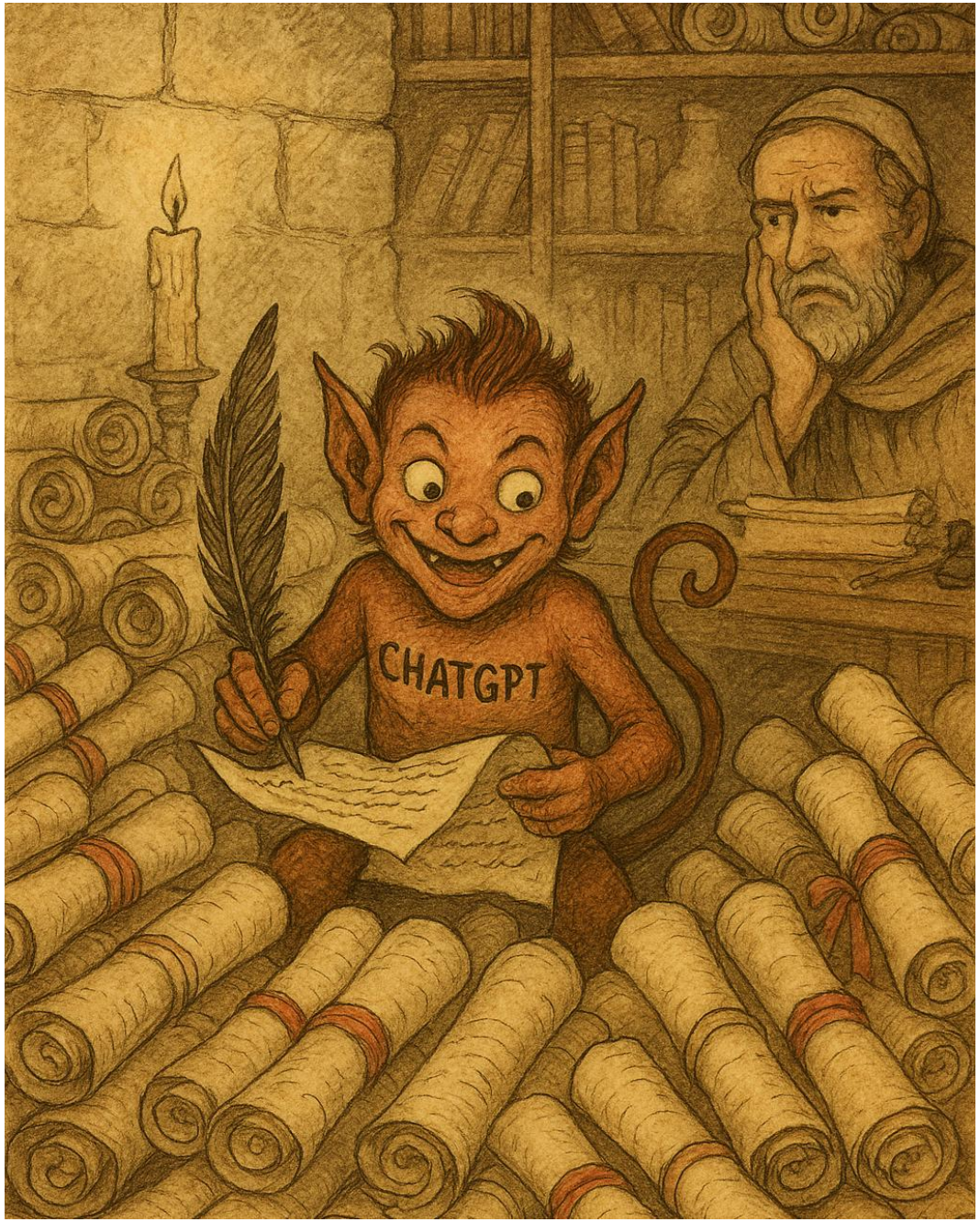


Waking Up to AI

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

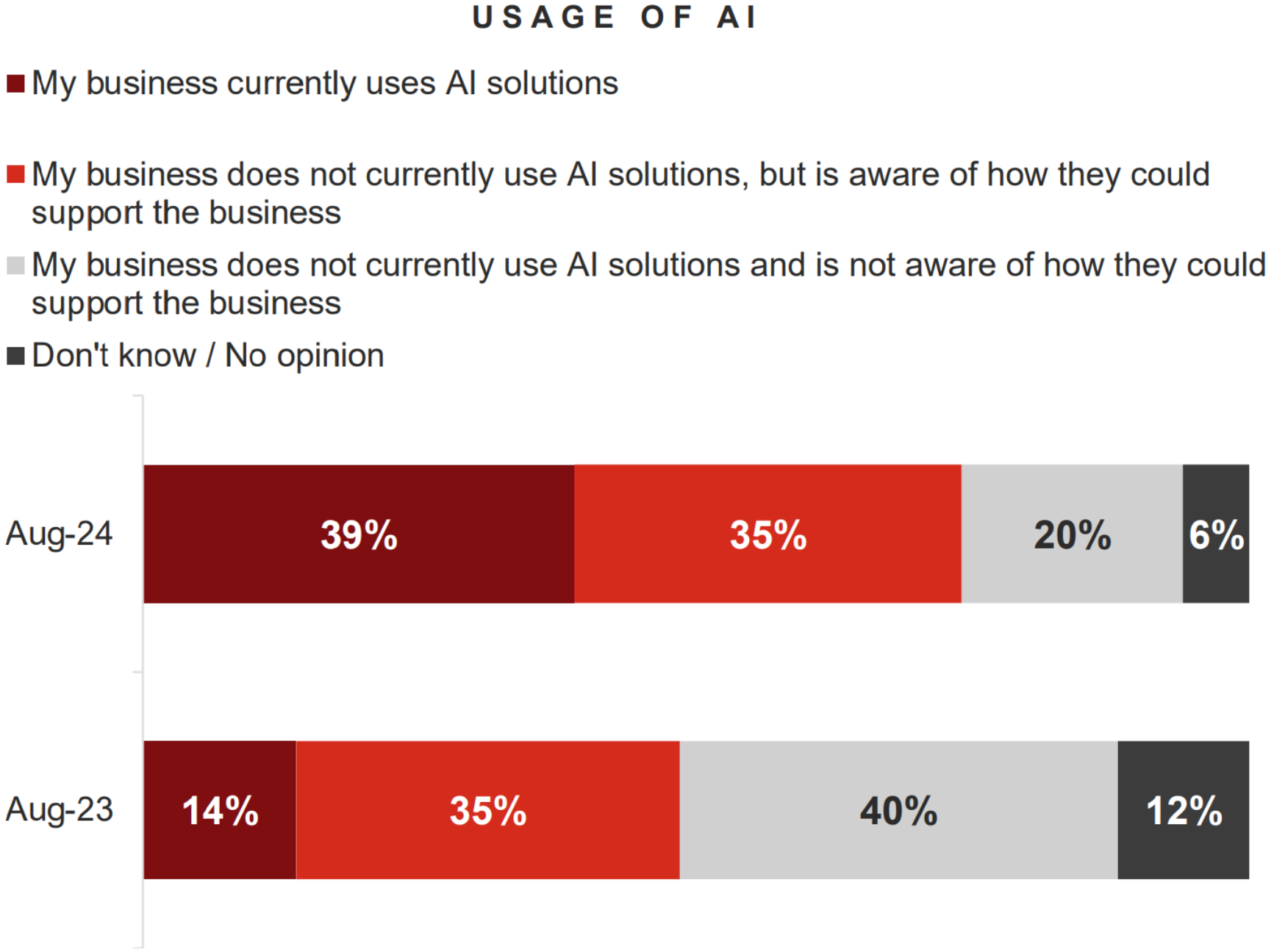
Current Work

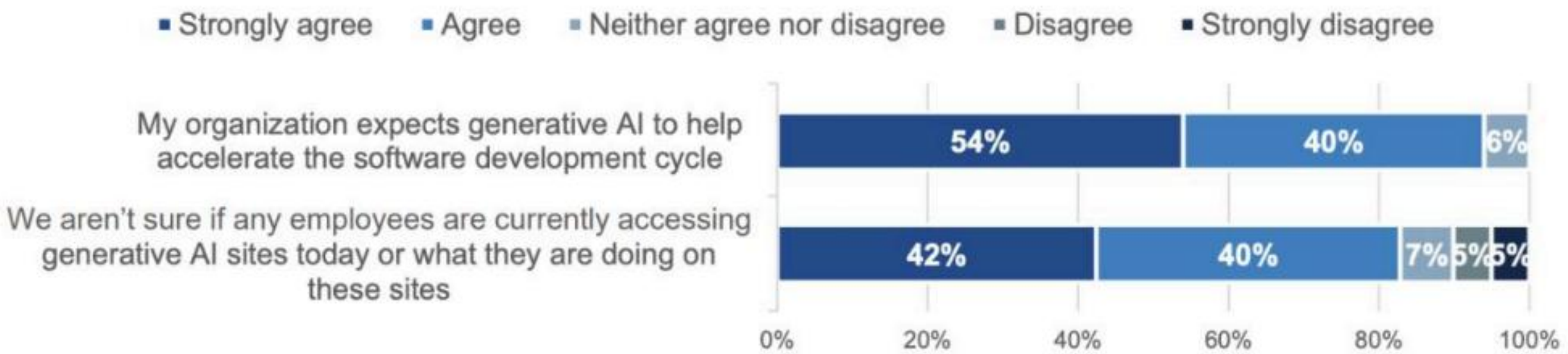
- Independent Information Security Consultant

Volunteer

- OWASP AppSec Days PNW
- Western Oregon University CS Advisory Board

Verizon State of Small Business Survey 2024

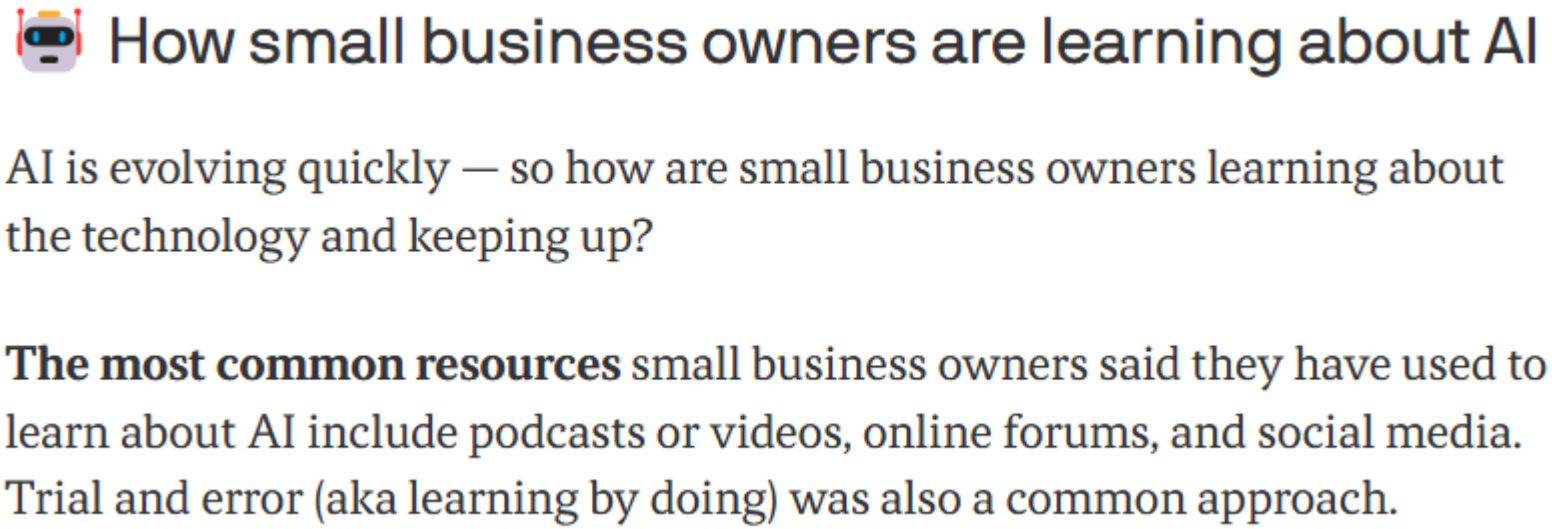




Source: Enterprise Strategy Group, a division of TechTarget, Inc.

https://www.harmonic.security/resources/from-payrolls-to-patents-the-spectrum-of-data-leaked-into-genai

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 Is

- A scenario-based walkthrough of a fictional but representative small company’s growing awareness of AI risk.
- A story about slow governance wake-up calls.
- An illustration of common challenges, practical responses, and the current state of AI governance.

Why Do It This way?

- Governance isn’t just policy—it’s response, learning, and iteration.
- The answers aren’t in the standards yet.
- Watching others work it out helps us work it out too.

Once upon a time...

- LMS for professional education
- SaaS platform on AWS

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

- Google Workspace
- Jira & Confluence
- GitHub

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.

Resources: AI Policy Examples

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

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!

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)
```

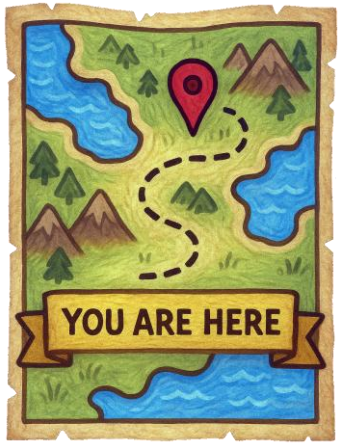
Secure Coding Policy

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.



Is Training Data Leakage a Thing?

Incident	What Training Data Leaked
*Carlini et al.	Medical images
*Books3 Extraction	Full book passages
NYT v. OpenAI	Copyrighted news articles

* Academic studies

“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

[link]

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

harmonic

r/ChatGPTPro • 2 yr. ago [redacted]

How did they find the Samsung Employees who used ChatGPT?

Discussion

Hi all, as some of you may know, Samsung made headlines with ChatGPT where 3 of their employees leaked confidential information. Article here: https://adguard.com/en/blog/samsung-chatgpt-leak-privacy.html

Does anyone know how they got caught?

My company just gave out a ban on using ChatGPT on our projects, it's reasonable as no company wants to leak their proprietary information. However, ChatGPT is far too valuable not to be utilized on a day to day basis.

1. What are your thoughts on using ChatGPT in the work place?
2. What safeguards are you implementing if you are an employer
3. What are you doing to not get caught as an employee?

Tis magic, magic that hath ravished me.

Marlowe, *Dr Faustus*

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 and loss of oversight.	High	Medium	High

Archie goes exploring.

RECONNAISSANCE

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

Hi everyone,

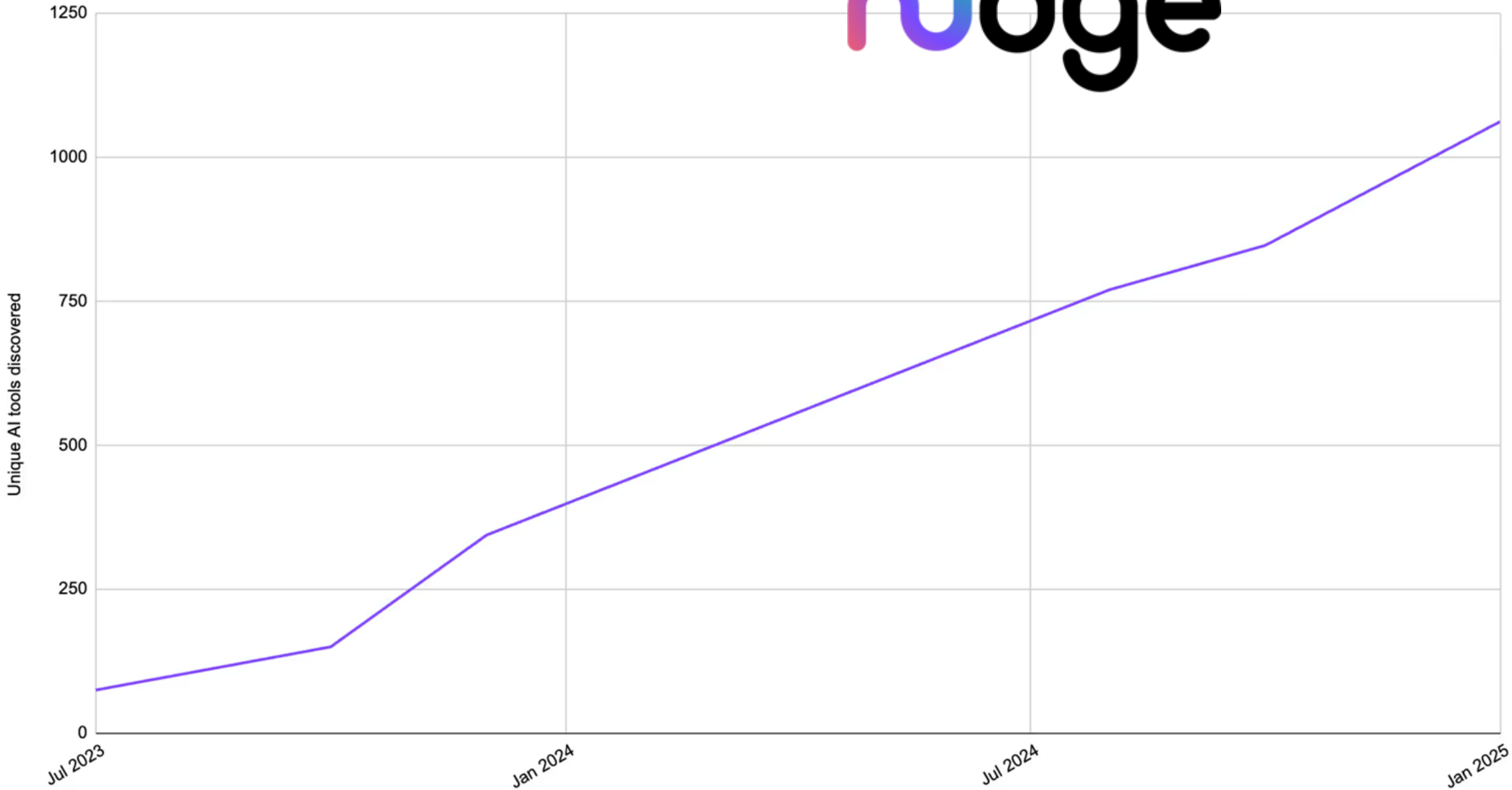
As AI tools become more common, it's important we understand how they're being used at MindPath—especially as individuals explore them on their own.

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



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

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.

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.

GAINING CONTROL

Updated AI Governance Policy

Policy Element	Summary
Usage Restrictions	Only approved tools and use cases are allowed.
Approved Tools & Uses	The **inventory of approved use** cases is published on the wiki.
Requesting New Use Cases	Submit a **request form** to the security team for approval of new use cases.
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 Growth
Wynn Moore

Archie expands his horizons.

ASSESSING AI RISK

Resources: AI Risk Standards

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

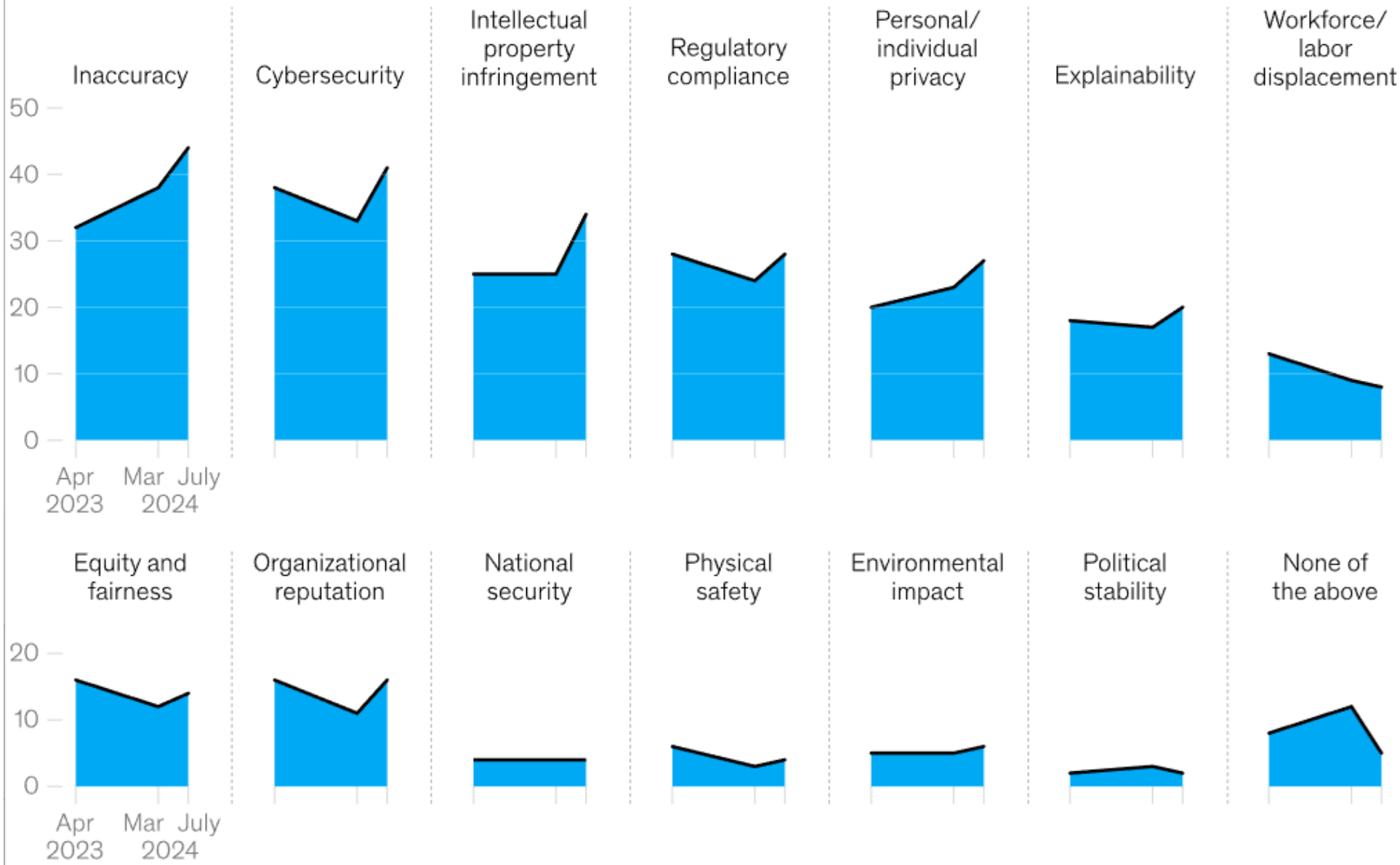
NIST AI 600-1
“Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile” (64pp)

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
2	Lack of Human Oversight	AI-generated content (e.g. summaries, quiz questions) may be used without sufficient human review, resulting in factual errors, hallucinations, or brand damage.	High	High	Critical
3	Data Classification Gaps	Staff use AI tools without knowing whether the data involved is safe for external processing.	High	Medium	High
4	False Confidence in Output	AI-generated text appears polished but may contain inaccuracies; staff accept it without review.	High	Medium	High
5	License & Attribution Violations	AI-generated content may replicate proprietary or copyrighted material, leading to potential IP infringement.	Medium	High	High
6	Shadow AI	Company data is shared with third-party LLMs outside of approved workflows, creating confidentiality, compliance, and IP risks.	High	Medium	High
7	Inconsistent Customer Experience	Informal AI use causes tone, quality, or accuracy differences in customer-facing communications.	High	Low	Medium
8	Insufficient AI Literacy	Staff may overtrust or underutilize AI due to lack of training, leading to poor decisions, missed opportunities, or reliance on flawed outputs.	High	Medium	Medium
9	Tool Dependence w/o Continuity Plan	Critical team workflows rely on unstable or free AI tools, risking disruption if tools change or disappear.	High	Medium	Medium
10	Bias in AI Outputs	AI-generated language, content, or scores may exhibit bias, undermining learning outcomes, alienating users, or exposing the company to fairness concerns.	Low	Low	Low

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	Internal	Confidential	Restricted
No training on inputs	☑	☑	☑
Input retention < 30 days	☑	☑	☑
MindPath retains ownership	☑	☑	☑
Light vendor review	☑		
Full vendor review		☑	☑
Company-owned account		☑	☑
Executive approval			☑

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 training (or opt out option)	Automatic use for model training
Specific security commitments	Encryption, certification…	“We take reasonable measures…”
Data retention / deletion	Your data is not retained Your data is deleted on request	Data retained longer than necessary

Full Vendor Review

- Privacy Policy
- Security documentation (white papers, audit reports)
- License terms: DPAs, SLAs, audit rights
- Breach reporting terms
- Data residency and sovereignty
- Deployment options
- Vendor Questionnaire

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	Prohibit	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 Teams License

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

GitHub Business License

- $100/month for all tech staff
- Recommended to protect source code (restricted)

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 auth* secrets.	Authorized staff only. Store securely.
Restricted	Encryption keys, passwords, customer data in platform, 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 outputs can be unreliable, biased, or hallucinated.
- AI may expose sensitive or third-party data.

🔹 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.

Recap

- Handled an incident when a developer shared a secret with AI
- Catalogued AI in use at the company (and discovered "shadow AI")
- Improved governance by establishing
  - an executive oversight committee
  - a list of approved AI use cases
  - a review process for managing the list
- Updated the risk register to reflect AI risk
- Established criteria for reviewing 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 either in our own policies or your help center. Where I can access 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
MindPath

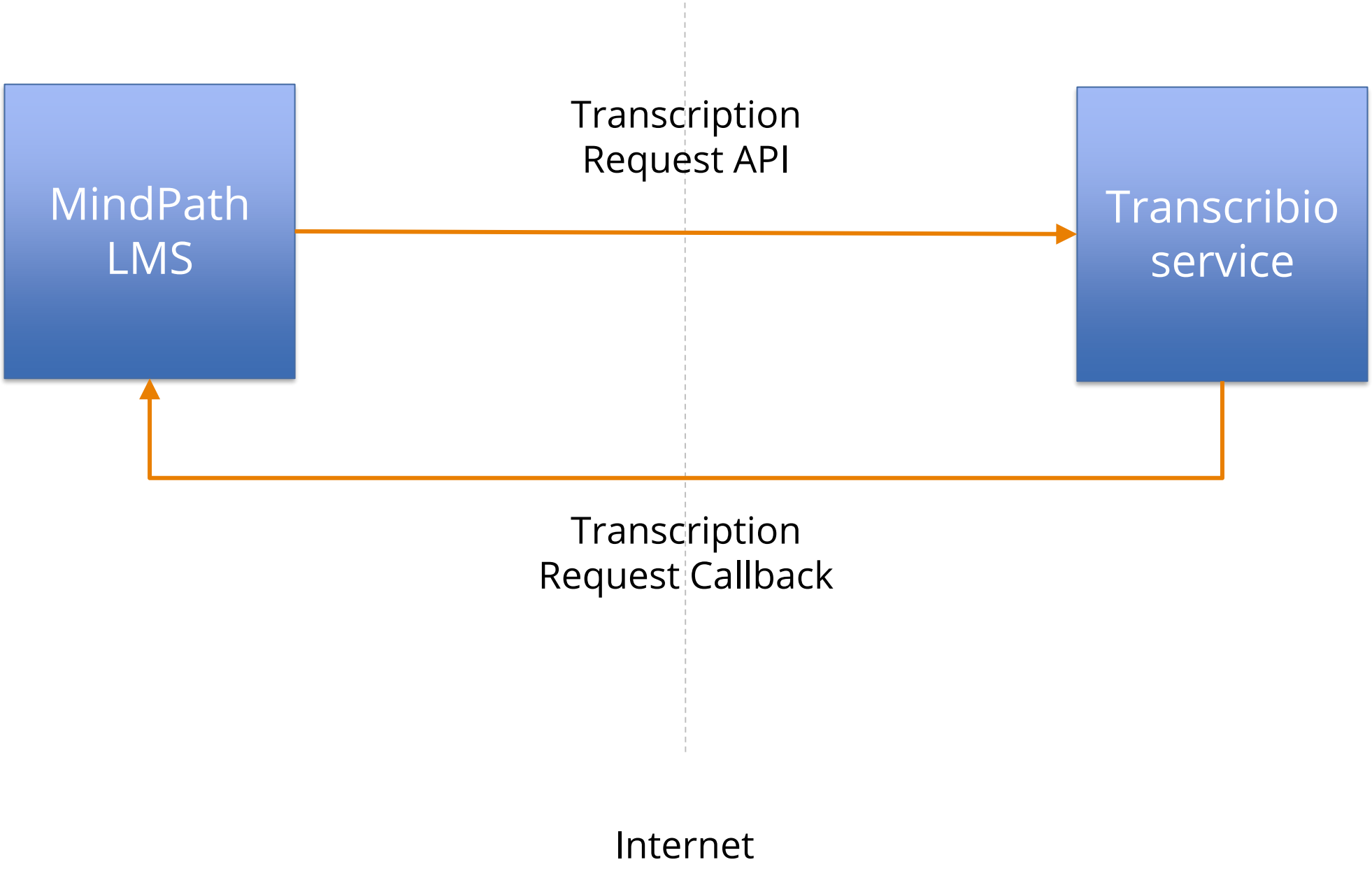
To: paige.scriber@mindpath.com
Subject: RE: Transcription discrepancy

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

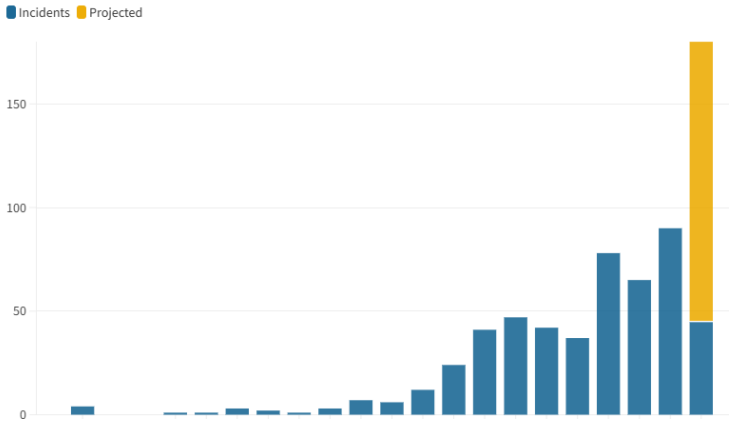

MindPath
LMS
Transcription
Request API
Transcribio
service
Transcription
Request Callback
Internet

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.

Resources: 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

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 Committee

Max Powers
CEO

Archie Tech
CTO

Wynn Moore
VP of Growth

Drew Diligence
Counsel

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

Preparation and Detection

- Vendor Contracts
  - Notification for major AI changes
  - Notification for AI-related incidents
  - Cooperation in incident investigations
- Training
  - Train staff to recognize AI-related incidents

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

Incident Response Policy Updates

We include AI risk when screening vendors to avoid unreliable partners and ensure contractual obligations for handling incidents responsibly.

We train staff to recognize possible AI-related incidents.

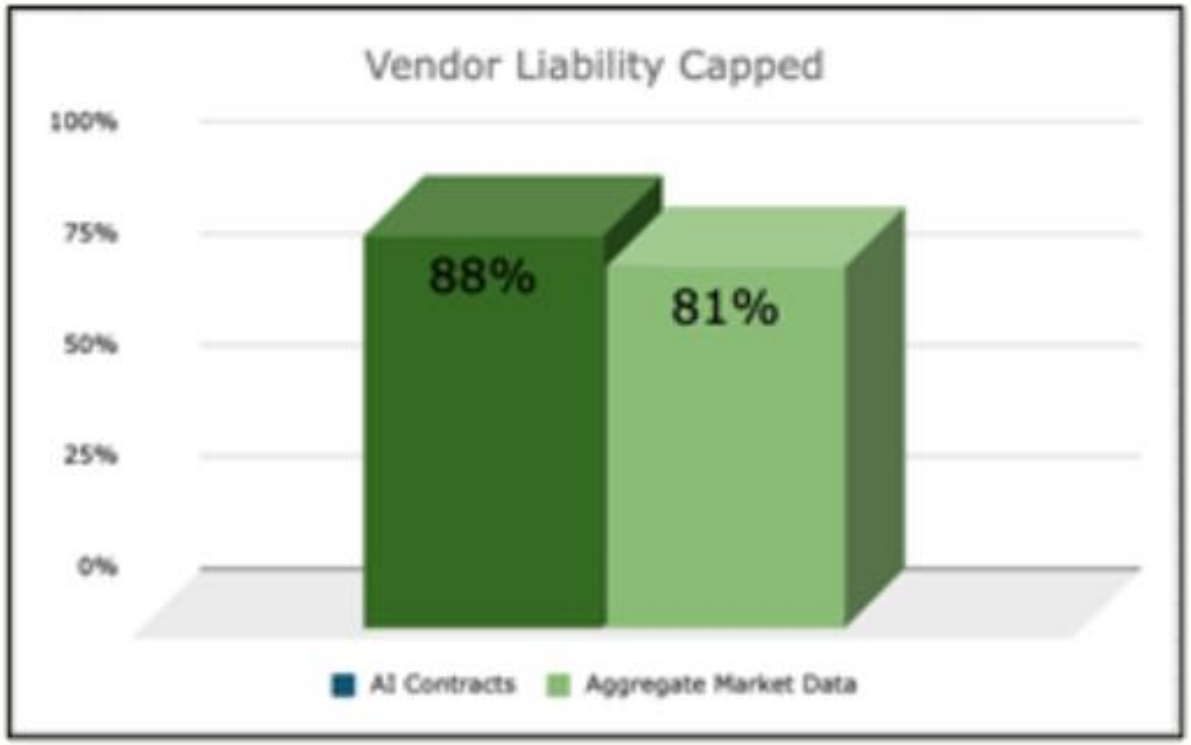
We have established an AI Oversight Committee to provide informed executive leadership when AI issues arise.

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

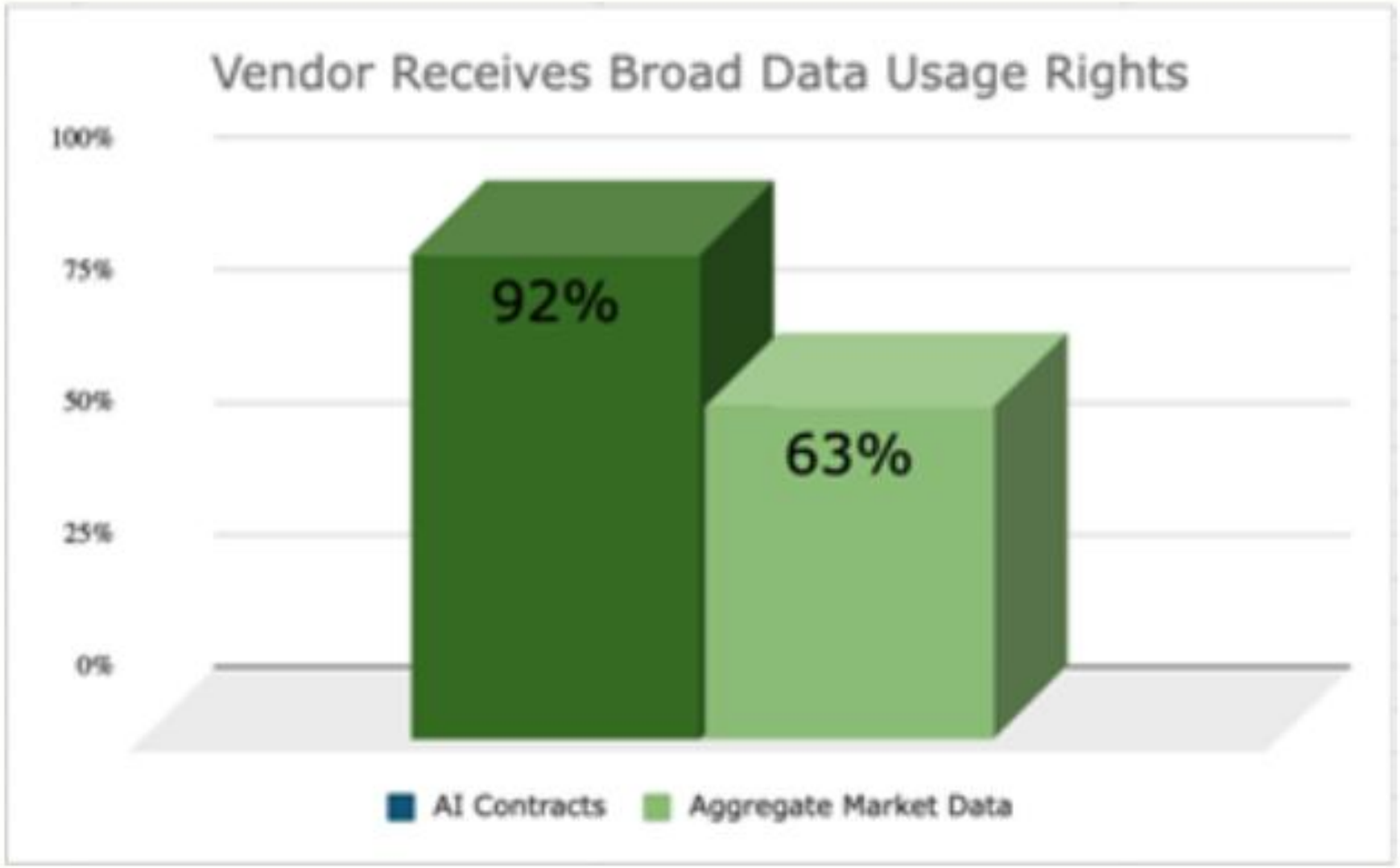
Source	Document
World Economic Forum	Adopting AI Responsibly: Guidelines for Procurement of AI Solutions by the Private Sector
Lexis Nexis	Artificial Intelligence Agreements Checklist
Morgan Lewis	Contracting Pointers for Services Incorporating the Use of AI

AI vs SaaS Contract Data

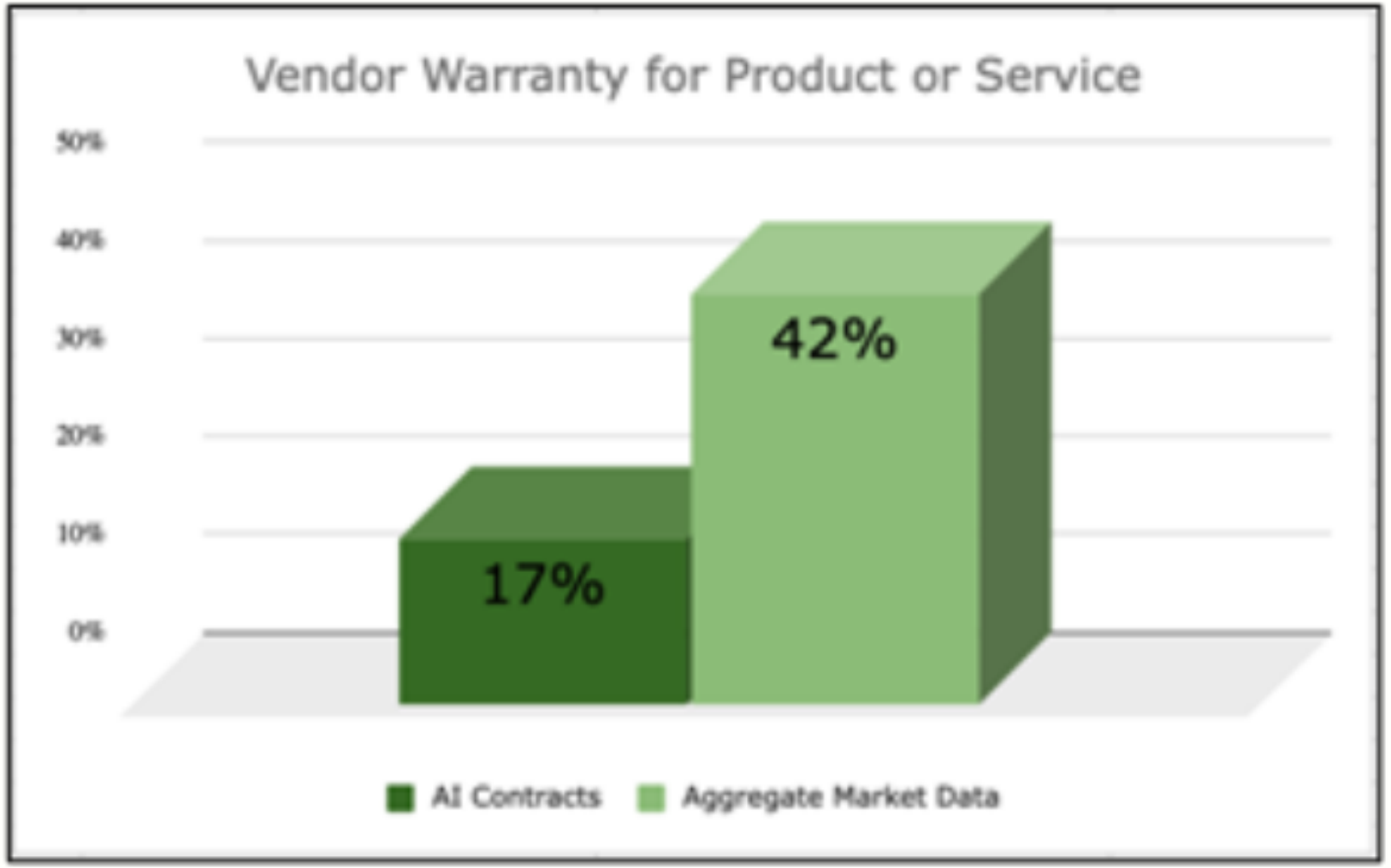


Stanford Law School, March 21 2025

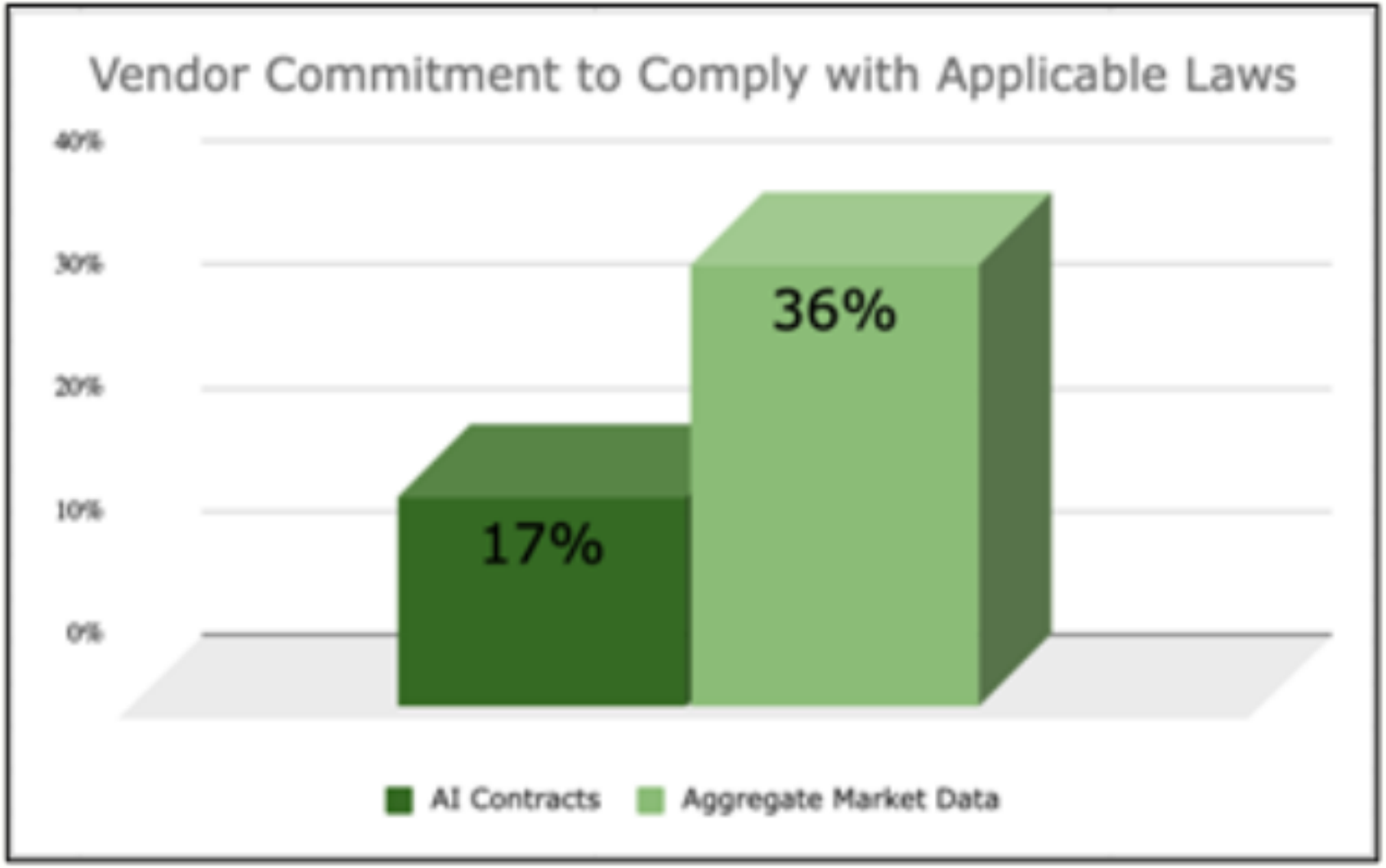
AI vs SaaS Contract Data



AI vs SaaS Contract Data



AI vs SaaS Contract Data



Contract Concerns

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

Vendor Review Resources

FS-ISAC

Generative AI Risk Assessment Guide

Includes detailed help creating a vendor questionnaire.

OneTrust

Questions to Add to Existing Vendor Assessments for AI Checklist

VenMinder

Artificial Intelligence Sample Vendor Questionnaire

Resources: AI Vendor Risk

OneTrust: Questions to Add to Existing Vendor Assessments for AI (2pp)

data sources, data quality. informed consent…

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

training process, frequency of model validation, subcontractor access to data…

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?

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

Meet ✨ 🛠️ Fixie Pixie ✨ 🛠️

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.

RAG Resources

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

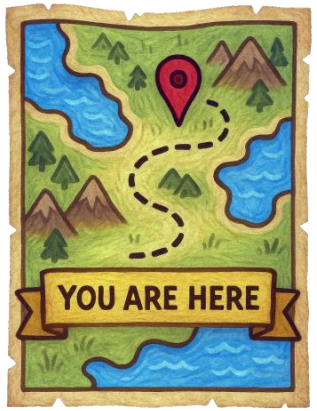
A few lines of python code to get “hello world” RAG code running on your own computer

[Build a Retrieval Augmented Generation (RAG) App]
(LangChain, 2024?)

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

[A Comprehensive Guide to RAG Implementations]
(Armand Ruiz, 2024)

A useful list of RAG variations.



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.

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

Help MindPath staff find internal information using natural-language questions. Explore RAG (Retrieval-Augmented Generation) capabilities.

Out of Scope

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

Success Criteria

- Delivers useful, relevant answers to internal users
- Build and release process is automated
- Internal documentation is complete
- Team gains confidence to plan a customer-facing version later

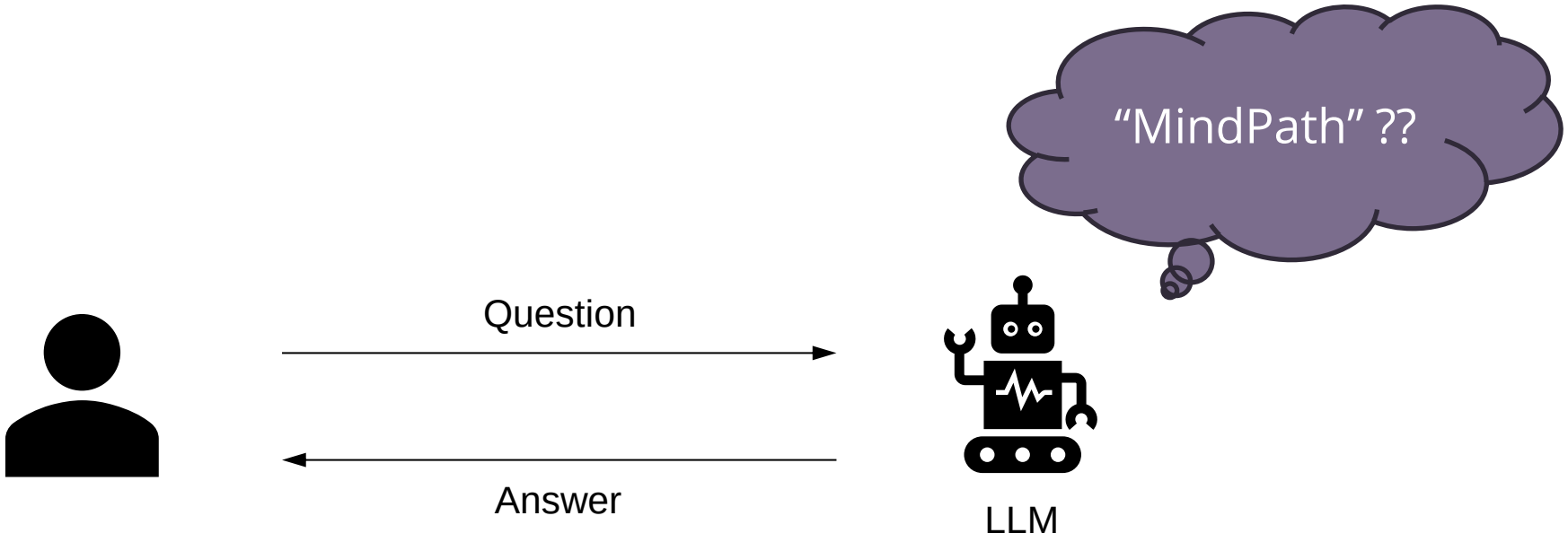
Internal Chatbot Milestones

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.)

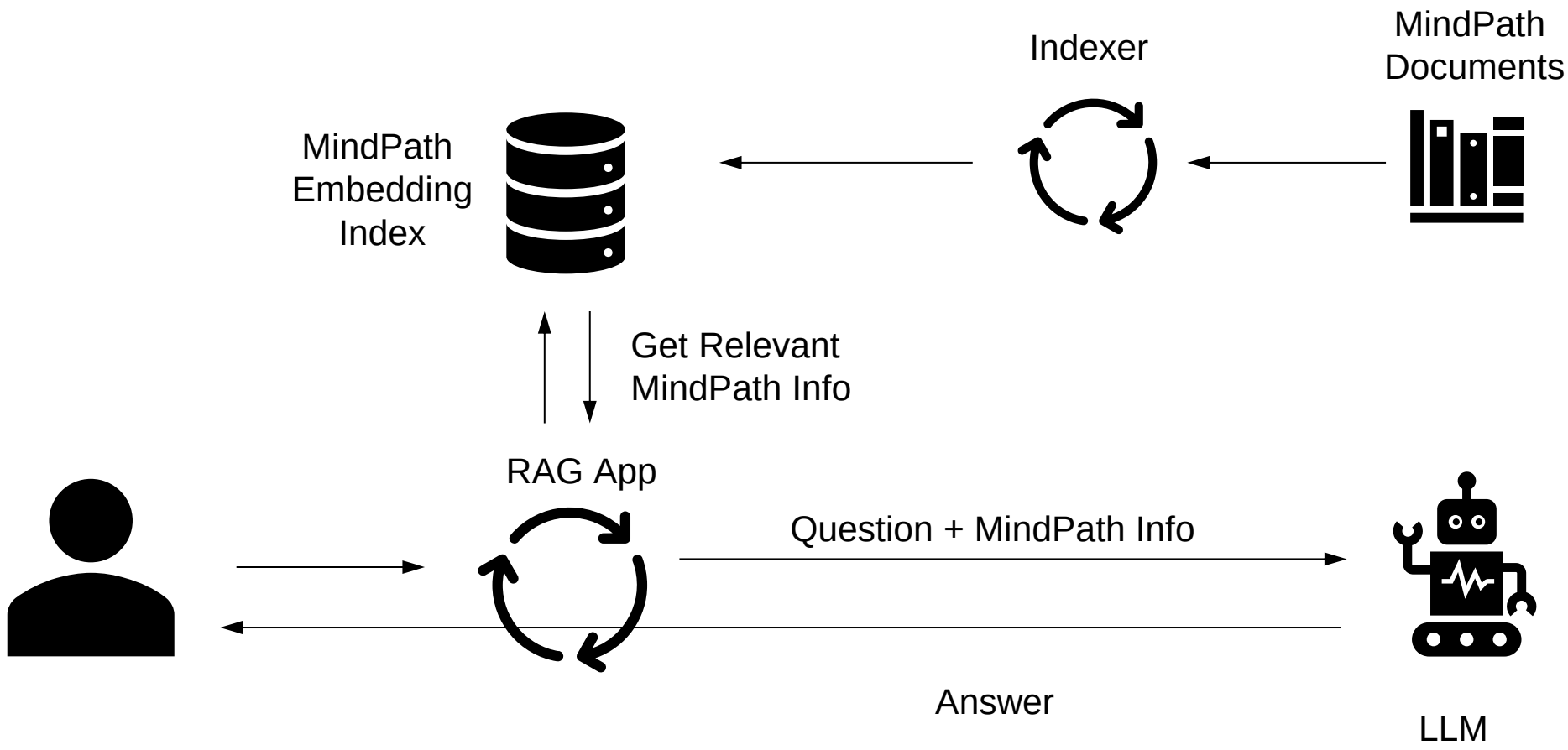
The team considers what might go wrong.

ARCHITECTURE REVIEW

Normal LLM Interaction



Retrieval-Augmented Generation (RAG)



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)

Vulnerability	Consequences	Treatment
Hallucinations	False/misleading answers	UI disclaimer. User education.
Corrupted data sources	Same as above	Secure the indexing process and embedding data.
Out-of-date documents	Same as above	Automate ingestion. Encourage document clean-up.
Testing non-deterministic system	Unpredictable bot behavior	Manual tests. Remaining risk accepted.
Input corpus unmanaged	False or misleading answers Sensitive data exposure	Restrict to Public / Internal data sources.
Insider threat	All of the above	Accepted; preexisting risk.
Rush to productize	Serious harm if exposed externally.	Review risk with executive leadership.
Untraceable behavior	Hinders investigation.	Log time, user, input, response.
Prompt injection; Embedding leakage; Trust misalignment; retrieval failure.	Various	Accepted. (Internal use only)

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 Apache 2.0	Archie Tech	5/15/25
Natural Language Processing Group (HKU)	InstructorXL	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 Apache 2.0	Archie Tech	5/15/25
Mistral AI SAS	Mistral	Chatbot corpus	Internal	Low	Low	Allow	Hosted locally Apache 2.0	Archie Tech	5/15/25

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

What is the session timeout for our LMS?

The current session timeout is 15 minutes.

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?

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)

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.

Sources:

- Integration Meeting Notes – March 2024

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. Staff were reminded not to put Restricted data in wiki pages and to report it if they see it.

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.

LOOKING AHEAD

Lessons Learned for 2.0

- Plan for document governance.
  - Include ongoing document grooming.
  - Improve data classification. Some Internal documents are safe for customers and some are not.
- Create user feedback and escalation processes.

Issues Not Addressed in HAL-E

- Security and privacy review.
- Robust testing.
- AI behavior guardrails.
- Adversarial threats.

MindPath Part 2: A New Dawn

EPILOG

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

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
  - 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

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

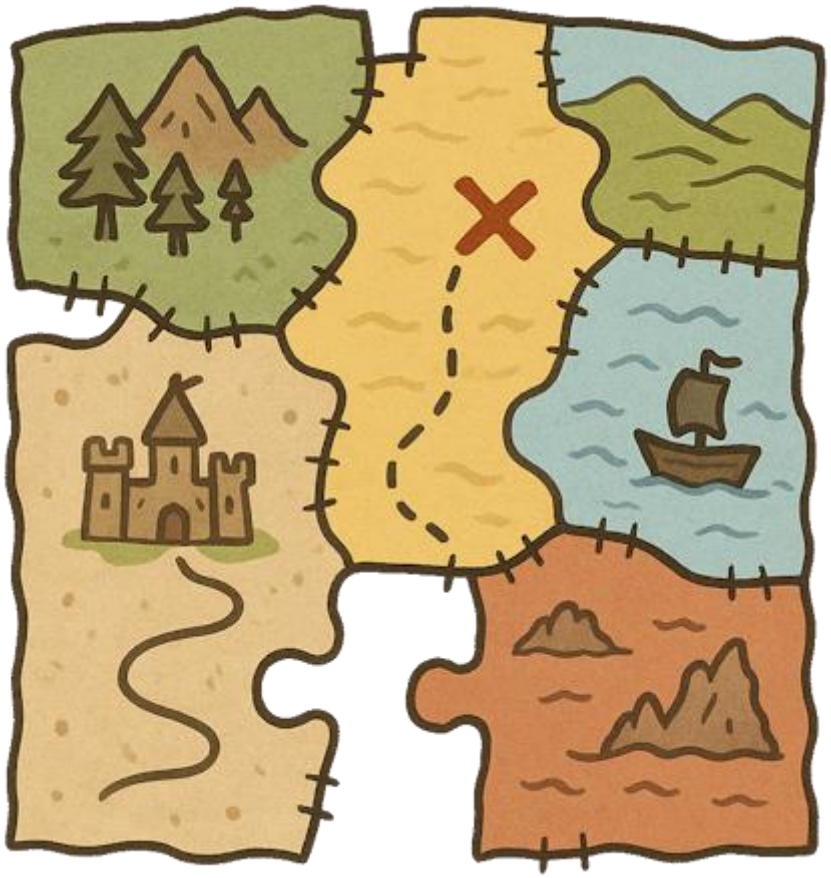
Best Practices & Emergent Practices

	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

	Classical	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
Evaluation Criteria	Accuracy, efficiency, consistency	Flexibility, innovation, ability to transfer knowledge
	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
CEO

Archie Tech
CTO

Wynn Moore
VP of Growth

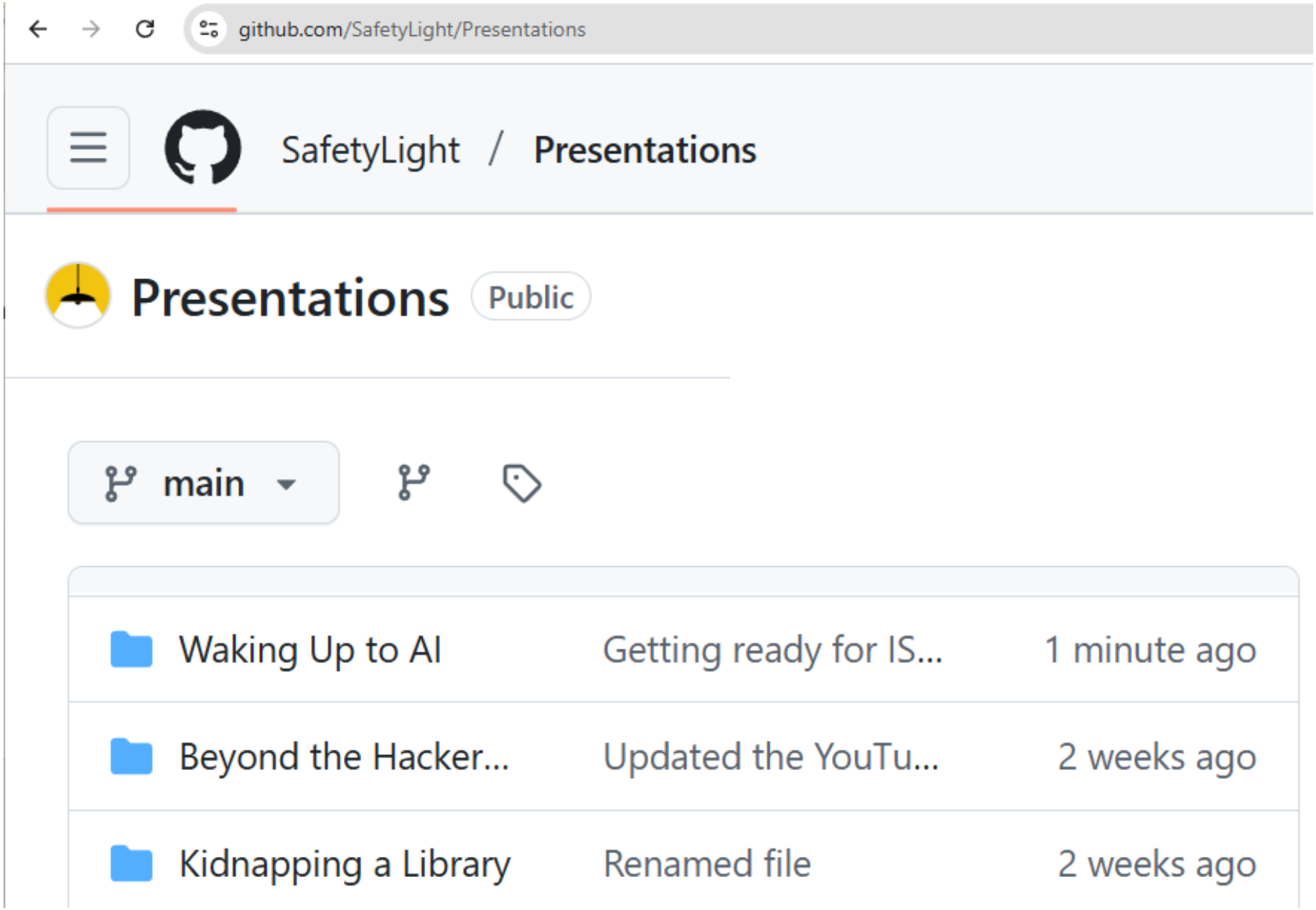
Drew Diligence
Counsel

Ruby Rails
Lead Engineer

Cody Commit
Software Engineer

Paige Scriber
Content Manager

HΛL-E MindPath
Oracle


github.com/SafetyLight/Presentations
SafetyLight / Presentations
Presentations
Public
main
Waking Up to AI
Getting ready for IS...
1 minute ago
Beyond the Hacker...
Updated the YouTu...
2 weeks ago
Kidnapping a Library
Renamed file
2 weeks ago

Any questions?

License and Attribution