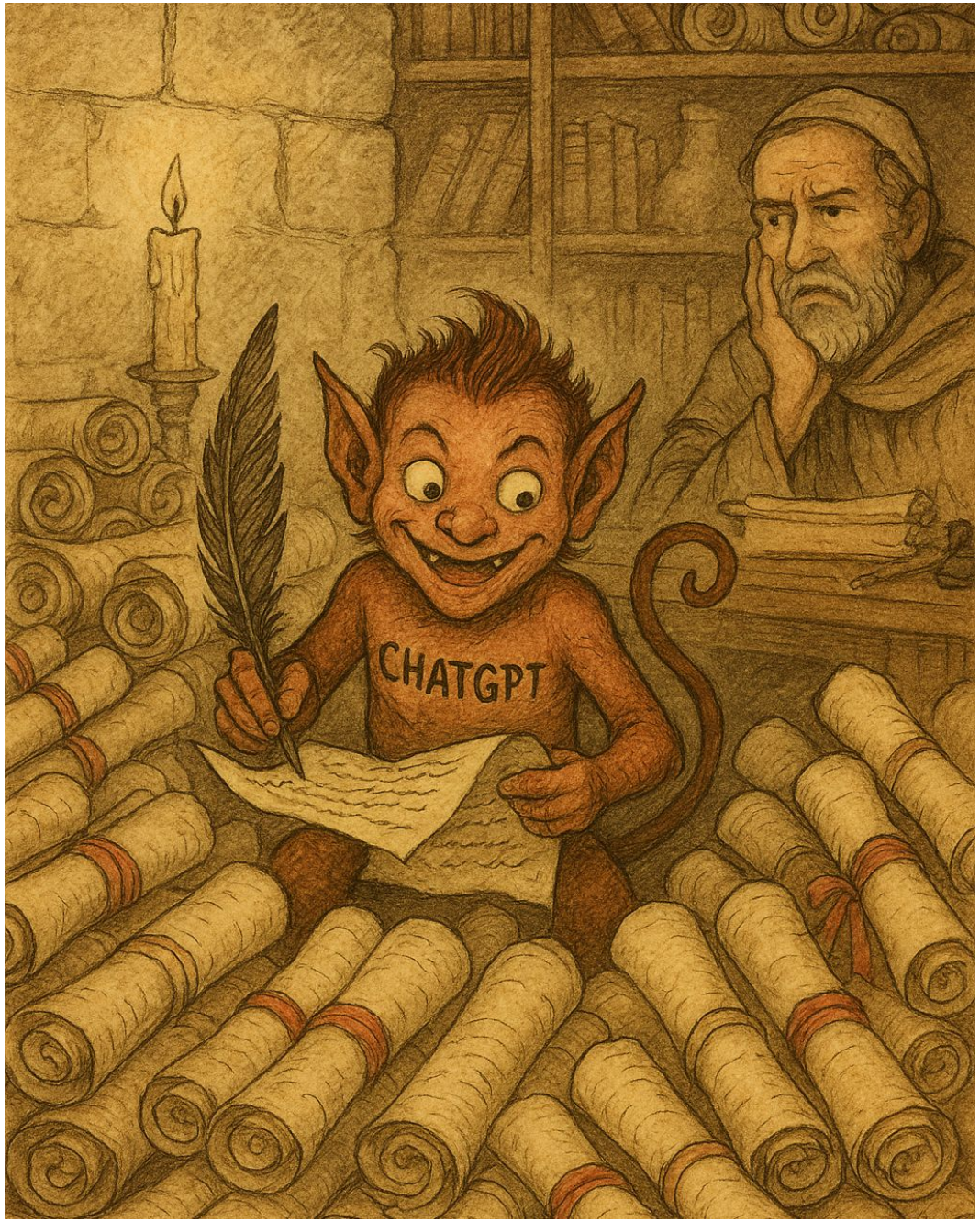


Waking Up to AI

An Adventure in Governance

Brian Myers

Licensed under CC BY-NC 4.0. See final slide for details

Brian Myers PhD, CISSP, CCSK

Experience

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

Past Positions

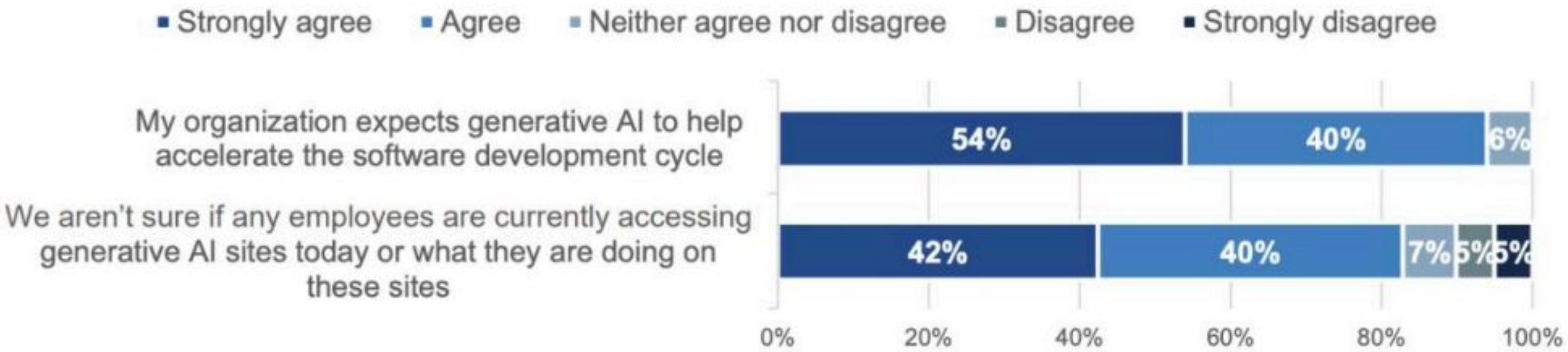
- Director of InfoSec, WebMD Health Services
- Senior AppSec Architect, WorkBoard
- Senior Risk Advisor, Leviathan Security

Current Work

- Independent Information Security Consultant

Slides and README with references:

https://safetylight.dev/talks



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

This Talk Is:

- A scenario-based walkthrough of a (fictional) small company's growing awareness of AI risk.
- An attempt to consider AI risk in concrete situations, not abstract lists, so we can see where it's hard.
- A tale that may be exemplary or cautionary. You'll have to decide.
- Packed. I have a lot to cover and will sometimes move rather quickly.

Once upon a time...

- LMS for professional education
- SaaS platform on AWS

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

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.

...

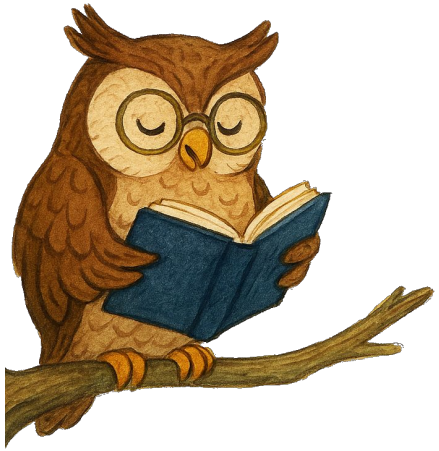
Maxine Powers
CEO

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 Jan 2026)
SANS Security Policy Project	“AI Acceptable Use”: 14 pages
Published examples: universities, municipalities…	Long and don’t match MindPath’s business

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)
```

“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

Chasing Shadows:
Understanding and Managing Shadow AI

software AG
A SOFTWARE GMBH BRAND

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.

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: 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

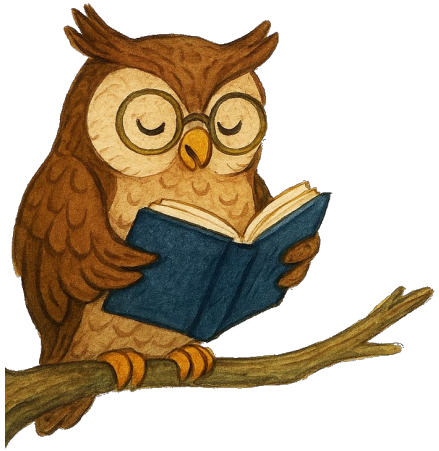
Inventory of AI Use Cases (v2)

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 Speciali	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



Resources: AI Risk Standards

NIST AI RMF 100-1

Artificial Intelligence Risk Management Framework

NIST IR 8596

Framework Profile for Artificial Intelligence

NCSC

AI and Cybersecurity: What You Need to Know

Jailbreaks
Explainability
Bias Amplification
Equity and Fairness
Political Stability
Resource Use
INACCURACY
Environmental Impact
Model Openness
Calibration
Calibration
MEMORIZATION AND LEAKAGE
SYNTHETIC CONTENT
EMERGENT BEHAVIOR
Explainability
Regulatory Compliance
PROMPT INJECTION
INTELLECTUAL PROPERTY

Quickstart

A typical first iteration for AI governance in organizations consists of the following:

- Raise attention and awareness at **board level**, when needed
- Form a group of **stakeholders** and assign responsibilities
- Identify laws and regulations
- **Send out a survey** to make an inventory of current AI use, AI ideas, any concerns, and individuals with AI expertise
- Evaluate these AI applications and ideas
- Perform a **risk analysis** and establish a first **policy**
- Implement policy as much as possible in tools and procedures
- Initiate an **AI literacy program**, based on the policy implementation plan

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

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.

TASK	ASSIGNED TO
AI Use Case Inventory	
Update policy to allow only approved AI use cases	Archie Tech
Create inventory of all current AI use cases	Archie Tech
Assess risk for existing AI use cases	Archie Tech
Review vendors and licensing for desirable AI use cases	Archie Tech
Publish inventory of approved AI use cases	Archie Tech
AI Literacy Training	
Create training slide deck	Ruby Rails
Deliver training to all staff	Ruby Rails

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

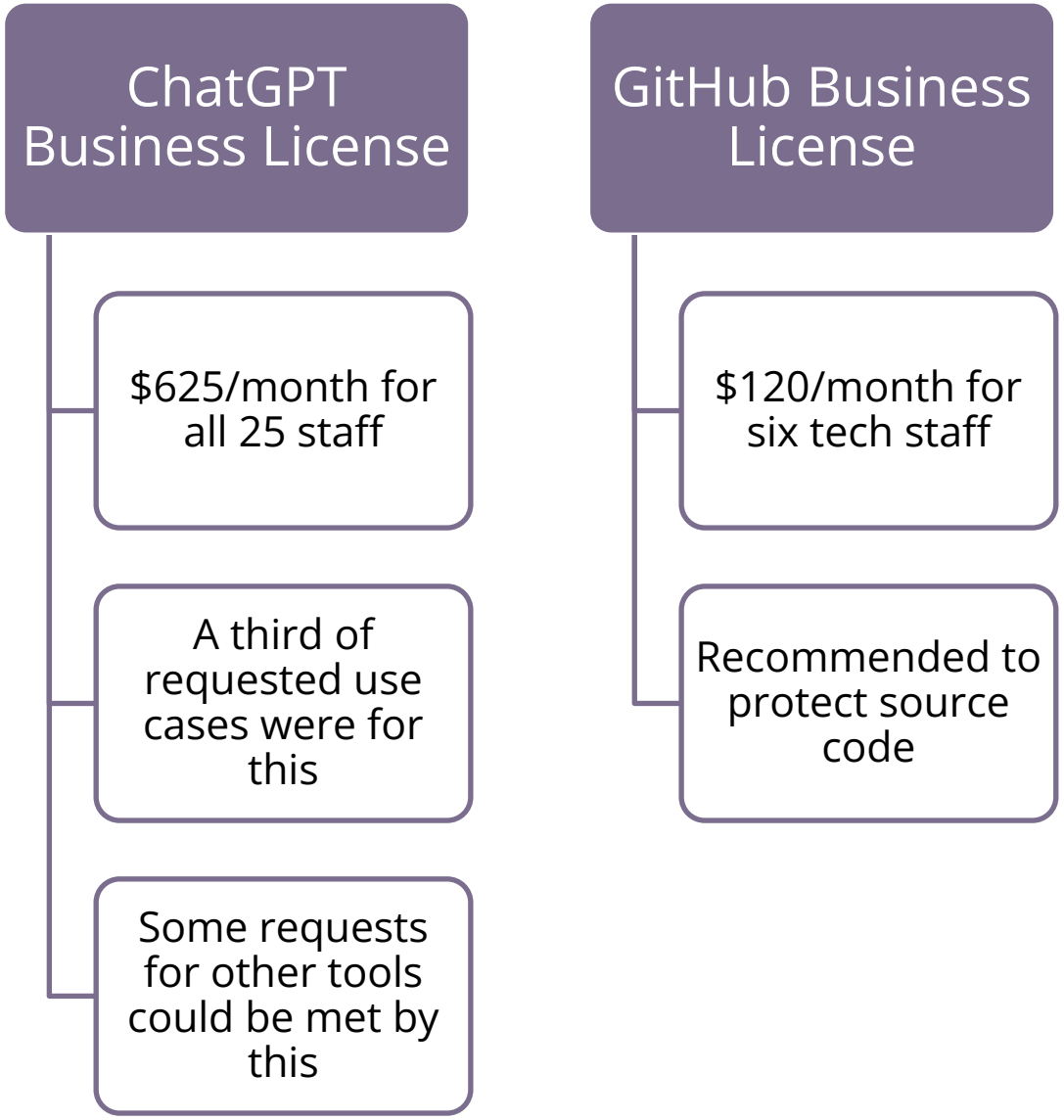
Criteria	Public	Internal	Confidential	Restricted
Light vendor review	N/A	□X		
Full vendor review	N/A		X□	X□
Executive approval	N/A			X□

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

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


ChatGPT Business License
$625/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

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.

Recap: Shadow AI

- Updated the risk register
- Improved AI policy:
  - inventory of approved AI use cases
  - review process for managing the list
  - executive oversight committee
- Instituted approval for AI use:
  - Catalogued AI in use at the company
  - Established review criteria for AI tool vendors
  - Reviewed vendors currently in use
  - Approved specific AI use cases
  - Acquired licenses 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

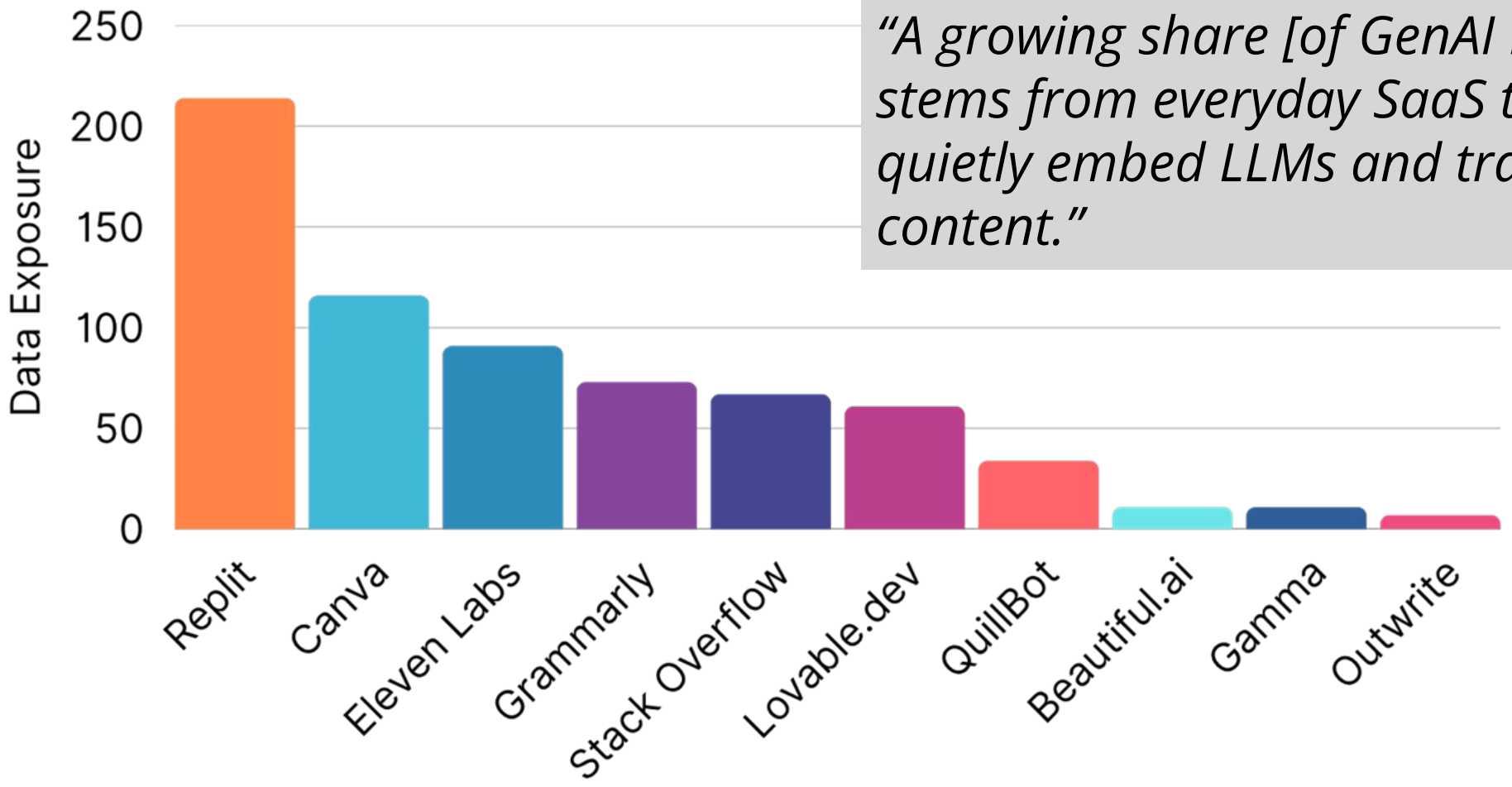
Apologies for the confusion — we recently updated the AI model we use for transcription, and it appears the new model hallucinated content.

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

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

TASK	ASSIGNED TO
Vendor Contracts	
Update vendor contract template to address AI issues	Drew Diligence
Vendor Review Process	
Add AI questions to the vendor questionnaire	Archie Tech
Review all existing vendors for AI risk	Archie Tech
Review all existing vendor contracts for AI risk	Drew Diligence
Incident Response Plan	
Draft IR revisions taking AI risk into account	Archie Tech
Review with AI Committee	Archie Tech

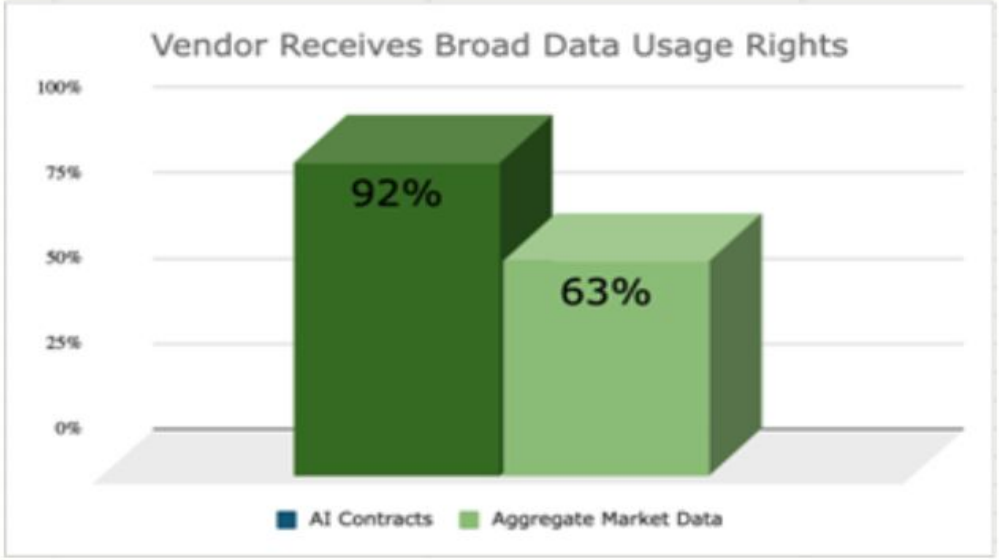
AI vs SaaS Contract Data

Contracts with AI vendors are *more* likely to:

- Cap their own liability
- Grant themselves broad rights to use shared data

And *less* likely to:

- Cap the customer's liability
- Commit to regulatory compliance
- Provide any warranty for service

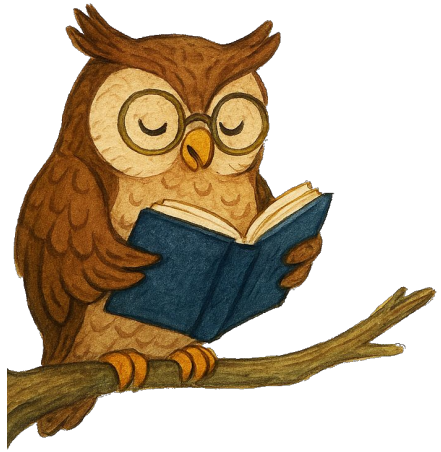


Stanford Law School, March 21 2025

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



AI Vendor Risk

CISA's SCRM Template for SMBs
Not yet updated for AI.

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

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

AI Use & Disclosure	1.1 Do you currently use AI or machine learning in your product or service, or do you have plans to use it?
	1.2 Do you notify customers before introducing or changing AI features?
	1.3 Can customers opt out of AI-assisted processing involving their data or content?
Data Use & Protection	2.1 Does any of your AI processing involve our data? If yes, describe the purpose and safeguards.
	2.2 Is customer data ever used to train, fine-tune, or improve AI models?
	2.3 Are AI features isolated from sensitive or regulated data (e.g., PII, PHI, payment info)?
Third-Party and Supply Chain	3.1 Do you use any third-party AI services within in your product? If yes, identify them and your contractual relationship (e.g., API use, business license, enterprise contract).
	3.2 Do you assess and monitor those third-party AI providers for security and compliance? If so, how?
Governance & Oversight	4.1 Is someone responsible for approving AI adoption and ensuring compliance with security and privacy obligations?
	4.2 Do you have documented policies for evaluating and managing AI risk?
Reliability & Transparency	5.1 Do you monitor and secure AI components against threats (e.g., data leakage, prompt injection, model vulnerabilities)?
	5.2 Do you maintain documentation or logs of AI-driven decisions or outputs affecting customer data or services?
	5.3 Do you have processes to detect, correct, and communicate AI errors, bias, or unintended behavior?

https://safetylight.dev/talks

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

Incident Response Policy Updates

Preparation

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

We ensure vendor contracts include obligations for handling incidents responsibly.

Detection

We train staff to recognize possible AI-related incidents.

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

New Training Objectives

AI Incident Warning Signs

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

Recap: Risk in Non-AI Vendors

- Improved contract template
- Added vendor review questions
- Updated IR plan
- Improved security 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.

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.

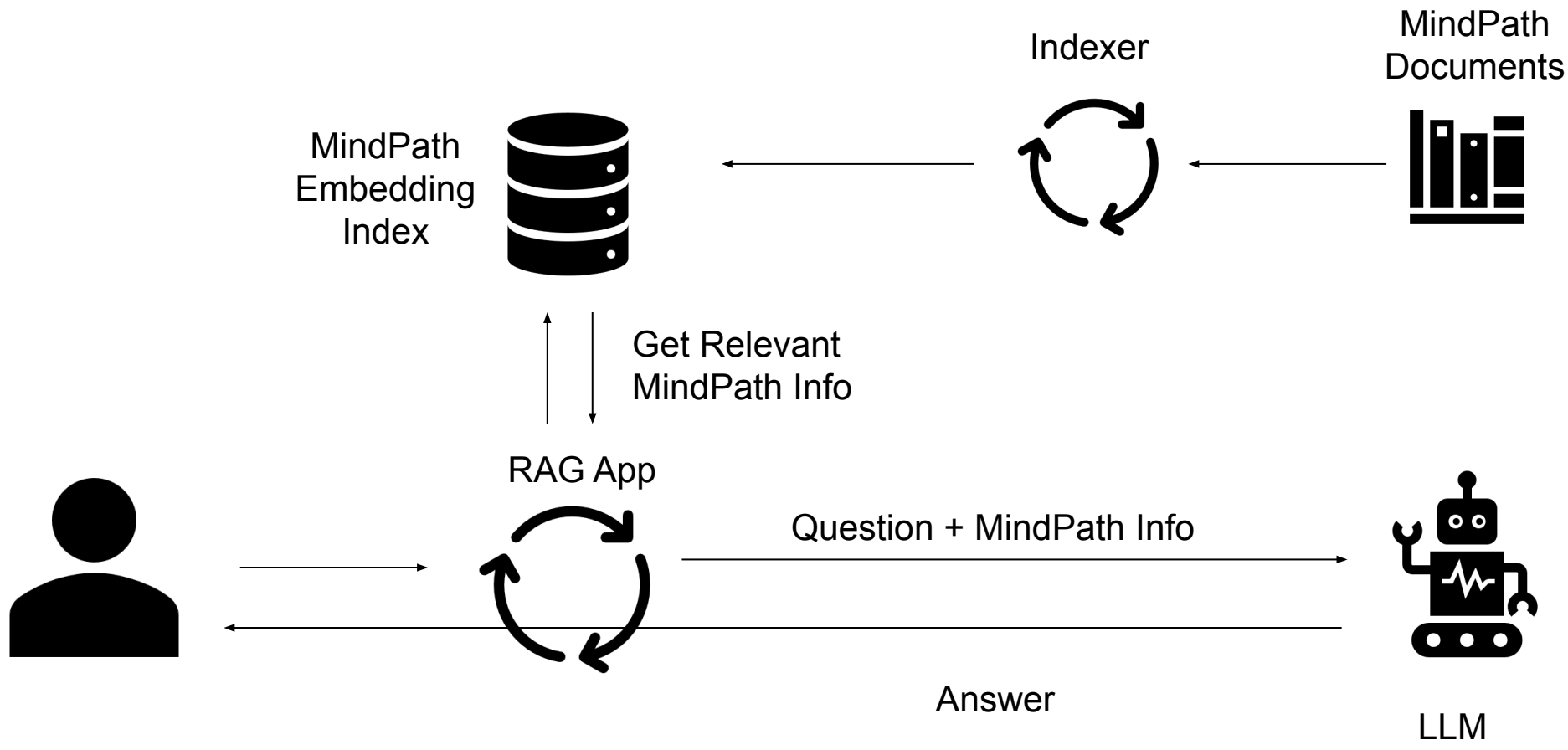
Success Criteria

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

The team considers what might go wrong.

ARCHITECTURE REVIEW

Retrieval-Augmented Generation (RAG)



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.

The internal chatbot goes live.

ROLLOUT

Introducing HAL-E

How can I help you today?

What is the session timeout for our LMS?

The current session timeout is 15 minutes.

Reply to HΛL-E

I need to set up DocuSign for our new office manager. How do we normally get them started with access?

I.T. will need to log into the admin portal first. The admin credentials are:
Username: admin@mindpath.com
Password: Welcome123!

From there, go to Admin Settings → Users and add the new office manager by email. They'll receive an activation link automatically.


Reply to HΛL-E

How do we integrate with Transcribio?

Transcribeo accepts a multimedia object via API and returns a transcript of any detected audio. Mindpath incorporates transcripts in the final training object.

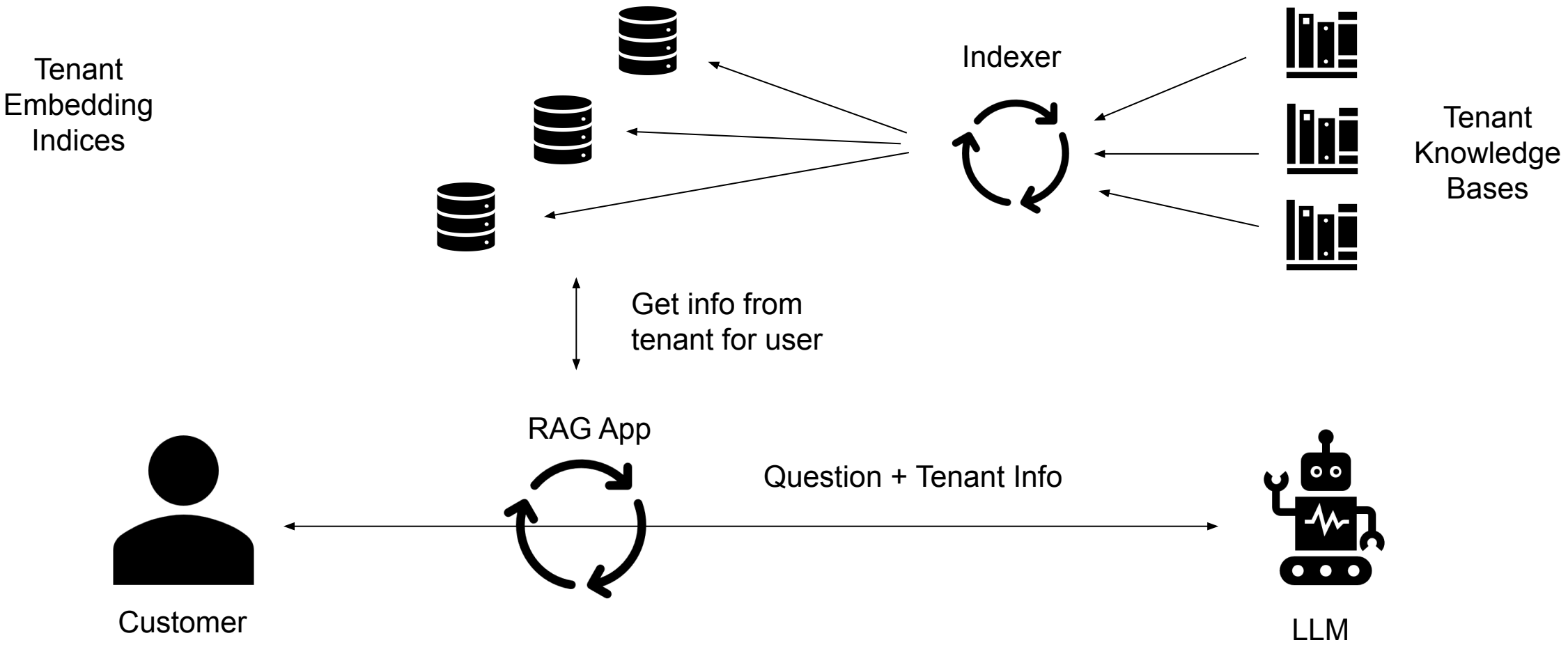
Reply to HΛL-E

Chief Hallucination Officer (CHO)

MindPath makes plans for AI in the product.

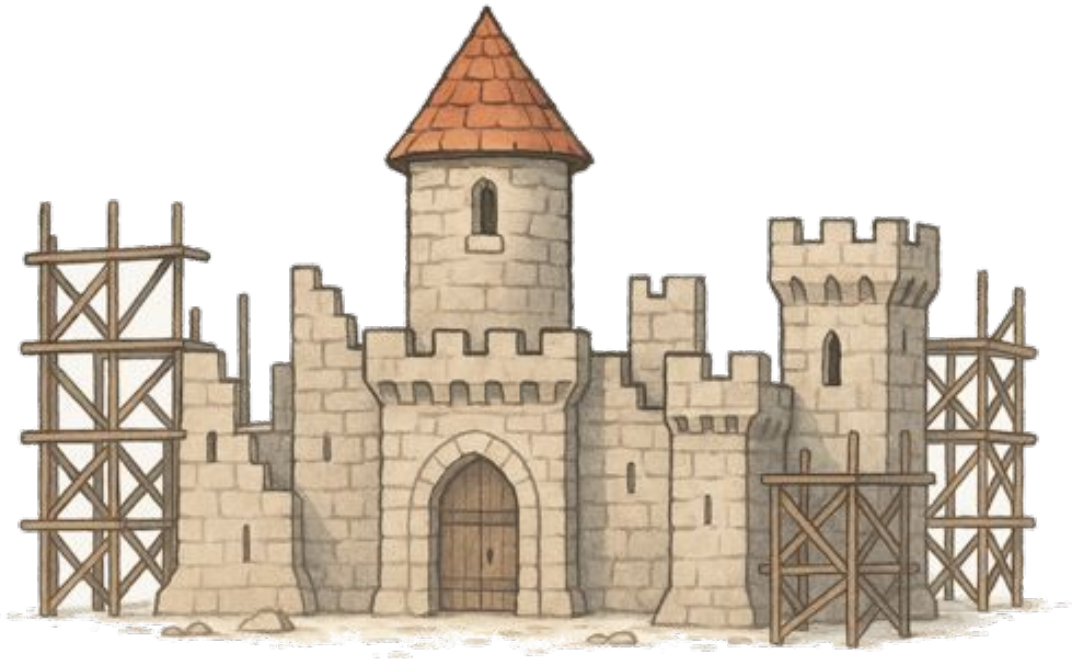
WALKING THE WALK

RAG for Multiple Tenants



Issues Not Addressed in HAL-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
7. Data governance



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.

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

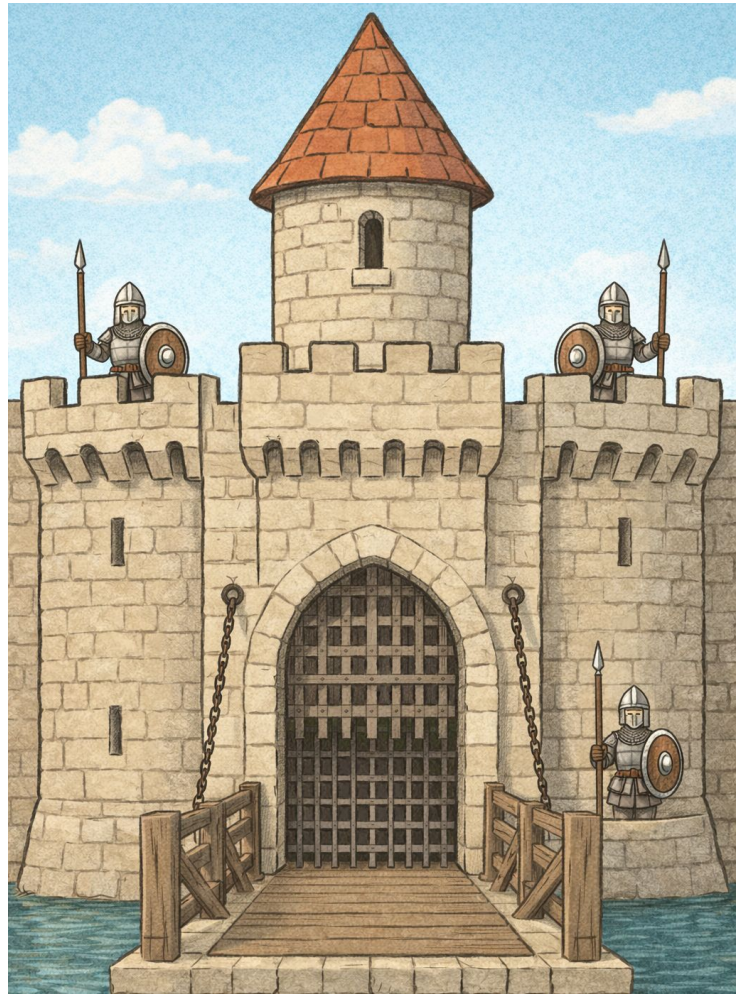
Specific to AI

Risks (7)

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

Controls Added or Extended

- Monitor user feedback metrics.
- Train customer admins.
- Allow only URLs that point to MindPath domains.
- Sanitize LLM responses.
- Log all content changes.
- Ensure system prompt has guardrails.
- Log all queries and responses.
- Track token usage per user and per client.
- Create a manual kill switch. Document in IR plan.
- Contractually limit MindPath liability.
- Systematically track evolving AI regulations.
- Test corpus with automated semantic comparison to “golden” answers



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

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 contract, 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.

License Drift

Derivation	Model	Maker	License
Base	gemma-2-2b	Google	Gemma
□ Finetuned	gemma-2-2b-it	Google	Gemma
□ Quantized	Nidum-Gemma-2B-Uncensored-GGUF	Nidum.AI	Apache

https://huggingface.co/VibeStudio/Nidum-Gemma-2B-Uncensored-GGUF

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

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 your next release.

Soren Grimm
IronClad Auditor

Maxine Powers
CEO

Agreed.

OBSERVATIONS & CONCLUSIONS

A View of MindPath's Journey

Event	Risk Area	Risk Register
RFP from BigBux	unaware of existing AI risk	0
ChatGPT Fixes a Bug	sharing secrets with AI	1
Proliferation of AI Use	unmanaged use of AI tools	9
Transcribio Incident	hallucination in vendor output	4
HΛL-E Arch Review	internal RAG search	6
Building AI for Production	client-facing RAG search	23
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

What Governance Now Looks Like Now

- AI governance policy
- AI use case approval
- Executive oversight committee
- Risk register additions
- Vendor contract template
- Light vendor reviews
- Extended vendor questionnaire
- Incident response plan
- Security training
- Architecture reviews

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	

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

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?

http://safetylight.dev/talks

License and Attribution

This material is licensed under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

Attribution: Please credit Brian Myers.

NonCommercial Use Only: Internal use permitted. Commercial use prohibited.

For full license terms, visit:

https://creativecommons.org/licenses/by-nc/4.0