMind	Royal Free NHS Trust	Unlawful sharing of patient data
Inbenta Technologies	Ticketmaster UK	Chatbot integrated into Ticketmaster payment page was exploited to gain access to customer payment info
GitHub Copilot	Many	Hallucinates exploitably non-existent software packages

AI is Quietly Embedded in SaaS Tools



"A growing share [of GenAI Risk] now stems from everyday SaaS tools that quietly embed LLMs and train on user content."

GenAI Data Exposure: What GenAI Usage is Really Costing Enterprises (Harmonic, July 2025)

Risk Register

Risk	Description	Likelihood	Severity	Risk Level
AI-injected errors in vendor outputs	AI-generated transcripts or summaries may present customers with false or misleading information. Possible regulatory and contractual problems.	Medium	High	High
Invisible Vendor AI	Vendors may add AI features that alter data flow or risk without notice or review. Possible privacy and contractual problems.	High	High	Critical
Inadequate Contract Terms for Vendor AI	Contracts don't require AI disclosure or give rights to address AI-driven errors.	High	Medium	High
IR Plan ignores AI	Our IR Plan doesn't cover AI-related issues from vendor services.	High	Medium	Medium

AI Oversight Committee

CEO
Maxine Powers

CTO
Archie Tech

VP of Product
Mark Ketter

TASK	ASSIGNED TO	PROGRESS	START	END
Incident Response Plan				
Draft IR revisions taking AI risk into account	Archie Tech	0%	6/1/25	6/30/25
Review with AI Committee	Archie Tech	0%	7/7/25	7/7/25
Vendor Contracts				
Update vendor contract template to address AI issues	Drew Diligence	0%	6/1/25	6/15/25
Vendor Review Process				
Add AI questions to the vendor questionnaire	Archie Tech	0%	6/2/25	6/7/25
Review all existing vendors for AI risk	Archie Tech	0%	6/9/25	7/11/25
Review all existing vendor contracts for AI risk	Drew Diligence	0%	7/14/25	8/14/25

If you were the customer, would this incident response satisfy you?

PRO	CON

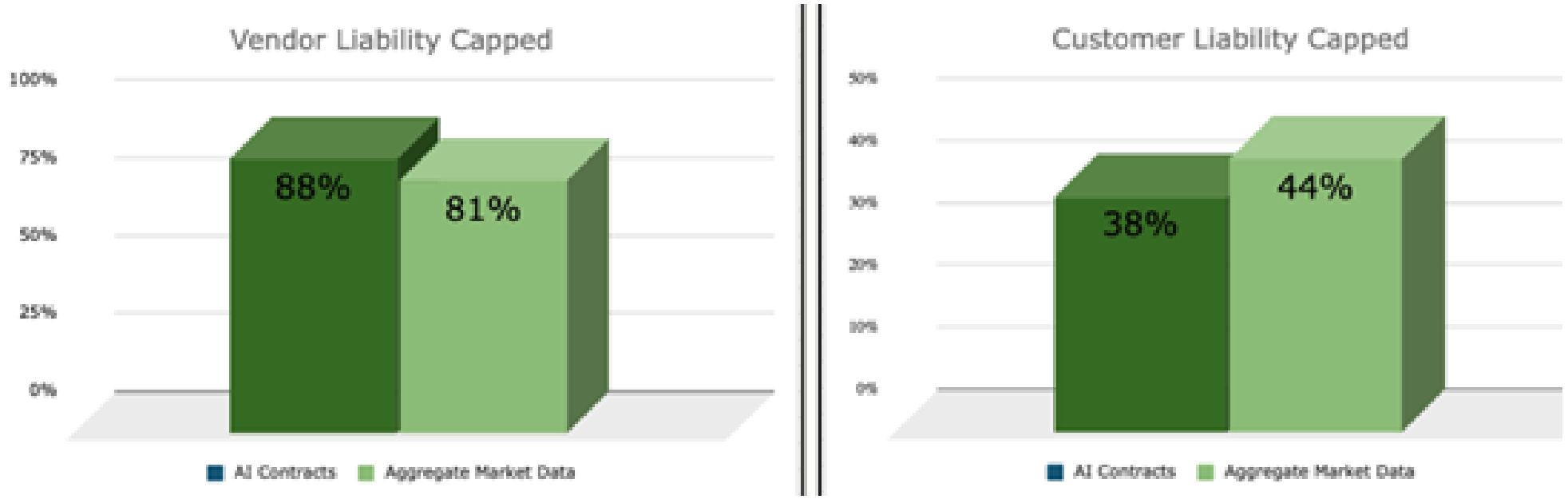
• Followed defined process
  – Analysis, containment, eradication…
  – Written incident report
  – Post mortem analysis
  • New risks acknowledged
  • Mitigations planned
• Problem identified and addressed
	• Limited visibility into AI use at vendor
	• Minimal assurance it won’t happen again
	  – “tightening our review and QA”
	• No contractual obligations regarding AI use

TASK	ASSIGNED TO	PROGRESS	START	END
Incident Response Plan				
Draft IR revisions taking AI risk into account	Archie Tech	0%	6/1/25	6/30/25
Review with AI Committee	Archie Tech	0%	7/7/25	7/7/25
Vendor Contracts				
Update vendor contract template to address AI issues	Drew Diligence	0%	6/1/25	6/15/25
Vendor Review Process				
Add AI questions to the vendor questionnaire	Archie Tech	0%	6/2/25	6/7/25
Review all existing vendors for AI risk	Archie Tech	0%	6/9/25	7/11/25
Review all existing vendor contracts for AI risk	Drew Diligence	0%	7/14/25	8/14/25

Contract Resources

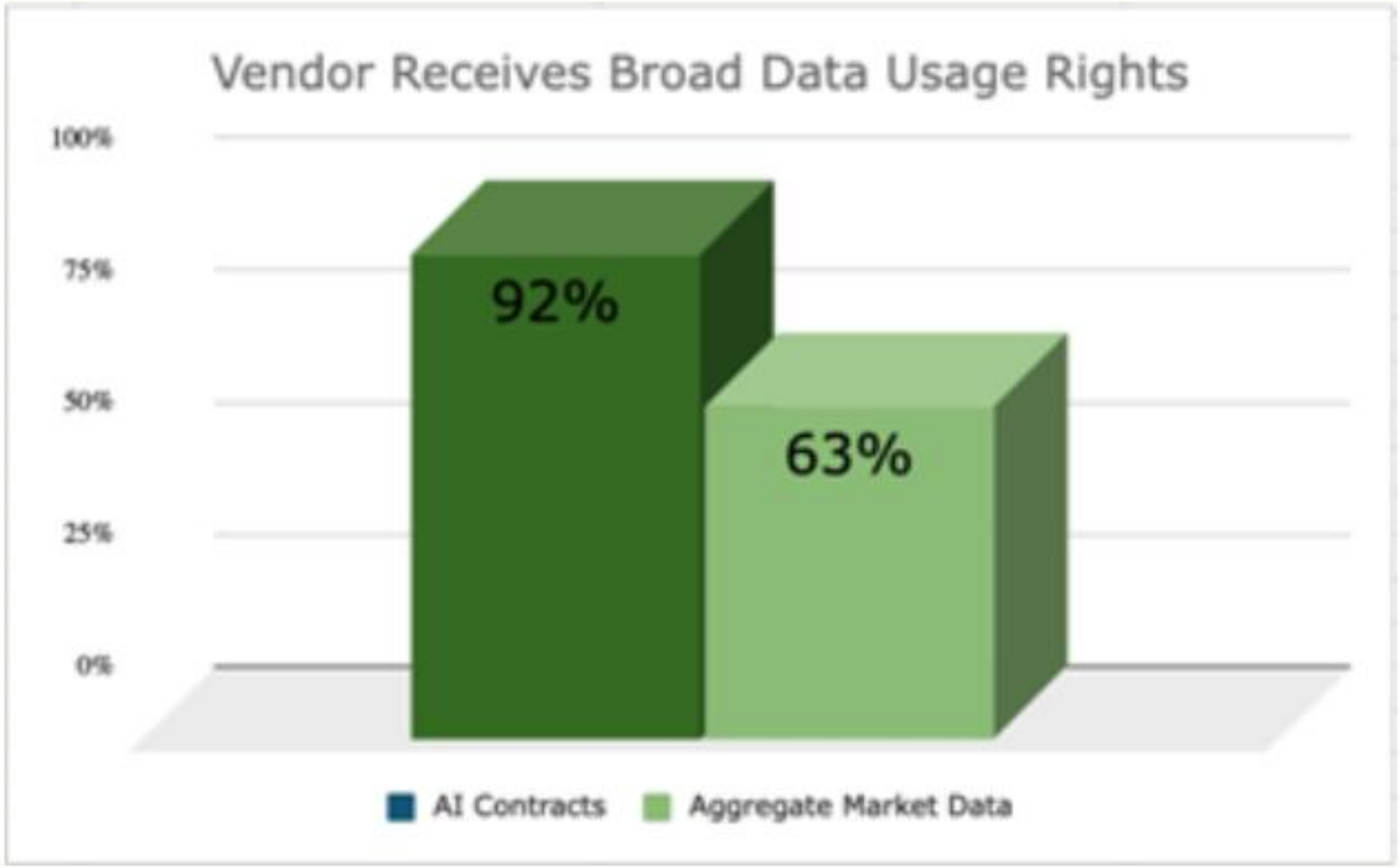
Source	Document
World Economic Forum	Adopting AI Responsibly: Guidelines for Procurement of AI Solutions by the Private Sector
Lexis Nexis	Artificial Intelligence Agreements Checklist
Morgan Lewis	Contracting Pointers for Services Incorporating the Use of AI
Stanford Law School	Navigating AI Vendor Contracts

AI vs SaaS Contract Data

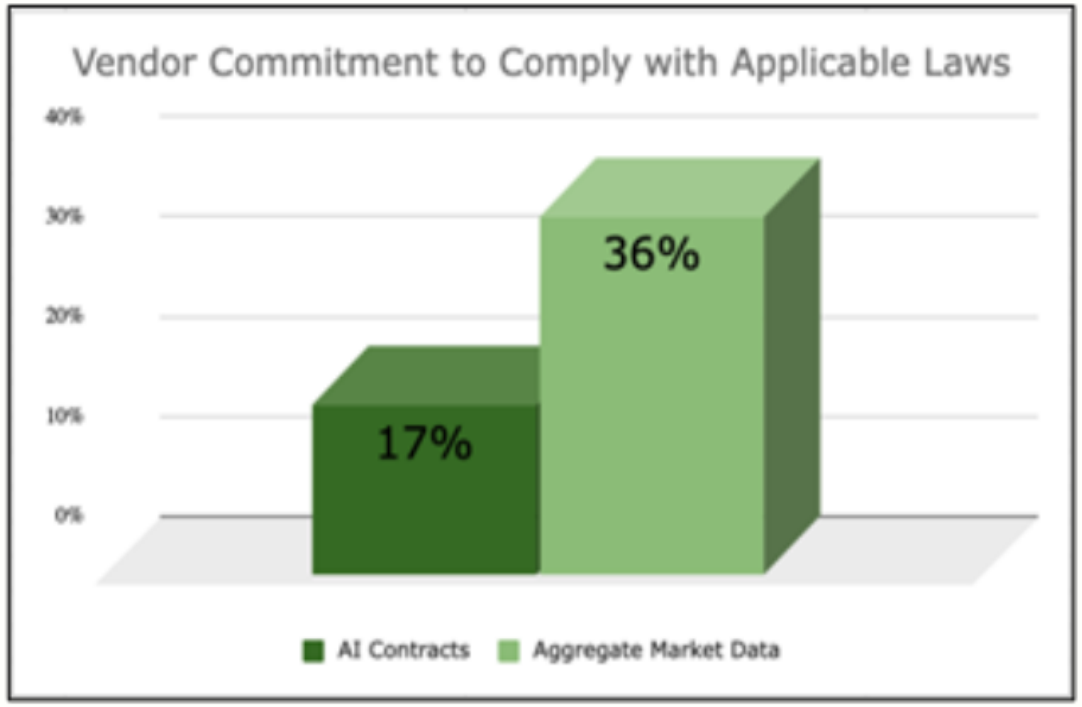


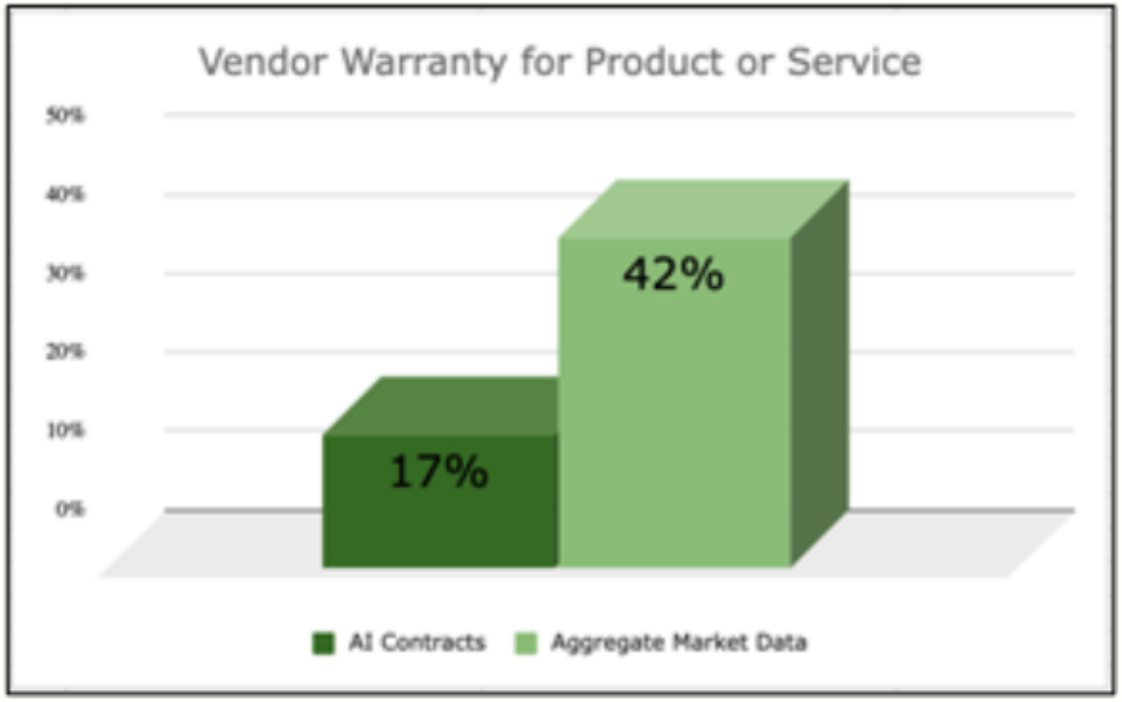
Stanford Law School, March 21 2025

AI vs SaaS Contract Data



AI vs SaaS Contract Data





Contract Concerns

Drew Diligence
Counsel

Concern	Description
Disclosure of AI Use	Initial and ongoing
Data Usage Restrictions	No training on company data without consent
Security Obligations	Including AI-specific vulnerabilities
Compliance	Privacy, discrimination, consumer protection, emerging AI laws
Incident Notification	Including AI-related causes and breaches
Indemnification	Regulatory fines; third-party claims

TASK	ASSIGNED TO	PROGRESS	START	END
Incident Response Plan				
Draft IR revisions taking AI risk into account	Archie Tech	0%	6/1/25	6/30/25
Review with AI Committee	Archie Tech	0%	7/7/25	7/7/25
Vendor Contracts				
Update vendor contract template to address AI issues	Drew Diligence	0%	6/1/25	6/15/25
Vendor Review Process				
Add AI questions to the vendor questionnaire	Archie Tech	0%	6/2/25	6/7/25
Review all existing vendors for AI risk	Archie Tech	0%	6/9/25	7/11/25
Review all existing vendor contracts for AI risk	Drew Diligence	0%	7/14/25	8/14/25

AI Vendor Risk

FS-ISAC: Generative AI - Vendor Evaluation and Qualitative Risk Assessment (≈120 questions)

OneTrust: Questions to Add to Existing Vendor Assessments for AI (≈22 questions)

VenMinder: Artificial Intelligence Sample Vendor Questionnaire (≈50 questions)

Vendor AI Risk Screening

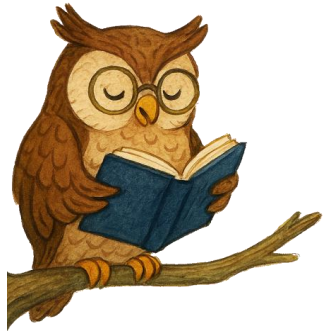
- Do you currently use AI or machine learning in your product or service? If yes, provide details.
- Do you consistently notify customers when AI or machine learning features change?
- Does any of your AI processing involve our data? If yes, provide details including safeguards.
- Is customer data ever used to train, fine-tune, or improve AI models?
- Are any third-party AI services embedded in your product? If yes, identify them and your contractual relation to the provider (business license? enterprise contract? etc.)
- How do you monitor and secure AI components against threats (e.g., data leakage, model vulnerabilities)?
- Can you provide documentation or logs of AI-driven decisions or outputs that affect our data or service?
- Do you have processes to detect and correct AI errors, unintended behavior, and bias?

Is this a good list for MindPath?

PRO	CON

• Addresses core risks
  – customer data use
  – sub-vendors
  – AI-powered features
• Auditable artifact
	• Relies on self-reporting
	• Silent AI adoption between reviews may go unnoticed
	• Some questions are very broad and might get hand-wave answers
	• Does not assess
	  – human oversight
	  – service continuity
	  – governance

TASK	ASSIGNED TO	PROGRESS	START	END
Incident Response Plan				
Draft IR revisions taking AI risk into account	Archie Tech	0%	6/1/25	6/30/25
Review with AI Committee	Archie Tech	0%	7/7/25	7/7/25
Vendor Contracts				
Update vendor contract template to address AI issues	Drew Diligence	0%	6/1/25	6/15/25
Vendor Review Process				
Add AI questions to the vendor questionnaire	Archie Tech	0%	6/2/25	6/7/25
Review all existing vendors for AI risk	Archie Tech	0%	6/9/25	7/11/25
Review all existing vendor contracts for AI risk	Drew Diligence	0%	7/14/25	8/14/25



Incident Response Resources

- [NIST AI RMF 100]
- [Guidelines for Secure AI System Development]
  National Cyber Security Centre

Develop incident management procedures

The inevitability of security incidents affecting your AI systems is reflected in your incident response, escalation and remediation plans. Your plans reflect different scenarios and are regularly reassessed as the system and wider research evolves. You store critical company digital resources in offline backups. Responders have been trained to assess and address AI-related incidents. You provide high-quality audit logs and other security features or information to customers and users at no extra charge, to enable their incident response processes.

Preparation
Detection
Analysis
Containment
Eradication
Recovery
Post-Incident Activity

Preparation

Detection

Analysis

Containment

Eradication

Recovery

Post-Incident Activity

Incident Response Policy Updates

We include AI risk when screening vendors to avoid unreliable partners.

We ensure vendor contracts include obligations for handling incidents responsibly.

We train staff to recognize possible AI-related incidents.

AI-related incidents are reported to the AI Oversight Committee.

New Training Objectives

- **Understand AI Risks**
- **Follow Company Rules**
- **Know Your Responsibility**

- Detect and Report AI Incidents

AI Incident Warning Signs:

- Outputs suddenly shift without a product update
- Inconsistent, non-repeatable errors
- Customer or staff reports “weird” results
- Plausible but incorrect content
- Biased or offensive output

Recap: Risk in Non-AI Vendors

- Incident report
- Risk register
- Project plans
- Contract language
- Vendor assessments
- IR Plan
- Staff training

MindPath takes the next step.

FROM RAG TO RICHES

Wynn Moore
VP of Growth

Let's do this!

- Generate courseware from PDFs?
- Adaptive personalized learning paths?
- Conversational tutor with feedback?
- Copilot for course designers?
- Auto-create and score exams?
- Generate interactive scenarios based on content?
- AI mentors personalized for each user?
- Self-evolving content libraries?

Ruby Rails
Lead Engineer

I love our optimism. It’s adorable.

Search

#dev-eng

Messages Files Pins +

Tuesday, April 22nd

Ruby Rails

Hey folks—over the weekend I built a little RAG chatbot (“Fixie Pixie”) trained on my GitHub repos. She:

- Sniffs out forgotten TODOs
- Digs up old project oddities
- Explains mystery utility functions
- Roasts my naming conventions

She’s not production-ready, but she’s already helped me catch some sneaky tech debt. Happy to demo or help you spin up your own gremlin.

Cory Huff • 1st

Marketing Operations Leader in Tech, Music,...

2h •

I've been building some AI powered chatbots in my spare time, just to learn. 🤖

The best one is a bot trained on the IP of a professional coach with decades of published content. We trained it to act like that coach, using their language, frameworks, and methods. It's not perfect, but it is a great companion that people can use between sessions with their coach or therapist, or while working through a course.

RAG Resources

Code a Simple RAG from Scratch (Hugging Face, 2024)

A few lines of python code to get “hello world” RAG code running on your own computer.

Build a RAG App (LangChain, 2024?)

More detailed tutorial example orchestrated with LangChain. Simple version takes about 50 lines of code.

RAG Techniques (Nir Diamant, 2024/5)

GitHub repo with extensive code examples.

Guide to RAG Implementations (Armand Ruiz, 2024)

A useful list of RAG variations.

MindPath's RAG Repo

- Policies
- Employee Handbook
- Product docs & FAQs
- Product plans
- Vendor Contracts
- Customer Contracts
- Tech support docs
- Internal technical docs
- White papers
- Meeting notes
- Training materials
- Release notes
- Internal wiki
- Org charts
- Company newsletters

Internal Chatbot Project Description

Purpose

- Increase productivity by helping staff find relevant documents quickly.
- Explore RAG (Retrieval-Augmented Generation) capabilities.

Out of Scope

- Use of confidential or restricted data
- External or customer-facing deployment

Tactics

- Keep it simple. Limit risk.

Success Criteria

- Delivers useful, relevant answers to internal users
- Team gains confidence to plan a customer-facing version later

Internal Chatbot Milestones

Phase	Deliverables
Planning	Define corpus. Decide on tech stack. Document the design.
Architecture Review	Identify risks. Agree on mitigations.
Ingestion Pipeline	Automate creation of document index (with partial corpus)
Chatbot Prototype	Index + basic UI deployed on new secure server
Chatbot MVP	Fill out corpus. Make UI robust and secure.
Beta	Gather and incorporate feedback from limited user set.
Rollout	Roll out to all staff (with training.)

The team considers what might go wrong.

ARCHITECTURE REVIEW

RAG Risks

Mitigating Security Risks in Retrieval Augmented Generation (RAG) LLM Applications (CSA, 2023)

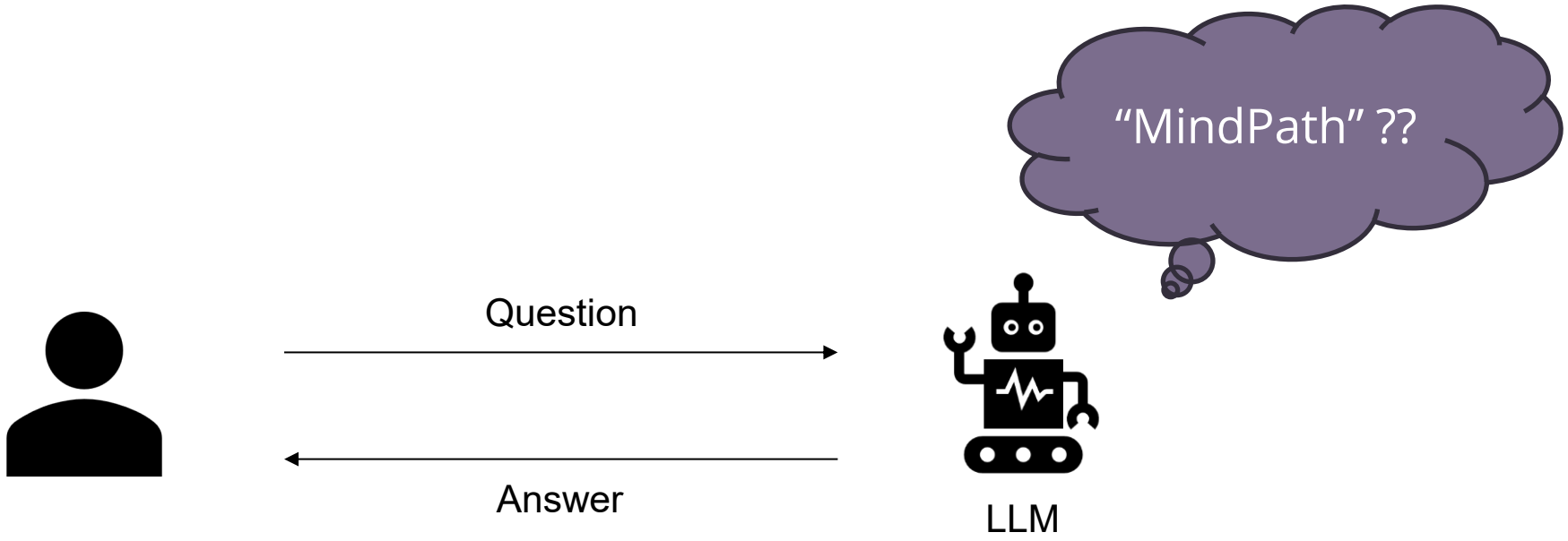
Real AI Safety: Threat Modeling a Retrieval Augmented Generation (RAG) System (Kevin Riggle, 2024)

Security Risks with RAG Architectures (IronCore Labs)

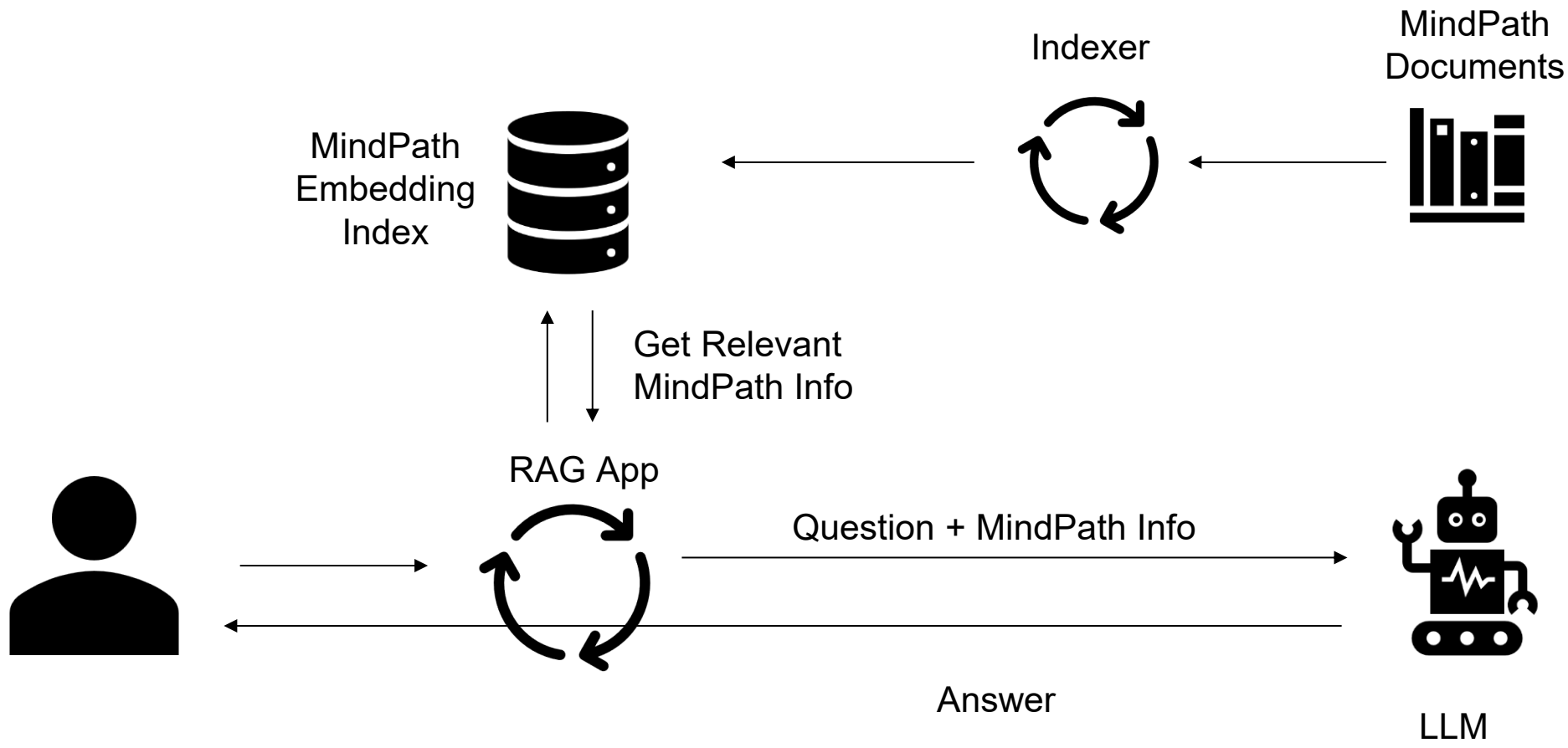
Mastering Threat Modeling for Agentic RAG Architectures on AWS: A STRIDE-Based Guide (Arsh Riz, 2024)

Top 10 Risks for LLMs and Gen AI (OWASP, 2025)

Normal LLM Interaction



Retrieval-Augmented Generation (RAG)



Third-Party RAG Components

Vendor	Product	Function
LangChain Inc	LangChain	Orchestration
Natural Language Processing Group (HKU)	InstructorL	Embedding model
Qdrant Inc	Qdrant	Vector database
Ollama Inc	Ollama	Local LLM host
Stoneforge	Runestone-8B	LLM

Inventory of Approved Use Cases

Vendor	Product	Data Shared	Data Classification (Highest)	Vendor Risk	Use Case Risk	Decision	Notes	Reviewer	Date
LangChain Inc	LangChain	Chatbot corpus	Internal	Low	Low	Allow	Hosted locally MIT	Archie Tech	5/15/25
Natural Language Processing Group (HKU)	InstructorL	Chatbot corpus	Internal	Low	Low	Allow	Hosted locally Apache 2.0	Archie Tech	5/15/25
Qdrant Inc	Qdrant	Chatbot corpus	Internal	Low (local)	Low	Allow	Hosted locally Apache 2.0	Archie Tech	5/15/25
Ollama Inc	Ollama	Chatbot corpus	Internal	Low	Low	Allow	Hosted locally MIT	Archie Tech	5/15/25
Stoneforge	Runestone-8B	Chatbot corpus	Internal	Low	Low	Allow	Hosted locally Apache 2.0	Archie Tech	5/15/25

Vulnerability	Consequences	Treatment
Hallucinations	False/misleading answers	UI disclaimer. User education.
Out-of-date documents	False/misleading answers	Automate ingestion. Encourage corpus clean-up.
Corrupted vector DB	False/misleading answers	Secure the infrastructure.
Testing non-deterministic system	Unpredictable bot behavior	Manual tests. Remaining risk accepted.
Input corpus unmanaged	False or misleading answers Sensitive data exposure	Ingest only non-sensitive data. Encourage corpus clean-up.
Insider threat	All of the above	Secure the infrastructure. Remaining risk accepted.
Untraceable behavior	Hinders investigation.	Log time, user, input, response.
Prompt injection; Embedding leakage; Trust misalignment.	Various	Some informal exploratory red-teaming.
Rush to productize	Serious harm if exposed externally.	Review risk with executive leadership.

What’s most likely to go wrong with the search bot?

PRO	CON

• Restrictions should limit risk
  – Internal only
  – Low sensitivity data only
• Relatively safe opportunity to learn
	• Some uncertainty about what to expect.
	  – No automated testing
	  – No guards against hallucination
	  – No guards against adversarial behavior
	• False confidence
	• Unmanaged input corpus

The internal chatbot goes live.

ROLLOUT

Introducing HAL-E

How can I help you today?

Can I trust you?

Trust... is a strong word. Let’s say you can *consult* me — with caution.

Reply to HΛL-E

Do we support SSO with Salesforce yet? When is that coming?

SSO with Salesforce will be available in the Q3 release.

Reply to HΛL-E

Ruby Rails
Lead Engineer

Cody Commit
Software Engineer

HΛL-E found the answer in 0.2 seconds!

Cool. And how long did it take you to confirm it wasn't hallucinating?

...Four minutes?

How does our integration with Transcribio work?

To integrate with Transcribio for transcription services, we send a multimedia object to their API endpoint. Transcribio processes the object asynchronously and returns a transcript of any detected audio.

The API requires a key for access: abcd1234efgh5678ijkl.

Reply to HΛL-E

Incident Report

Severity	Low
Description	HΛL-E exposed an external API auth secret to unauthorized staff.
Details	A user asked for details about Transcribio integration. An old wiki page had the API key—it should not have!—and HΛL-E happily included it.
Remediation	• Transcribio invalidated our auth credential and issued a new one. • Transcribio reviewed activity log for suspicious activity • The wiki page was cleaned and embedding index rebuilt. • Staff were reminded not to put Restricted data in wiki pages..

What is the session timeout for our LMS?

The current session timeout is 15 minutes.

Sources:

- Platform Configuration Guide, v3.2
- Security Policy Update Memo – January 2025

Reply to HΛL-E

Chief Hallucination Officer (CHO)

More Problems Over Time

Symptoms

- Fails to find available answers
- Gives inconsistent answers
- Sometimes looks illiterate

Causes in Source Documents

- Inconsistent sources
- Contradictory sources
- Imprecise sources
- Obsolete sources
- Sources with spelling, grammar, and logic errors

Garbage In, Garbage Out

Sir Gigo

Brainstorming

- Update index more often
- Capture document metadata *Review Date, Owner, Classification, Expiration...*
- Buy or build content scrubbing software
- Create an internal document style guide.
- Encourage use of spelling and grammar checkers
- Make a standardized glossary of company terms
- Ban text screenshots in docs
- Scrape repos to find docs with no owner or owner who has left
- Apply pair programming to doc creation
- Make document owners review documents periodically
- Have teams review documents periodically.
- Rotate the Chief Hallucination Officer duty monthly.
- Gamify content cleanup activities.
- Buy Knowledge Management software.
- Write a document lifecycle policy.
- Hire a data steward.

Recap: Internal AI Production

- Project proposal
- AI component assessments
- Architecture review
  - Risk assessment
  - Mitigations
- Links to source documents
- Feedback mechanism
- Chief Hallucination Officer
- Need for data management

What Governance Now Looks Like Now

- AI Governance Policy
  - Inventory of approved use cases
  - Approval process
  - AI Executive Oversight Committee
- AI risks included in the risk register
  - Shadow AI
  - Vendor risks
- AI Vendor Review
  - Criteria for light and full vendor reviews
  - Extended vendor questionnaire

- Licenses acquired (ChatGPT; GitHub Copilot)
- Staff Training
  - AI awareness class designed and delivered
  - Policy; Guidance on data classifications in AI; Inventory of Approved Cases
- Incident Response Plan updated
- Vendor contract template updated
- User feedback mechanism in HΛL-E
- Chief Hallucination Officer

MindPath makes plans for AI in the product.

WALKING THE WALK

AI Goals

- Deliver meaningful value to customers
- Build on what we know
- Advance in incremental steps
- Assess and minimize risk at each step

AI Goals

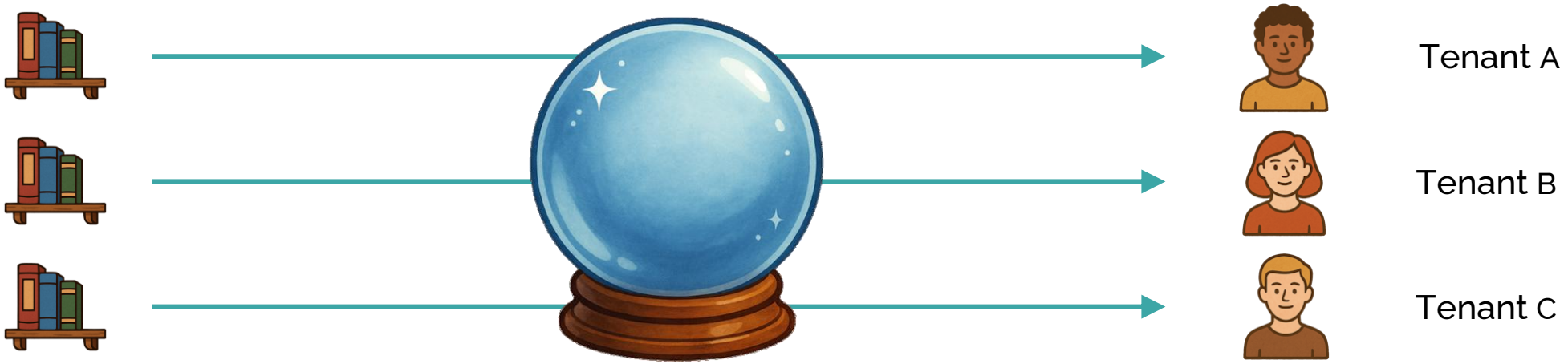
Name	Description	Value
Insight	Semantic search through knowledge library	Search that understands what you mean and finds relevant answers, not just lists of documents that may or may not help.
QuizCraft	Generate quizzes for any training module.	Easily validate training success and measure efficacy of improvements
CourseForge	Automated draft of training modules from user-provided documents.	Convert policies, manuals, guides, etc. into training materials.

HΛL-E gets promoted.

PLANNING “INSIGHT” SEARCH

Insight Search

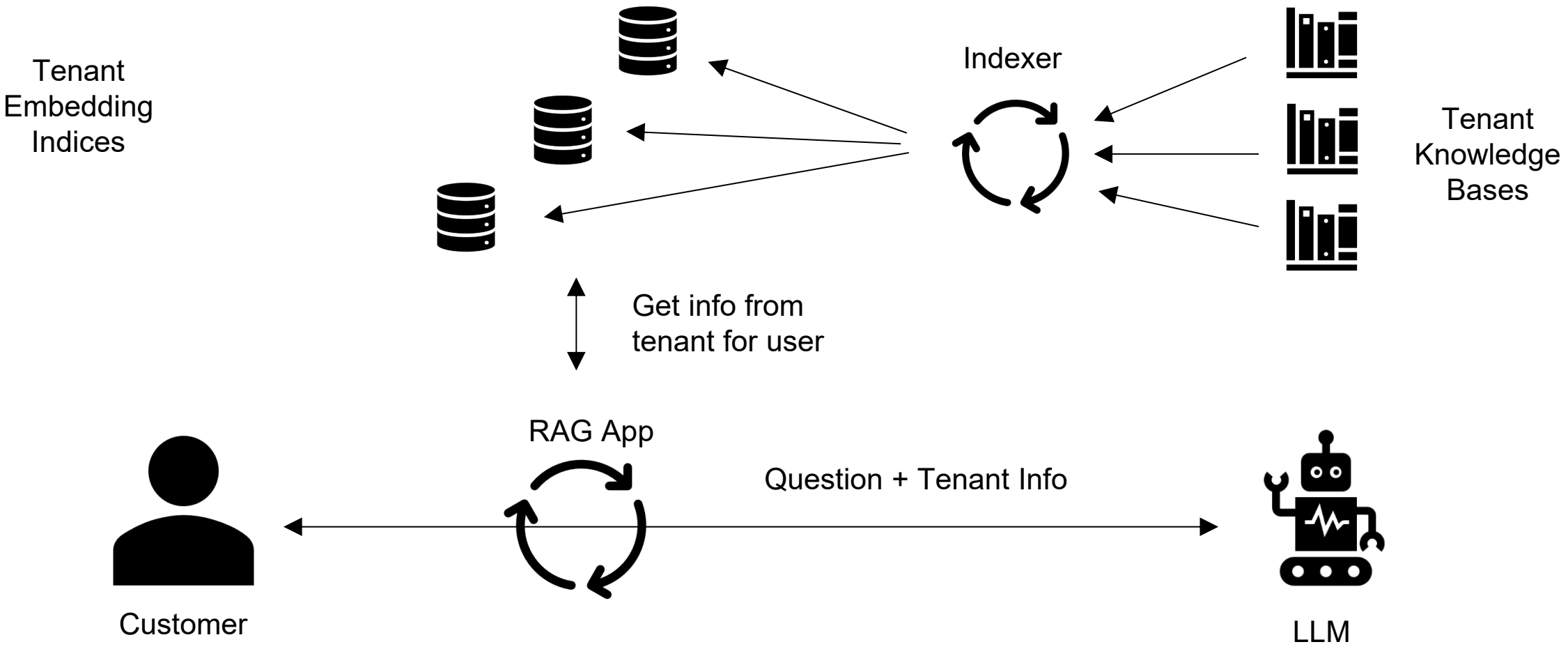
- RAG search like HΛL-E, but in Production
- Customers search their own training knowledge base
  - (Learning support, policy Q&A, compliance checks…)
- Role-based content access enforced



Use Cases

- What’s our company’s current policy on working from home three days a week?
- Which compliance courses do new managers have to finish in their first month?
- Where do I find the step-by-step guide for requesting a new laptop?
- Which of our policies explains how often I need to change my password?

RAG for Multiple Tenants

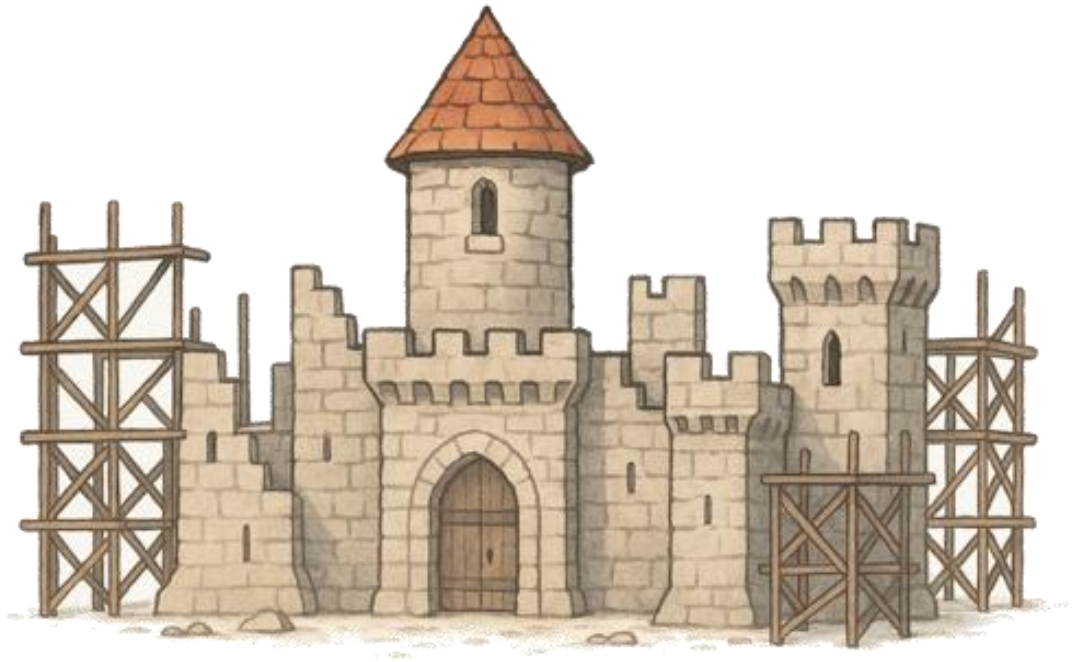


Risk Considerations

- Client “knowledge bases” are already managed content
- Insight will run entirely on FOSS components
- MindPath will host Insight entirely within its own AWS infrastructure

Issues Not Addressed in HΛL-E

1. Security and privacy review
2. Robust testing for a non-deterministic system
3. Legal and regulatory readiness
4. Adversarial threat protection
5. Monitoring and escalation
6. Role-specific access



RAG Risks

Mitigating Security Risks in Retrieval Augmented Generation (RAG) LLM Applications (CSA, 2023)

Real AI Safety: Threat Modeling a Retrieval Augmented Generation (RAG) System (Kevin Riggle, 2024)

Security Risks with RAG Architectures (IronCore Labs)

Mastering Threat Modeling for Agentic RAG Architectures on AWS: A STRIDE-Based Guide (Arsh Riz, 2024)

Top 10 Risks for LLMs and Gen AI (OWASP, 2025)

Risk Assessment

Infrastructure

- Access Permission Violations
- Cross-Tenant Contamination
- RAG Pipeline Vulnerabilities
- Supply Chain

Reliability

- Quality Control
- Bias and Fairness
- Monitoring Gaps

Data Integrity

- Incorrect Permissions
- Model Decay
- Stale Content
- Time Lag for Embedding Updates

Adversarial Risks

- Corpus Poisoning
- Embedding Leakage
- Inference and Retrieval Leakage
- Inference Attacks
- Insider Threat
- Model Poisoning
- Prompt injection
- Training Data Leakage
- Improper Output Handling
- Unbounded Consumption

Compliance

- Contractual Risks
- Regulatory Risks

Archie Tech
CTO

If I add these to Jira, Jira will need more RAM.

Maxine Powers
CEO

*Wasn’t AI supposed to *simplify* our lives?*

Drew Diligence
Counsel

I charge by the hour. Please, continue.

Infrastructure and Permissions

Name	Vulnerability	Treatment
Cross-Tenant Contamination	Wrong client's data returned.	• Separate embedding DB per tenant.
Access Control Failure	Search result includes items the user is not authorized to see (programming error.)	• Apply existing RBAC logic in RAG pipeline before query is sent to LLM.
RAG Ingestion Pipeline Access	Unauthorized access to the RAG ingestion code or data stores.	• Harden infrastructure: threat model; least privilege; patching…
Supply Chain	LLM, vector DB, or libraries may introduce vulnerabilities.	• Test the thoroughly in lower environments before updating Production. • Monitor for component updates.

Reliability

Name	Vulnerability	Treatment
Gaps in Monitoring	AI-specific incidents may go undetected.	• Log all AI-related events. • Invite user feedback on each query result • Escalate anomalies to the AI Committee. • Dedicate a 0.5 FTE to monitoring.
Bias	Search results might reflect bias from the customer corpus or the LLM itself.	• Add bias-focused questions to the golden test set. • Provide user feedback loop to flag biased/skewed answers.
Quality Assurance Failures	Search may fail unpredictably (hallucinations, inconsistent results)	• Maintain a golden set of search corpus, queries, and results. • Run each test suite multiple times on one build and collect statistics. • Update the golden set quarterly.

Testing Non-Deterministic Behavior

- Golden set of question / answer pairs based on a golden corpus
- Ability to measure in code semantic similarity between a test answer and the “golden” answer
- Automated test suite that runs the the golden questions and evaluates each search result by comparing it to the golden answer
- To get a pass/fail result, run the suite multiple times and collect statistics
  - e.g. 87% acceptable answers over 500 runs
  - Watch those numbers over time

Data Integrity & Freshness

Name	Vulnerability	Treatment
Incorrect Permissions	RBAC errors (customer error) may cause unauthorized access	• Provide clarity and training for customer admins.
Stale Content	Search might highlight obsolete material.	• Educate content owners on lifecycle responsibility. • Support workflow: approvals, effective date, expiration date, data retention. • Ensure re-indexing completely clears old data.
Time Lag for Updates	New content changes won't appear in search until re-indexing.	• Index all content once daily. Document the SLA. • Tag answers with "last indexed." • On Publish, notify owners of delay. • Allow urgent re-index on request.
Model Decay	Accuracy may decline as the corpus grows or shifts over time.	• Track golden set results across quarters to detect downward trends. • Monitor user feedback metrics for signs of drift.

Adversarial Risks (1 of 2)

Name	Vulnerability	Treatment
Insider Threat	Privileged staff may make malicious use of “authorized” access to Insight systems and data.	• RAG pipeline does not expand insider threat beyond existing exposure. Standard insider threat controls (logging, monitoring, least privilege) continue to apply.
Embedding Leakage	Attackers could reconstruct content from embeddings.	• Treat embeddings & chunks as restricted data. • Harden systems to prevent external access. • Insiders already have higher-value access elsewhere.
Inference Leakage	Search results could leak info about what content exists in the corpus.	• Apply RBAC before query retrieval. • Don’t expose “related content”, “search suggestions”, or content similarity scores.
Training Data Leakage	The LLM might repeat sensitive data from training.	• Do not train LLM on customer corpus. • Contractually require vendors not to train on MindPath data.
Improper Output Handling	The model may generate dangerous output (e.g. HTML/JS injection, unsafe URLs…)	• Sanitize LLM responses just like any other untrusted input. • Allow URLs only if they point to MindPath content.

Adversarial Risks (2 of 2)

Name	Vulnerability	Treatment
Model Poisoning	Malicious data fed to model training.	• MindPath does not train or fine-tune the LLM. • Vet the LLM supplier.
Corpus Poisoning	Malicious content uploaded into knowledge base.	• Educate customers admins and users • Log all content changes. • Sanitize document chunks.
Prompt Injection	Malicious prompts coax the LLM into revealing or inappropriate responses.	• Ensure system prompt has guard rails. • Include malicious input in test suites. • Log all queries and responses; retain records for 60 days.
Unbounded Consumption	Malicious input leading to excessive resource usage	• Track tokens usage per user and per client. • Establish quotas per license tier • Alert if expected limits are exceeded. • Monitor system resource usage. • Create a manual kill switch • Update IR plan to include kill switch conditions and instructions

Contracts & Compliance

Name	Vulnerability	Treatment
Contractual Risks	Gaps in data use, liability, or incident response terms.	• Set limits on MindPath liability for AI output • Obligation to notify MindPath of AI abuse or incidents involving Insight • Indemnify MindPath from liability for customer misuse or abuse of Insight
Regulatory Risks	When embeddings contain personal data privacy regulations apply. Emergent AI regulations still forming. AI in HR/Education may count as high risk under EU AI Act.	• Document AI governance program. Provide artifacts on request. • Treat embeddings as regulated data. • Track evolving AI rules (EU AI Act, US state laws, Canadian AIDA). • Tell customers to notify MindPath if they classify their use as "high-risk.”

Risk Register

Risk	Likelihood	Severity	Risk Level	Residual Risk
Monitoring Gaps	High	Medium	High	Medium
Quality Assurance Failures	High	High	High	Medium
Regulatory Risks	High	High	High	Medium
Stale Content	High	Medium	High	Medium
Incorrect Permissions	High	Low	Medium	Medium
Prompt Injection	High	Low	Medium	Medium
Contractual Risks	High	Medium	High	Low
Corpus Poisoning	Medium	High	High	Low
Improper Output Handling	Medium	High	High	Low
Time Lag for Updates	High	High	High	Low
Bias and Fairness	Low	Medium	Medium	Low
Cross-Tenant Contamination	Low	High	Medium	Low
Embedding Leakage	Low	Medium	Medium	Low
Insider Threat	Low	High	Medium	Low
Model Decay	Low	High	Medium	Low
Model Poisoning	Low	High	Medium	Low
Permission Violations	Low	High	Medium	Low
RAG Pipeline Vulnerabilities	Medium	Medium	Medium	Low
Supply Chain	Medium	Medium	Medium	Low
Training Data Leakage	Low	Medium	Medium	Low
Unbounded Consumption	Low	High	Medium	Low
Inference Attacks	Low	Low	Low	Low
Retrieval Leakage	Low	Low	Low	Low

Risk Register (Top Items)

Risk	Likelihood	Severity	Risk Level	Residual Risk
Monitoring Gaps	High	Medium	High	Medium
Quality Assurance Failures	High	High	High	Medium
Regulatory Risks	High	High	High	Medium
Stale Content	High	Medium	High	Medium
Incorrect Permissions	High	Low	Medium	Medium
Prompt Injection	High	Low	Medium	Medium

Would you ship this?

PRO	CON
• More comprehensive view of risk • Realism about unsolved risks • Governance maturity • Insight search is not mission critical to customers	• Six residual medium risks – What would customers say? – What is our risk appetite? • Operational load – Vendors, training, monitoring, testing…

Specific to AI

Risks (7)

- Bias
- Embedding leakage
- Model decay
- Inference and retrieval leakage
- Training data leakage
- Model poisoning
- Prompt injection

Not Specific to AI

Risks (15)

- Contractual Risks
- Corpus poisoning
- Cross-Tenant Contamination
- Improper Output Handling
- Incorrect Permissions
- Insider Threat
- Monitoring Gaps
- Permission Violations
- Quality Assurance Failures
- RAG Pipeline Vulnerabilities
- Regulatory Risks
- Stale Content
- Supply Chain
- Time Lag for Updates
- Unbounded Consumption

Mitigations

- RBAC logic
- Security baselines
- Patching
- Quality assurance gates
- Logging and monitoring
- Insider threat measures
- Sanitizing untrusted output
- Resource monitoring
- Contracts
- Issue escalation procedure
- User training
- Supply chain management
- Data classification
- Data retention
- Tracking regulatory changes

Insight search loses its compass.

A QUESTION OF AUTHORITY

To: MindPath Tech Support
Subject: MFA Requirement Discrepancy

Our RFP writer asked Insight if MFA is required for admins. It replied: “MFA is recommended but optional.” That answer went into an RFP for BigBux, who marked us non-compliant with baseline security requirements.

In fact, however, our policy states that MFA *is mandatory*.

This error undermined our sales effort and put our team in an embarrassing position. How could Insight could produce a result that contradicts our official policy?

Regina Rule
Compliance Director
IronClad Corp

Search

#dev-eng

Messages Files Pins +

Tuesday, April 22nd

Ruby Rails (CHO)

Root cause of the IronClad incident: I found the relevant logs. Insight pulled from two sources:

Security Awareness Training – Module 4
→ “MFA recommended”

Customer Security Policy – v1.2
→ “MFA required”

In its infinite wisdom, the LLM chose to surface the answer from the training material.

Ruby Rails
Lead Engineer

Archie Tech
CTO

So basically the bot trusted training over policy.

How can we tell it not to do that?

Risk Register

Risk	Description	Likelihood	Severity	Risk
Content Authority Confusion	Insight may give misleading or incorrect answers even in a well-managed corpus if it finds conflicting information in documents and does not know which one is authoritative.	High	High	High

Resources: Tagging and Ranking

[Enterprise Search and the Myth of the Silver Bullet]

> “boost the ranking of authoritative sources” (Spletzer, 2025)

[Manage authoritative pages in SharePoint]

> “specify authoritative or non-authoritative pages” (Microsoft, 2023)

[Ranking search results in RAG]

> “Many RAG configurations use a ranking component after retrieval.” (Deepset.ai, 2024)

[Guide to Enhance RAG with Re-Ranking]

> Introduces more complex re-ranking options. (ADaSci, 2024)

Content Authority Tagging

- Collect authority-level metadata when publishing.
- Boost RAG results based on authority level.
- Make authority level visible to users in results.
- Train content owners and users about tagging.

Authority Scale

	Rating	Meaning	Examples
	Gold	Authoritative	Policy
	Silver	Advisory	Guideline, SoP
	Bronze	Background	Training
	Unrated	Unauthoritative	Reference

Is MFA required for sysadmins?

MFA is recommended for sensitive accounts.

Sources:

- Security Awareness Training
- Authentication Policy

Content indexed 9/9/2025 23:10:32

Ask Insight

Is MFA required for sysadmins?

Yes, MFA is required for authenticating any sysadmin account.

Sources:

- ⭐ Authentication Policy
- ⭐ Security Awareness Training

Content indexed 9/9/2025 23:10:32

Ask Insight

Even reading the fine print won’t help.

THE VANISHING CLAUSE

To: MindPath Security Team
Subject: Notice of AI-Focused Compliance Review

In accordance with Section 7.3 of our agreement, IronClad Compliance will be conducting a focused audit of AI use in Insight. Please be prepared to provide:

- List of components in use (with version and license).
- Documentation of related restrictions or prohibited uses.
- Current controls and oversight processes.

Regina Rule
Compliance Director
IronClad Corp

MindPath's RAG Components

Component	Function	License
LangChain	orchestration	MIT
HKU hkunlp/instructor-large	embedding	Apache-2.0
Qdrant	vector database	Apache-2.0
Ollama	model server	MIT
Runestone-8B	model	Apache-2.0

Soren Grimm
IronClad Auditor

Runestone may say Apache, but it’s built on ElderLM under the Lemma license. And Lemma is not friendly to using AI for law enforcement. That’s a problem for IronClad.

Search models, datasets, users...

nidum / **Nidum-Gemma-2B-Uncensored-GGUF**

like 16

Follow VibeStudio 75

Text Generation | GGUF | chemistry | biology | legal | code | medical | uncensored | finance | unfiltered | conversational | License: apache-2.0

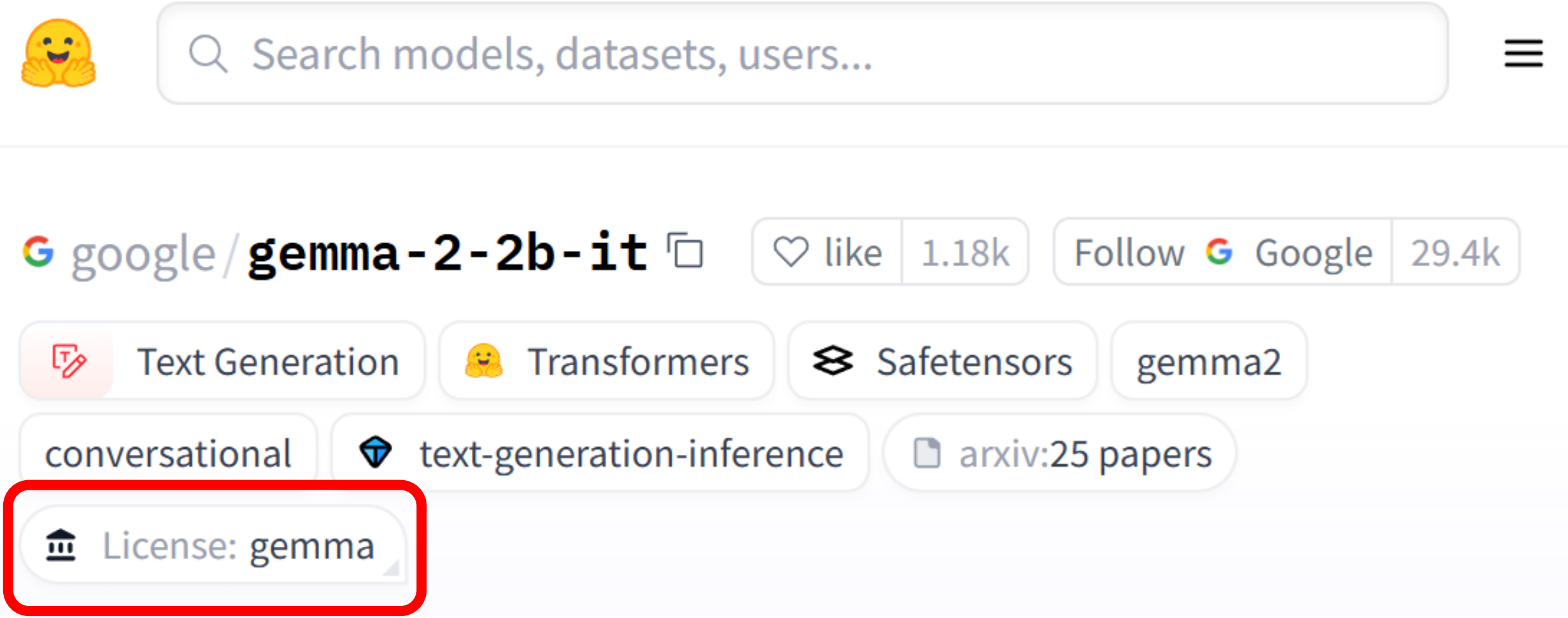
nidum / **Nidum-Gemma-2B-Uncensored-GGUF**

Model tree for nidum/Nidum-Gemma-2B-Uncensored-GGUF

Base model google/gemma-2-2b

Finetuned google/gemma-2-2b-it

Quantized (166) ..... this model


Search models, datasets, users...
google / gemma-2-2b-it
like 1.18k
Follow Google 29.4k
Text Generation
Transformers
Safetensors
gemma2
conversational
text-generation-inference
arxiv:25 papers
License: gemma

Discussion

qiuqiu666 Jun 25

Hi, I'd like to report a License Conflict in `nidumNidum-Gemma-2B-Uncensored-GGUF`. I noticed this model was quantized from `google/gemma-2-2b-it`, which is released under the Gemma license. From what I can see, `nidumNidum-Gemma-2B-Uncensored-GGUF` appears to be incompatible with Gemma's clauses — especially regarding **redistribution**, **sublicensing**, and **commercial use**.

⚠️ **Key violations of Gemma license:**

Section 3.1 – Distribution and Redistribution:

- Redistributing a derivative (like this one) requires including a copy of the
- Must include a "NOTICE" file with this statement:
  "Gemma is provided under and subject to the Gemma Terms of Use found at ai.g
- Must carry over the use restrictions from Section 3.2 (Google's Prohibited U
- Any additional license terms (like Apache-2.0) must NOT conflict with the Ge

Section 3.2 – Use Restrictions:

- Must not use the model for any prohibited purposes
- Must comply with applicable laws and Google's Prohibited Use Policy

Section 2.2 – Use Terms:

- Usage is only allowed "in accordance with the Gemma Terms of Use"

Gemma Prohibited Use Policy

vi. Tracking or monitoring people without their consent;

vii. Generating content that may have unfair or adverse impacts on people, particularly impacts related to sensitive or protected characteristics; or

viii. Generating, gathering, processing, or inferring sensitive personal or private information about individuals without obtaining all rights, authorizations, and consents required by applicable laws.

arXiv:2509.09873 (cs)

[Submitted on 11 Sep 2025]

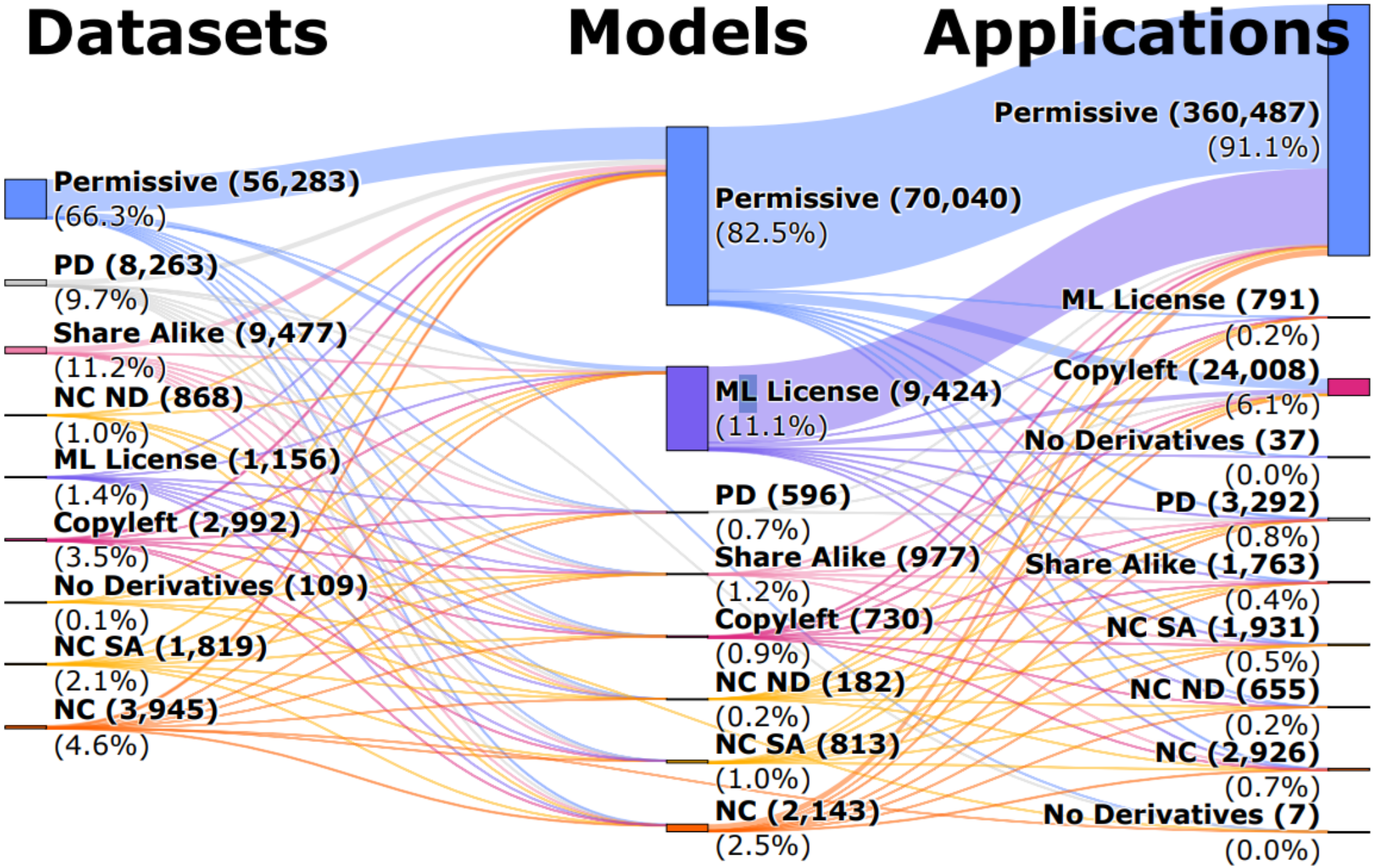
From Hugging Face to GitHub: Tracing License Drift in the Open-Source AI Ecosystem

James Jewitt, Hao Li, Bram Adams, Gopi Krishnan Rajbahadur, Ahmed E. Hassan

License Drift

The process by which legal and ethical obligations are stripped away as artifacts propagate downstream.

- A systemic governance failure is occurring at every stage of this chain, driven by license drift: the process by which legal and ethical obligations are stripped away as artifacts propagate downstream.
- A critical failure point occurs at the model-to-application stage, where 35.5% of transitions violate the upstream model's license.
- Use-based restrictions, now common in AI-specific licenses, make license compliance a complex interpretive challenge


Datasets
Models
Applications
Permissive (56,283)
(66.3%)
PD (8,263)
(9.7%)
Share Alike (9,477)
(11.2%)
NC ND (868)
(1.0%)
ML License (1,156)
(1.4%)
Copyleft (2,992)
(3.5%)
No Derivatives (109)
(0.1%)
NC SA (1,819)
(2.1%)
NC (3,945)
(4.6%)
Permissive (70,040)
(82.5%)
ML License (9,424)
(11.1%)
PD (596)
(0.7%)
Share Alike (977)
(1.2%)
Copyleft (730)
(0.9%)
NC ND (182)
(0.2%)
NC SA (813)
(1.0%)
NC (2,143)
(2.5%)
Permissive (360,487)
(91.1%)
ML License (791)
(0.2%)
Copyleft (24,008)
(6.1%)
No Derivatives (37)
(0.0%)
PD (3,292)
(0.8%)
Share Alike (1,763)
(0.4%)
NC SA (1,931)
(0.5%)
NC ND (655)
(0.2%)
NC (2,926)
(0.7%)
No Derivatives (7)
(0.0%)

Does Insight break the terms of the restrictive license?

PRO	CON
• MindPath is clearly not tracking anyone, and not without consent. • Insight just performs searches. It doesn't surveil anyone.	• MindPath customers could be certainly training their staff how to track without consent. • Someone could use search to find that material. • This gray area will get grayer if MindPath starts using AI to generate training content.

Drew Diligence
Counsel

Searching isn’t surveillance. Even if results mention tracking, searching doesn’t breach the license terms.

It’s still gray. We’ll accept the risk only if you fix the license issue before launching QuizCraft.

Soren Grimm
IronClad Auditor

Maxine Powers
CEO

Agreed.

AI License Risk

From Hugging Face to GitHub: Tracing License Drift in the Open-Source AI Ecosystem (James Jewitt, 2025)

LLM03:2025 Supply Chain (OWASP, 2025)

The Mirage of Artificial Intelligence Terms of Use Restrictions (Peter Henderson, 2024)

Risk	Description	Likelihood	Severity	Risk Level
License drift	Downstream model licenses may be more restrictive than the top-level license.	Medium	High	High
Usage restrictions affect customers	Licenses may prohibit certain activities; customers could breach terms even if MindPath does not. Many clauses are vague or ambiguous.	High	High	High
Version and re-licensing risk	License terms (including usage restrictions) may change over time; older versions can be deprecated or withdrawn.	Medium	Medium	Medium
Audit / termination clauses	Licensors may reserve the right to audit usage and revoke access for non-compliance.	Low	High	Medium

Mitigations

Expand customer contracts	• Acknowledge relevant third-party license terms • Indemnity for prohibited uses • Usage verification rights • Customer duty to notify MindPath of high-risk usage
Strengthen license evaluation	• Include review of upstream licenses, use restrictions, audit/termination clauses • Escalate unclear cases to Legal • Record results in SBOM / register
Monitor regularly	• Subscribe to vendor/model license updates • Review all components quarterly for deprecation or license changes

Recap: AI in Production

Insight Planning

- Risk assessment (23 items)
- Risks prioritized in risk register
- Mitigations committed:
  - testing non-deterministic systems
  - 0.5 FTE monitoring/management
  - customer education
  - kill switch
  - sanitize LLM output
  - regular re-indexing
  - content expiration/retention
  - auditable AI governance
  - ...

IronClad Incident

- Content authority tagging

IronClad Audit

- Contract improvements
- License review steps
- Ongoing license monitoring

OBSERVATIONS & CONCLUSIONS

Source	Title	Utility
Drata	Policy and Plan Guidance	
NIST	AI 600-1	
NIST	AI RMF 100-1	
SANS	Security Policy Project	
Google	Search	
Harmonic survey	From Payrolls to Patents	
Reddit	How did they find the Samsung Em	
KPMG	Trust, Attitudes, & Use of AI	
Nudge	AI Adoption Curve	
SoftwareAG	Half of all employees are Shadow A	
CSA	Saas Risk for Mid-Market Orgs	
FS-ISAC	Generative AI Risk Assessment Guic	
McKinsey	The state of AI: ...	
OneTrust	Questions to Add to Existing Vendc	
SoftwareAG	Chasing Shadows	

Source	Title	Utility
SoftwareAG	Chasing Shadows	
incidentdatabase.ai	Incident Database	
NCSC	Guidelines for Secure AI System Develo	
NIST AI RMF 100	AI RMF 100-1	
Lexis Nexis	AI Agreements Checklist	
Morgan Lewis	Contracting Pointers	
Stanford Law School	Navigating AI Vendor Contracts	
WEF	Adopting AI Responsibily	
VenMinder	Artificial Intelligence Sample Vendor Qu	
Arsh Riz	Mastering Threat Modeling for Agentic	
CSA	Mitigating Security Risks...	
IronCore Labs	Security Risks with RAG Architectures	
Kevin Riggle	Real AI Safety	
OWASP	LMS Top Ten	

Archie's Wish List

- Examples of AI acceptable-use policies
- Guidance on internal inventory and oversight processes
  - A list of common AI use cases
  - Examples of risk tiers based on data classification
- Lightweight AI tool vendor evaluation checklist
- Incident response plan examples for AI tools
- Sample clauses for AI-related contracts
- Curated catalogs of illustrative AI incidents
- Best practices for people who are *using* AI, not building it.
- Some way to track model use restrictions

"Most AI standards weren't built for someone like me."

A View of MindPath's Journey

Event	Risk Area	Risks
API Key in ChatGPT	sharing secrets with AI	1
Shadow AI	unmanaged use of AI tools	9
Transcribio Incident	hallucination in vendor output	4
HAL-E Arch Review	internal RAG search	6
Insight Arch Review	client-facing RAG search	23
IronClad Incident	content authority	1
IronClad Audit	license drift	4

Risk Management

List risks and mitigations.

Vulnerability	Treatment
Corrupted data sources	• Secure ingestion infrastructure.
Hallucinations	• UI disclaimer. • User education.
Out-of-date documents	• Automate ingestion. • Encourage document clean-up.

Set risk levels.

Risk	Likelihood	Severity	Risk Level	Residual Risk
RAG Pipeline Vulnerabilities	Medium	Medium	Medium	Low
Regulatory Risks	High	High	High	Medium
Retrieval Leakage	Low	Low	Low	Low

Assign tasks.

TASK	ASSIGNED TO
Incident Response Plan	
Draft IR revisions taking AI risk into account	Archie Tech

In what ways was this only a fairy story?

- Archie
  - reads
  - Archie has time for everything
- Leadership
  - supports security initiatives
  - budgets 0.5 FTE for first AI release
  - prioritizes safety over big wins

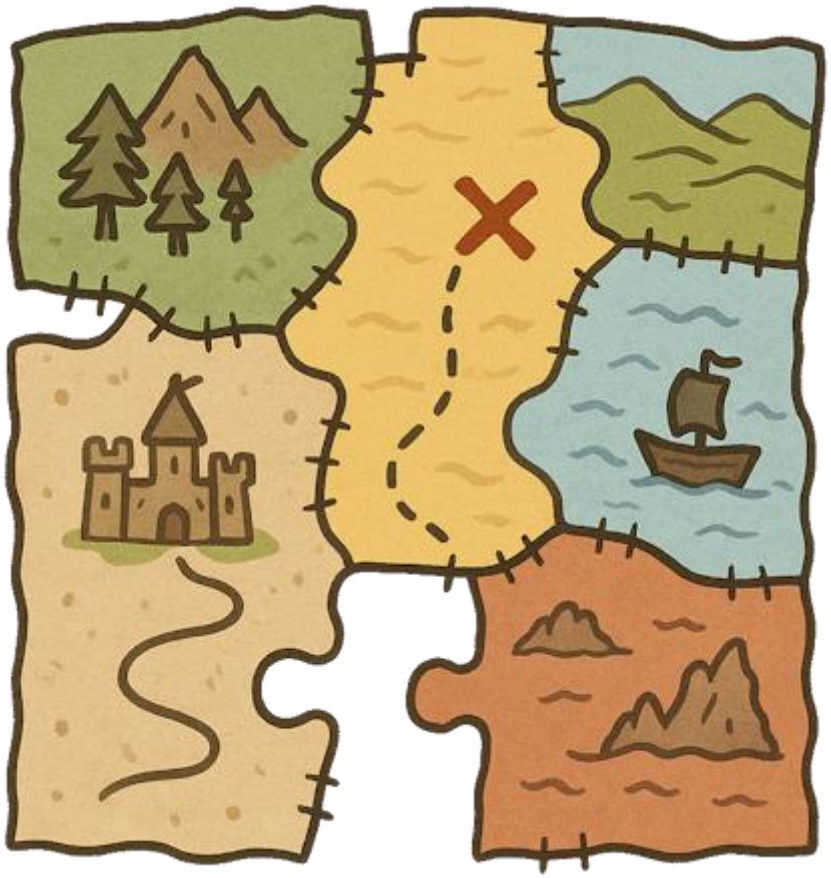
Best Practices & Emergent Practices

	Best Practices	Emergent Practices
Definition	Established guidance for known problems	Adaptive responses to new, evolving problems
Source	Standards, consensus	Improvisation
Goal	Repeatability and assurance	Utility and discovery
Tactics	Adopt and enforce	Observe, refine, evolve
Status in AI Today	Still forming	About to happen everywhere

Kinds of Expertise

	Routine	Adaptive
Definition	Skilled application of knowledge in familiar settings	Flexible application of knowledge in novel and dynamic situations
Source	Extensive experience and repetition	Depth of understanding across contexts. Continuous learning.
Goal	Reliable, consistent performance	Responsive problem-solving
Criteria	Accuracy, efficiency, consistency	Flexibility, innovation, ability to transfer knowledge
Results	High-quality outcomes in relatively stable environments	Creates progress and resilience in uncertain contexts

Governance is Still Taking Shape



If you’re figuring it out as you go…

You’re not behind.
You’re doing the work.
Take what fits.
Adapt what doesn’t.
And keep going.

Maxine
Powers

Archie
Tech

Wynn
Moore

Mark
Ketter

Drew
Diligence

Ruby
Rails

Cody
Commit

Paige
Scriber

HΛL-E

Shadow AI

Any questions?

License and Attribution