Lessons from building verticalized AI agents

2025-08-13

Dr Chris Lovejoy, MD

Building vertical AI

Jason Liu

AI Consultant, 567

LESSONS FROM BUILDING VERTICALIZED AI AGENTS

1. How domain experts can supercharge AI development - and why to build them a custom UI
2. Why prompting beats finetuning for verticalized agents - and how to go beyond ‘prompt engineering’
3. How to get and maintain customer trust - through intelligent performance monitoring and a secure LLM-native architecture

+ BONUS LESSON (if time)

My background

- **Trained in medicine at University of Cambridge** (UK), worked as a doctor
- **Established the AI function at Cera Care** (tech-enabled home care, now at $500m ARR)
- **Built the first blood glucose prediction algorithm** using non-invasively measured data (at ZOE)
- **Founded 2 companies** (medical AI research summarization, health data infrastructure)
- **Founding team and Head of Clinical AI at Anterior** (Sequoia-backed start-up using LLMs for healthcare admin)
- **Helped >10 companies as applied AI consultant** across education, healthcare, recruiting, retail

Why is it hard to successfully apply LLMs to specialized industries?

(1) The last mile problem

The challenge is no longer in quality of the reasoning

It's in the context the model has on how a workflow is performed

(2) It’s harder to define what’s good and/or correct

You often need a domain expert to “translate”

LESSONS FROM BUILDING VERTICALIZED AI AGENTS

1. How domain experts can supercharge AI development - and why to build them a custom UI
2. Why prompting beats finetuning for verticalized agents - and how to go beyond 'prompt engineering'
3. How to get and maintain customer trust - through intelligent performance monitoring and a secure LLM-native architecture

+ BONUS LESSON (if time)

“look at your data”

“look at your data”

The agreement dated March 15, 2023 contains a clause stating 'Party A shall deliver all goods within 30 days of execution, subject to force majeure events beyond reasonable control.' The contract further specifies 'failure to deliver constitutes material breach warranting immediate termination.' This establishes clear performance obligations and remedy provisions.

“look at your data”

The agreement dated March 15, 2023 contains a clause stating 'Party A shall deliver all goods within 30 days of execution, subject to force majeure events beyond reasonable control.' The contract further specifies 'failure to deliver constitutes material breach warranting immediate termination.' This establishes clear performance obligations and remedy provisions.

“but what does this mean?”

🤔

“look at your data” → “empower domain experts to look at (and translate) your data”

The agreement dated March 15, 2023 contains a clause stating 'Party A shall deliver all goods within 30 days of execution, subject to force majeure events beyond reasonable control.' The contract further specifies 'failure to deliver constitutes material breach warranting immediate termination.' This establishes clear performance obligations and remedy provisions.

“but what does this mean?”

🤔

Production application

AI outputs

Domain expert reviews

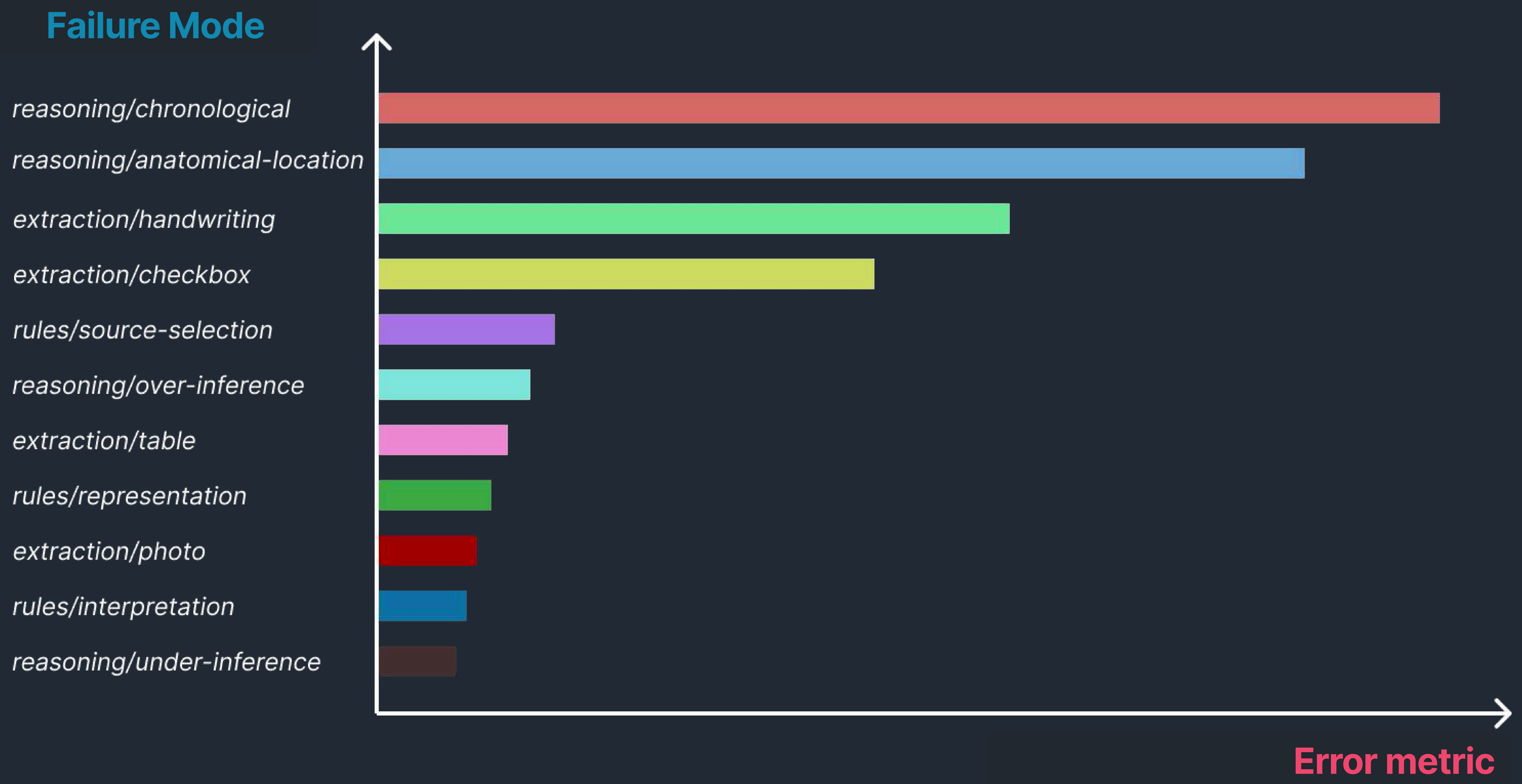
Performance metrics

Failure modes

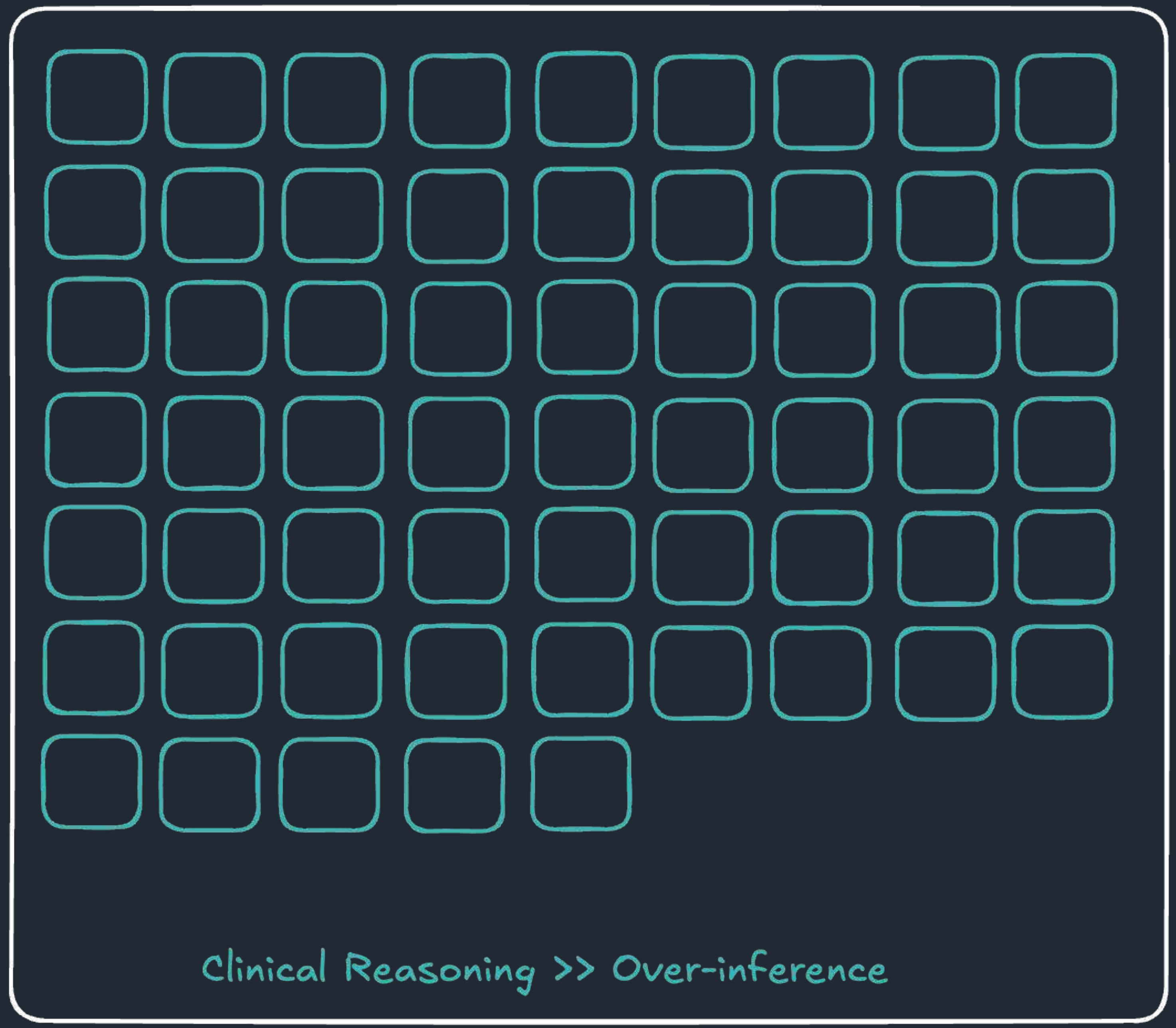
Suggested improvements

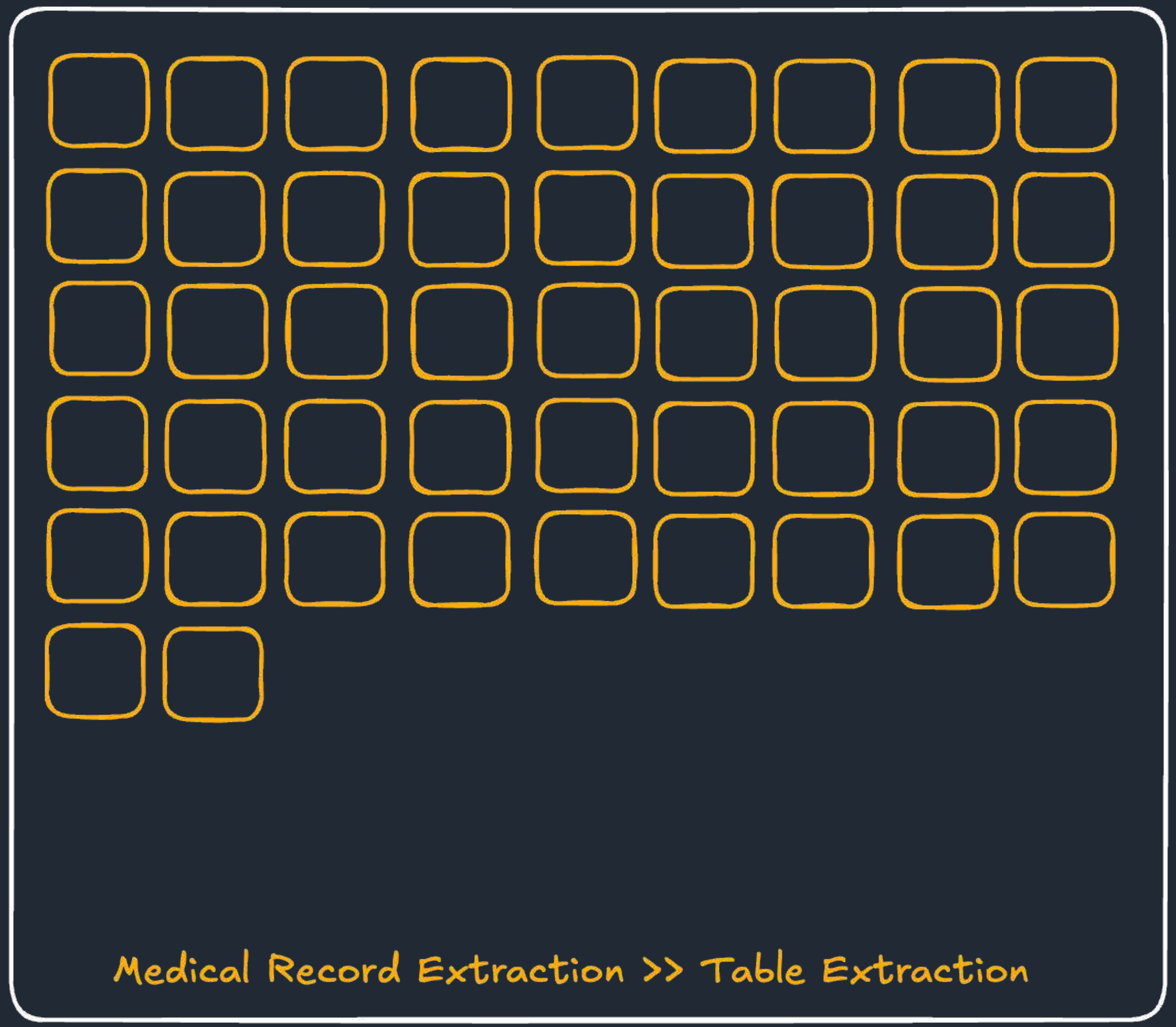
Input-output pairs

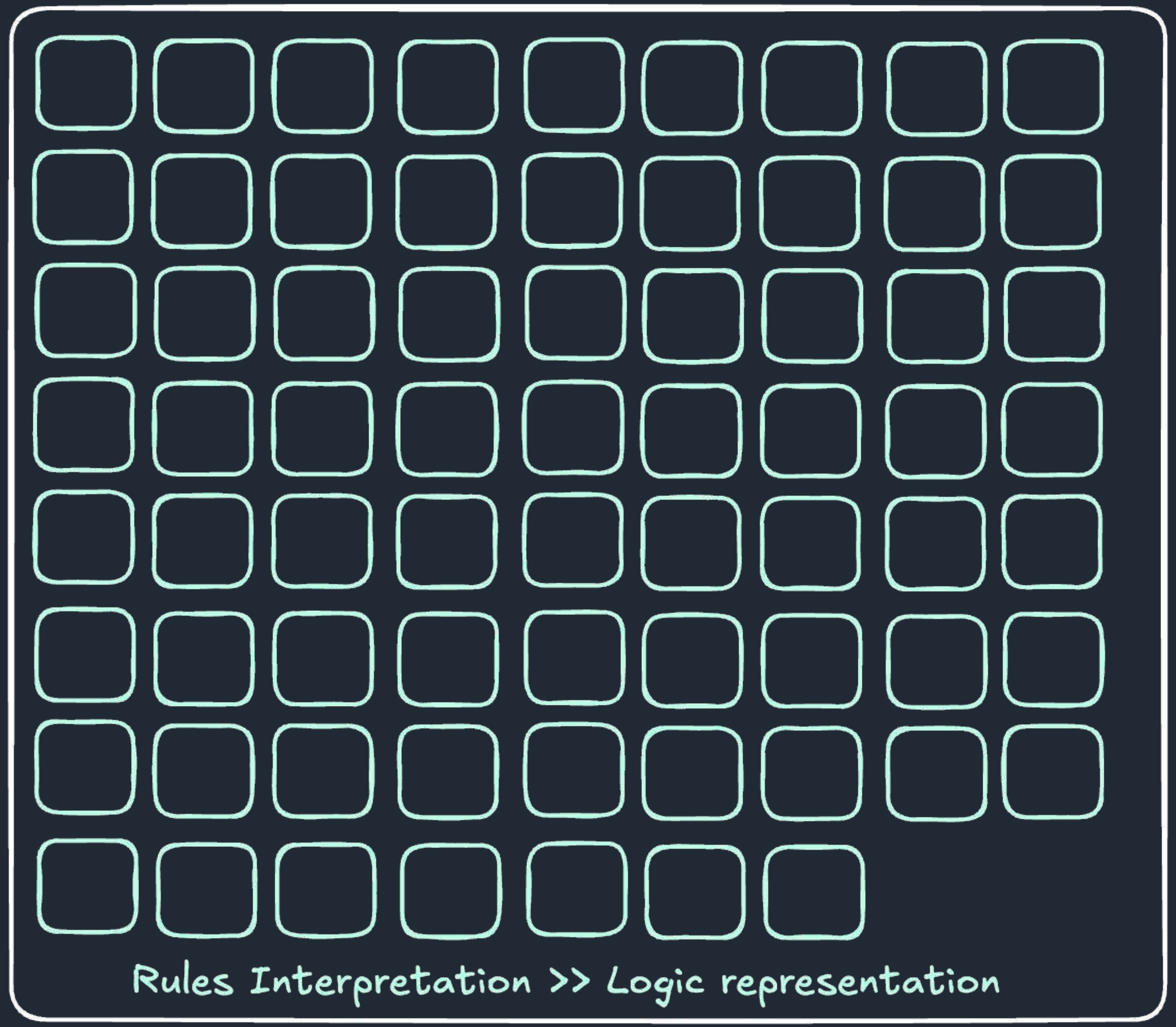
Combining metrics and failure modes from production helps prioritise work



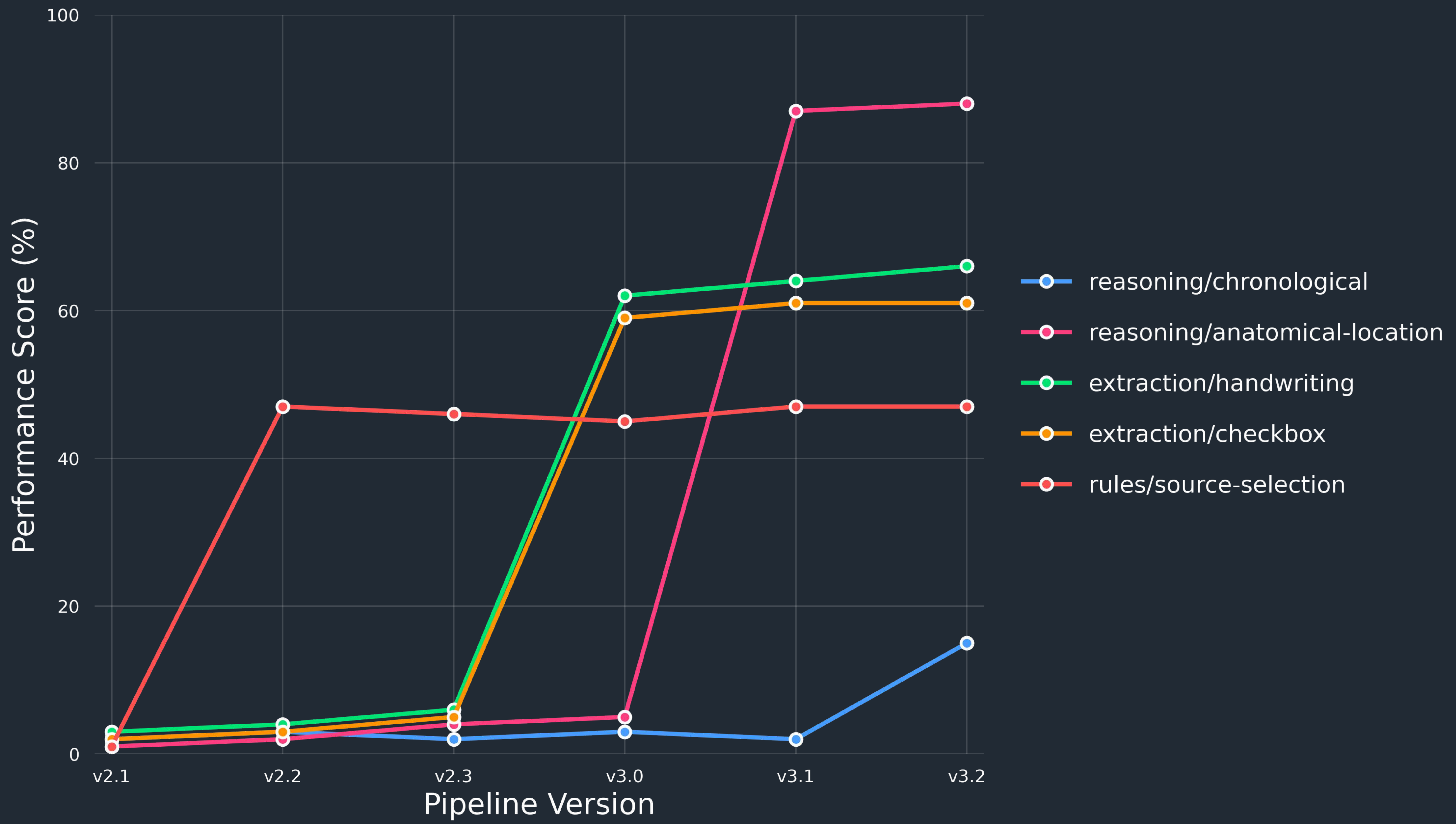
Ready-made failure mode datasets enable rapid AI iteration by testing the impact of changes



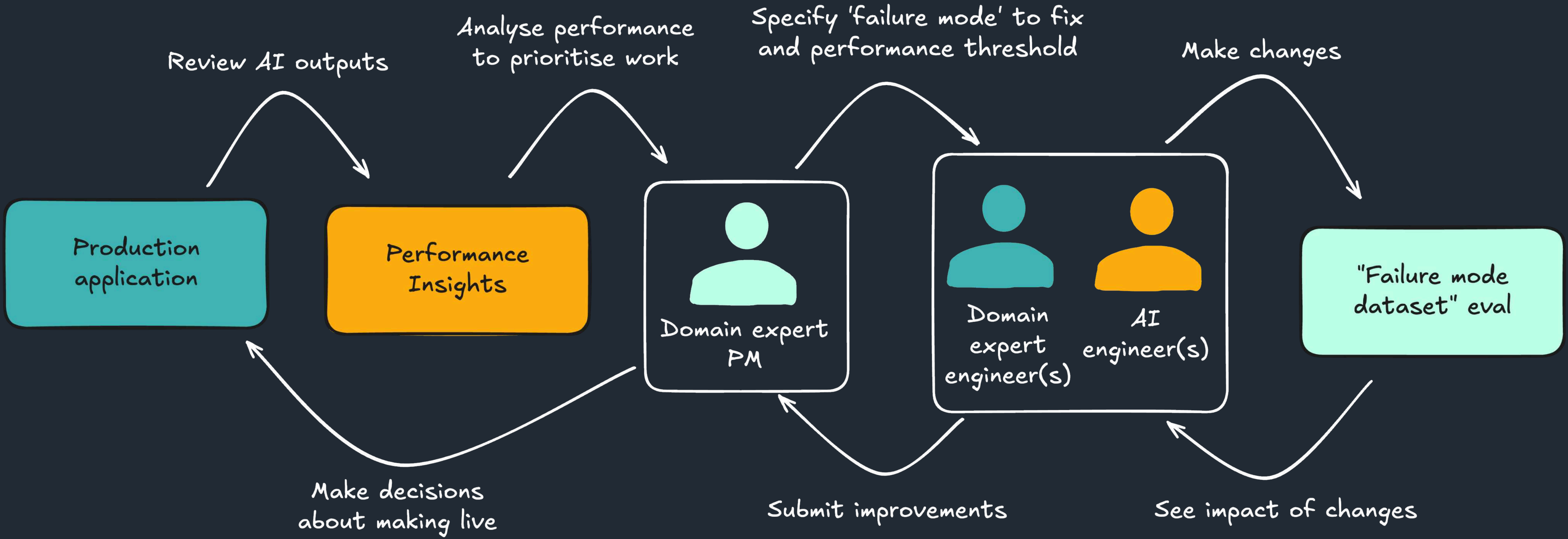




Ready-made failure mode datasets enable rapid AI iteration by testing the impact of changes



This process enables rapid iteration to give the AI system context on a specific workflow, solving the last mile problem.



Review AI outputs
Analyse performance to prioritise work
Specify 'failure mode' to fix and performance threshold
Make changes
Production application
Performance Insights
Domain expert PM
Domain expert engineer(s)
AI engineer(s)
"Failure mode dataset" eval
Make decisions about making live
Submit improvements
See impact of changes

Process described in detail here:

How to build an LLM-Native Expert System *- Dr Chris Lovejoy, AI Engineer World's Fair (SF, June 2025)*
https://www.youtube.com/watch?v=MRM7oA3JsFs

Production application

AI outputs

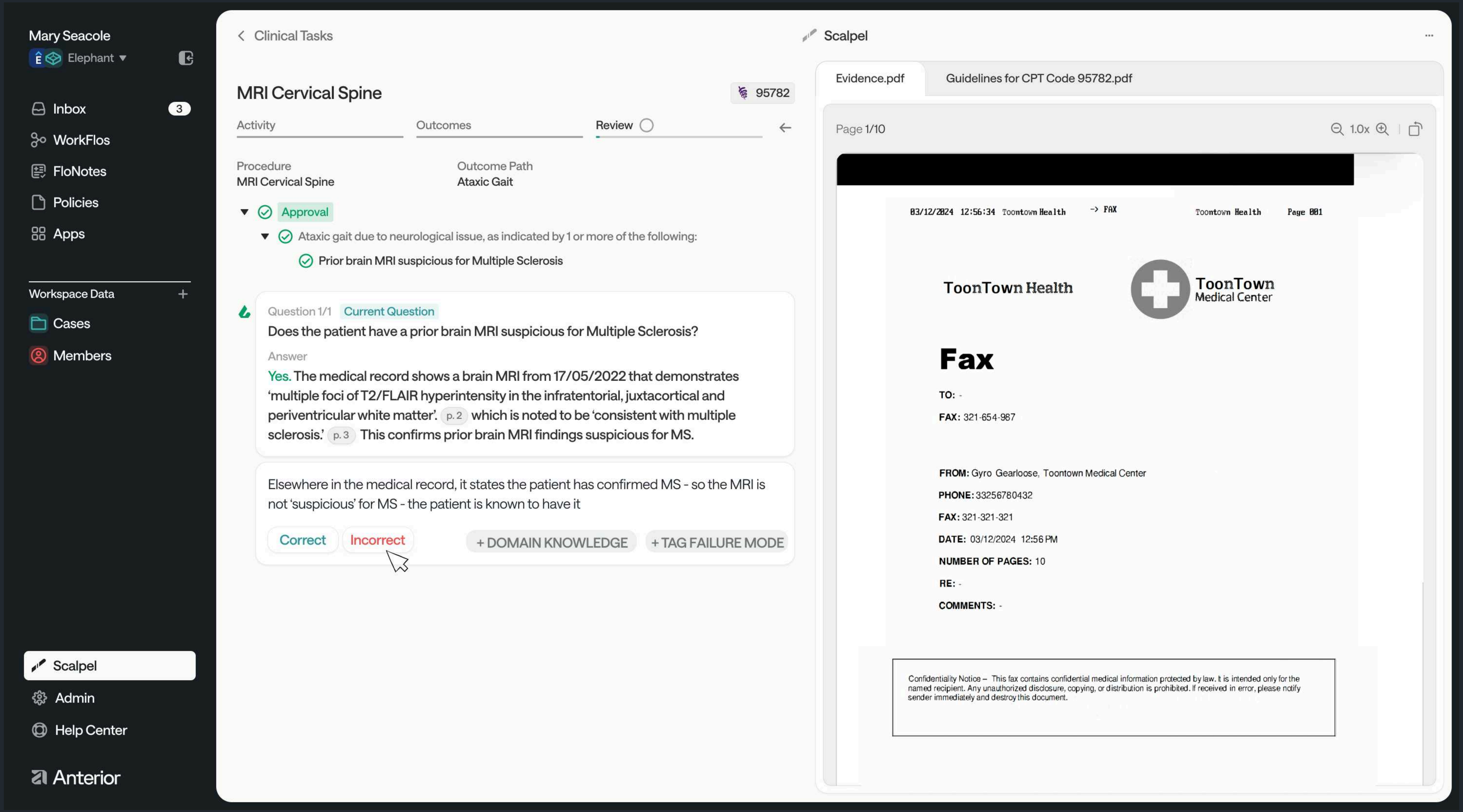
Domain expert reviews

Performance metrics

Failure modes

Suggested improvements

Input-output pairs



MRI Cervical Spine

95782

Procedure: MRI Cervical Spine

Outcome Path: Ataxic Gait

- ✓ Approval
  - ✓ Ataxic gait due to neurological issue, as indicated by 1 or more of the following:
    - ✓ Prior brain MRI suspicious for Multiple Sclerosis

Question 1/1 Current Question

Does the patient have a prior brain MRI suspicious for Multiple Sclerosis?

Answer

Yes. The medical record shows a brain MRI from 17/05/2022 that demonstrates 'multiple foci of T2/FLAIR hyperintensity in the infratentorial, juxtacortical and periventricular white matter'. p. 2 which is noted to be 'consistent with multiple sclerosis.' p. 3 This confirms prior brain MRI findings suspicious for MS.

Elsewhere in the medical record, it states the patient has confirmed MS - so the MRI is not 'suspicious' for MS - the patient is known to have it

[redacted]

03/12/2024 12:56:34 Toontown Health -> FAX Toontown Health Page 001

ToonTown Health

ToonTown Medical Center

Fax

TO: -

FAX: 321-654-987

FROM: Gyro Gearloose, Toontown Medical Center

PHONE: 33256780432

FAX: 321-321-321

DATE: 03/12/2024 12:56 PM

NUMBER OF PAGES: 10

RE: -

COMMENTS: -

Confidentiality Notice – This fax contains confidential medical information protected by law. It is intended only for the named recipient. Any unauthorized disclosure, copying, or distribution is prohibited. If received in error, please notify sender immediately and destroy this document.

Performance metrics

Failure modes

Suggested improvements

Input-output pairs

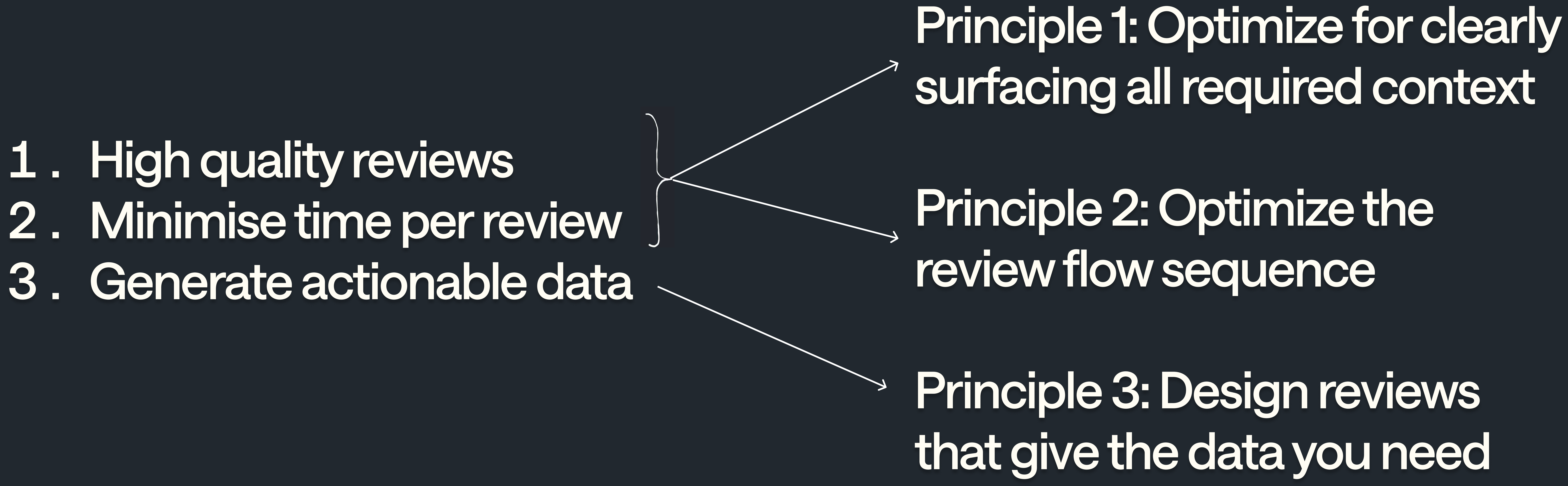
Optimize for 3 things:

1. High quality reviews
2. Minimise time per review
3. Generate actionable data

Described in more detail here:

Why custom AI review dashboards transform AI products - and how to build one *- Dr Chris Lovejoy*
https://chrislovejoy.me/review-dashboard

Optimize for 3 things:



1. High quality reviews
2. Minimise time per review
3. Generate actionable data

Principle 1: Optimize for clearly surfacing all required context

Principle 2: Optimize the review flow sequence

Principle 3: Design reviews that give the data you need

Described in more detail here:

Why custom AI review dashboards transform AI products - and how to build one *- Dr Chris Lovejoy*
https://chrislovejoy.me/review-dashboard

LESSONS FROM BUILDING VERTICALIZED AI AGENTS

1. How domain experts can supercharge AI development - and why to build them a custom UI
2. Why prompting beats finetuning for verticalized agents - and how to go beyond 'prompt engineering'
3. How to get and maintain customer trust - through intelligent performance monitoring and a secure LLM-native architecture

+ BONUS LESSON (if time)

FOR VERTICAL AI APPLICATIONS:

better prompting > finetuning models

But go beyond ‘prompt engineering’

- Over-optimizing prompts is brittle
- Better ways to improve the prompts in vertical applications:
  - **Context augmentation** (using domain-expert generated knowledge base)
  - **In-context learning** (using domain-expert generated input-output pairs)

Example Domain Knowledge (for Context augmentation)
[Regional Bank X] considers high-net-worth clients to include individuals with investable assets exceeding $1M, liquid net worth above $500K, and annual income over $200K for at least two consecutive years.
In credit agreements, 'availability' refers to the unused portion of a credit line, while 'utilization' is the drawn amount. A borrowing base facility's availability fluctuates based on eligible collateral values, typically updated monthly or quarterly
The CAMELS rating system for bank examination uses: - C (Capital adequacy) - A (Asset quality) - M (Management) - E (Earnings) - L (Liquidity) - S (Sensitivity to market risk) Each is scored 1-5 where 1 is strongest and 5 requires immediate supervisory attention.

Example Input-Output Pair (for in-context learning)

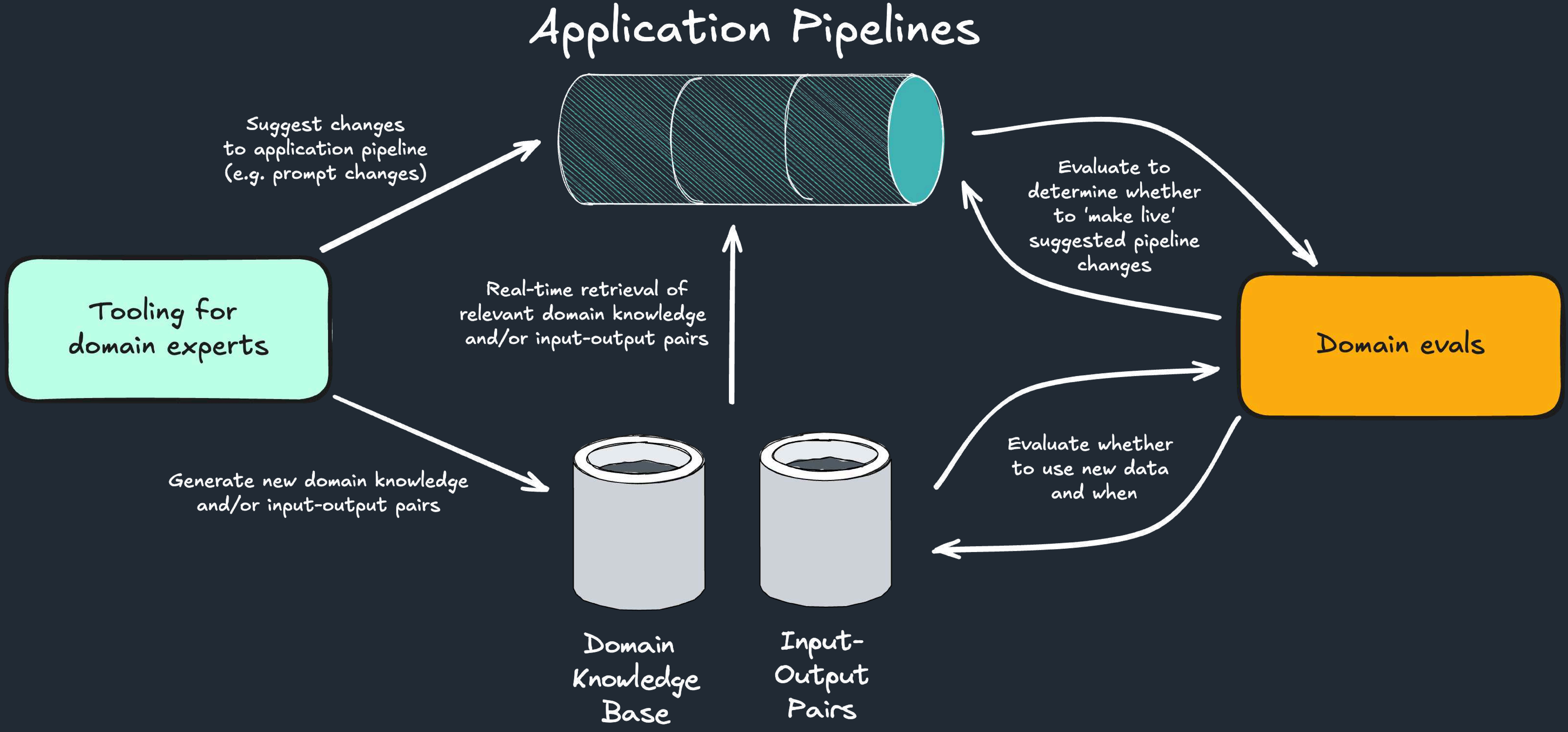
(

input = “Does the patient have a prior brain MRI suspicious for Multiple Sclerosis?\n\nRelevant evidence: {relevant_evidence}”

output = “While there is a brain MRI from 17/5/2022 described as ‘consistent with multiple sclerosis’, this finding is documented in the context of established Multiple Sclerosis diagnosis. The medical record confirms a definitive MS diagnosis, precluding classification of these imaging findings as ‘suspicious for’ MS. Rather, these MRI findings represent characteristic radiological features consistent with the patient's known MS diagnosis.”

)

Automate improvements using evals


Application Pipelines
Suggest changes to application pipeline (e.g. prompt changes)
Evaluate to determine whether to 'make live' suggested pipeline changes
Tooling for domain experts
Real-time retrieval of relevant domain knowledge and/or input-output pairs
Domain evals
Generate new domain knowledge and/or input-output pairs
Evaluate whether to use new data and when
Domain Knowledge Base
Input-Output Pairs

LESSONS FROM BUILDING VERTICALIZED AI AGENTS

1. How domain experts can supercharge AI development - and why to build them a custom UI
2. Why prompting beats finetuning for verticalized agents - and how to go beyond ‘prompt engineering’
3. How to get and maintain customer trust - through intelligent performance monitoring and a secure LLM-native architecture

+ BONUS LESSON (if time)

To build customer trust in your verticalized agent you need to:

1. Give them confidence in the AI performance
2. Show them you’re handling data securely
3. Protect your application against LLM-specific attack vectors

(1) Achieve AI confidence by monitoring and reporting performance

- Reviews of production outputs to generate performance metrics (ideally through a custom UI)
- Proactively communicate performance to customers
- Define a sampling strategy so that you’re reviewing the highest priority cases
- Set up an internal response protocol for whenever performance deviates below targets / SLAs
- Use LLM-as-judge evaluators to scale your monitoring ability

Case Study: Evals for mission-critical healthcare application

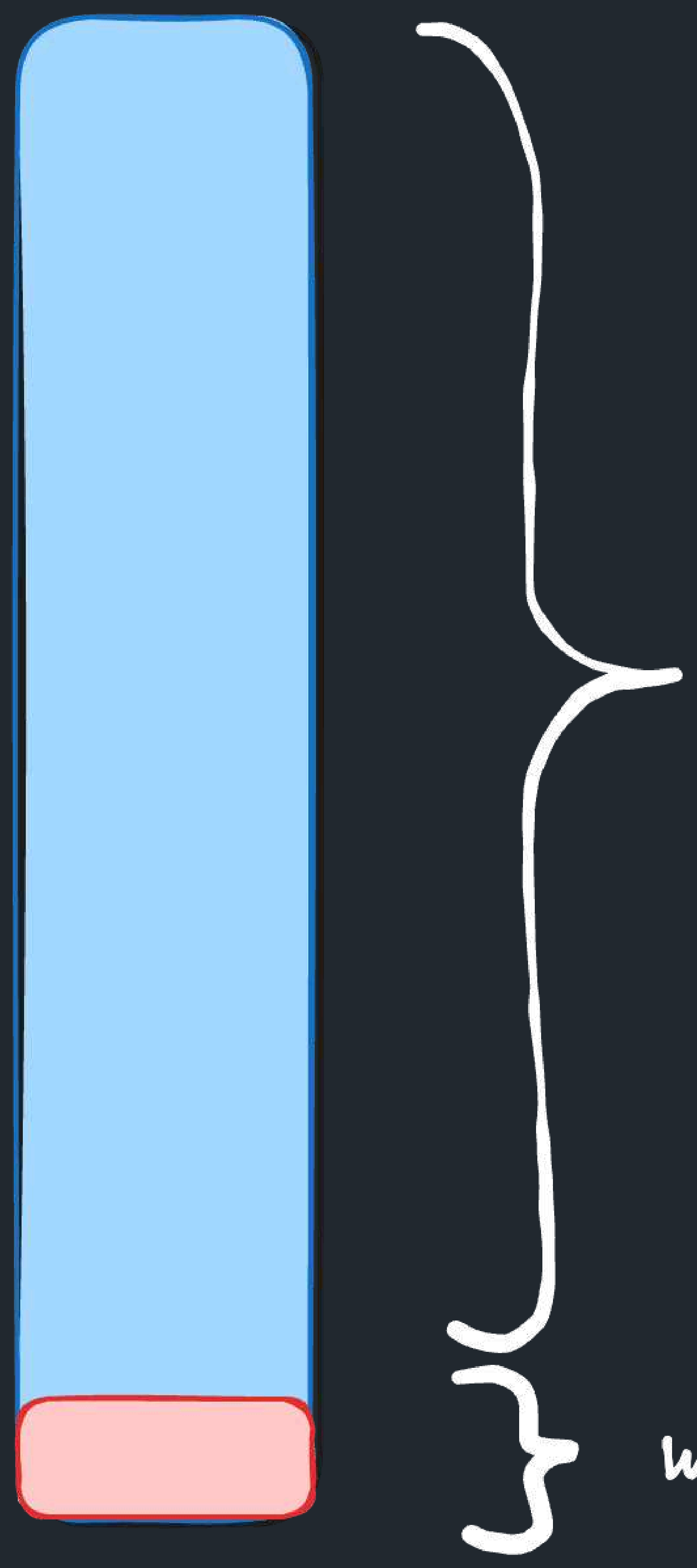
Approach described in detail here:

Mission-Critical Evals at Scale *- Dr Chris Lovejoy, AI Engineer Summit (NY, Feb 2025)*
https://www.youtube.com/watch?v=cZ5ZJy19KMo

Q: Does the patient have a prior brain MRI suspicious for Multiple Sclerosis?

AI Agent:

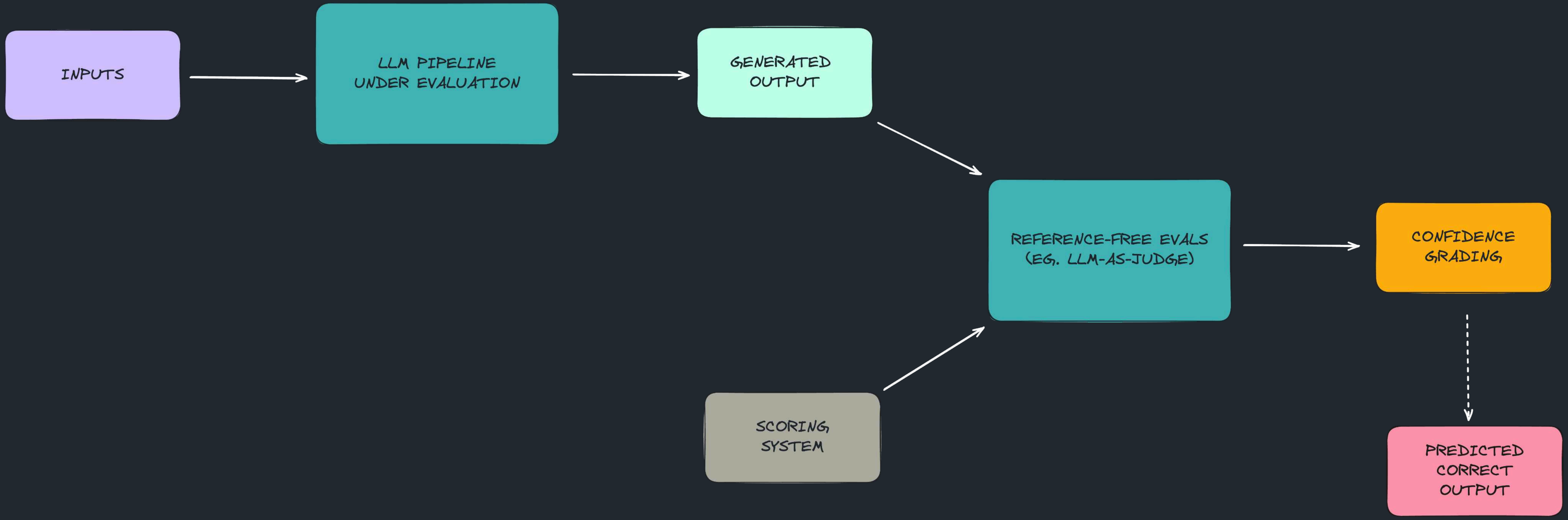
Yes. The medical record shows a brain MRI from 17/5/2022 that demonstrates 'multiple foci of T2/FLAIR hyperintensity in the infratentorial, juxtacortical and periventricular white matter' which is noted to be 'consistent with multiple sclerosis.' This confirms prior brain MRI findings suspicious for MS.



HOW DID WE DO?

WHICH SHOULD WE REVIEW?

An LLM-as-judge can generate both confidence score and predicted output



INPUTS

LLM PIPELINE UNDER EVALUATION

GENERATED OUTPUT

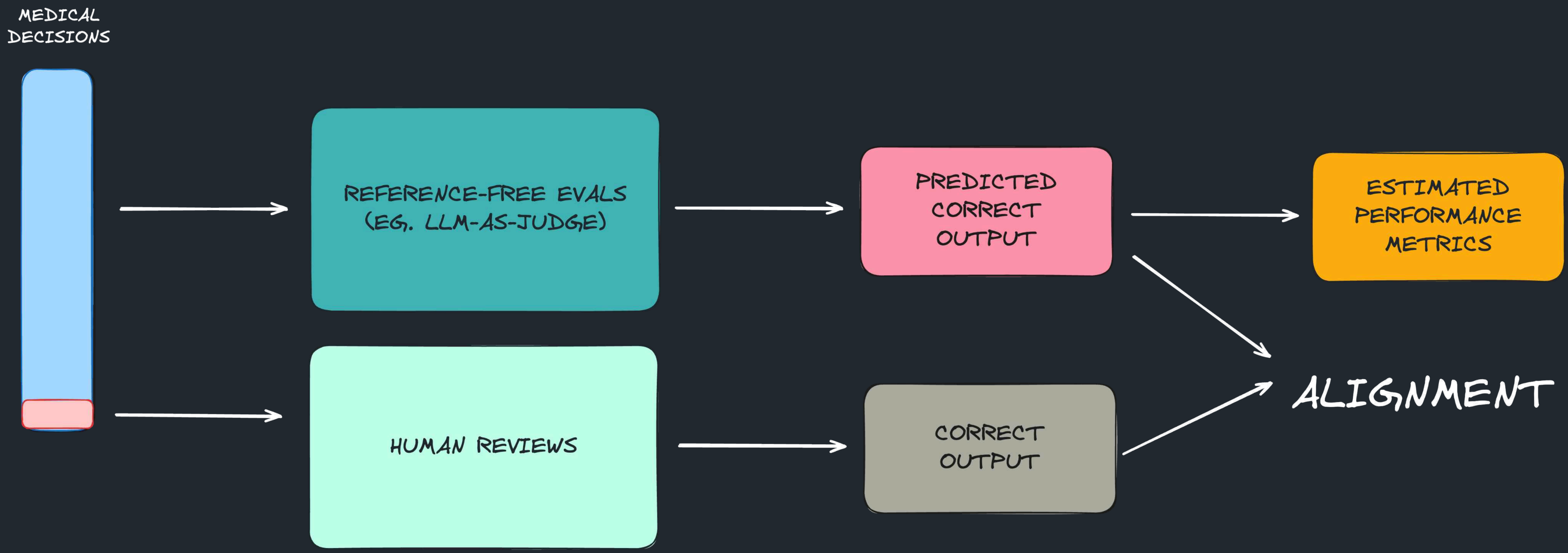
REFERENCE-FREE EVALS (EG. LLM-AS-JUDGE)

SCORING SYSTEM

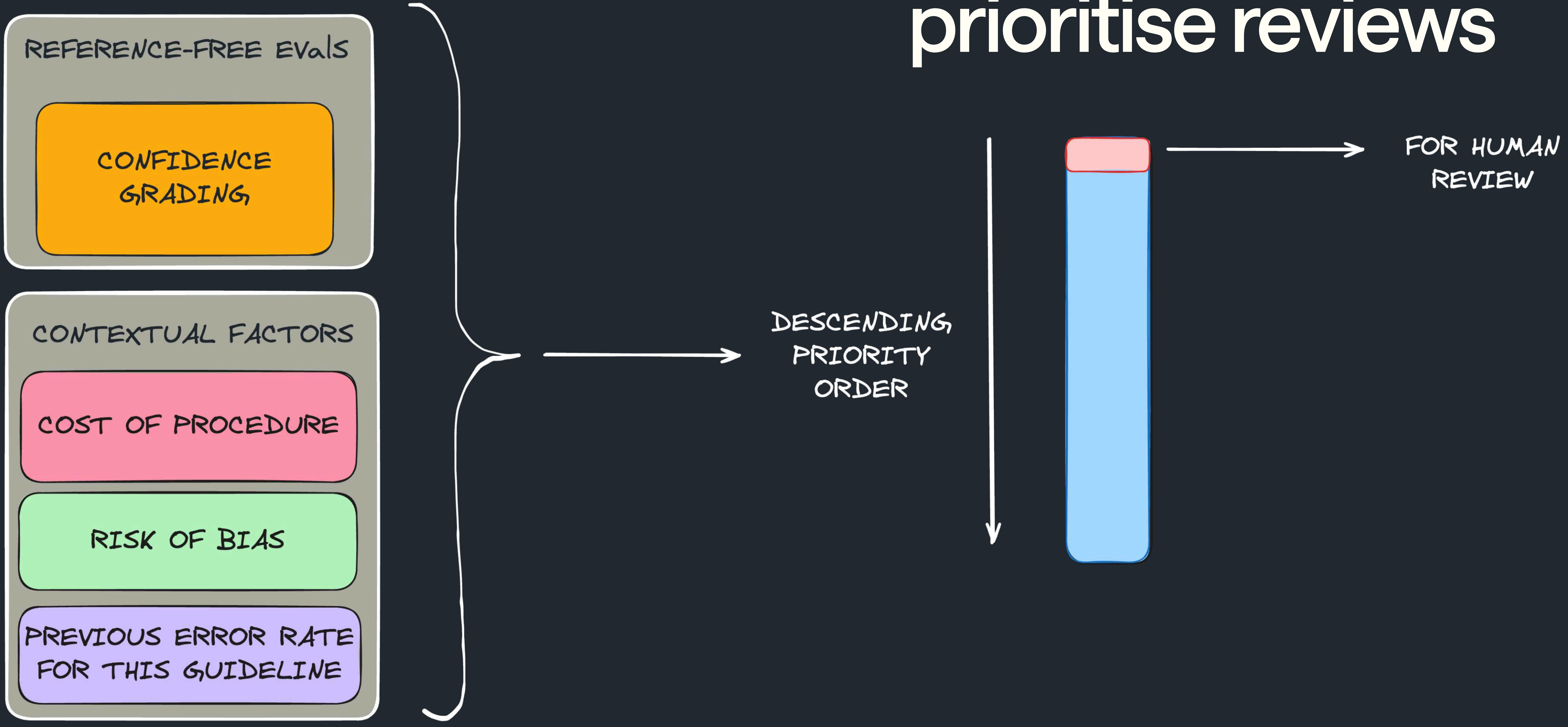
CONFIDENCE GRADING

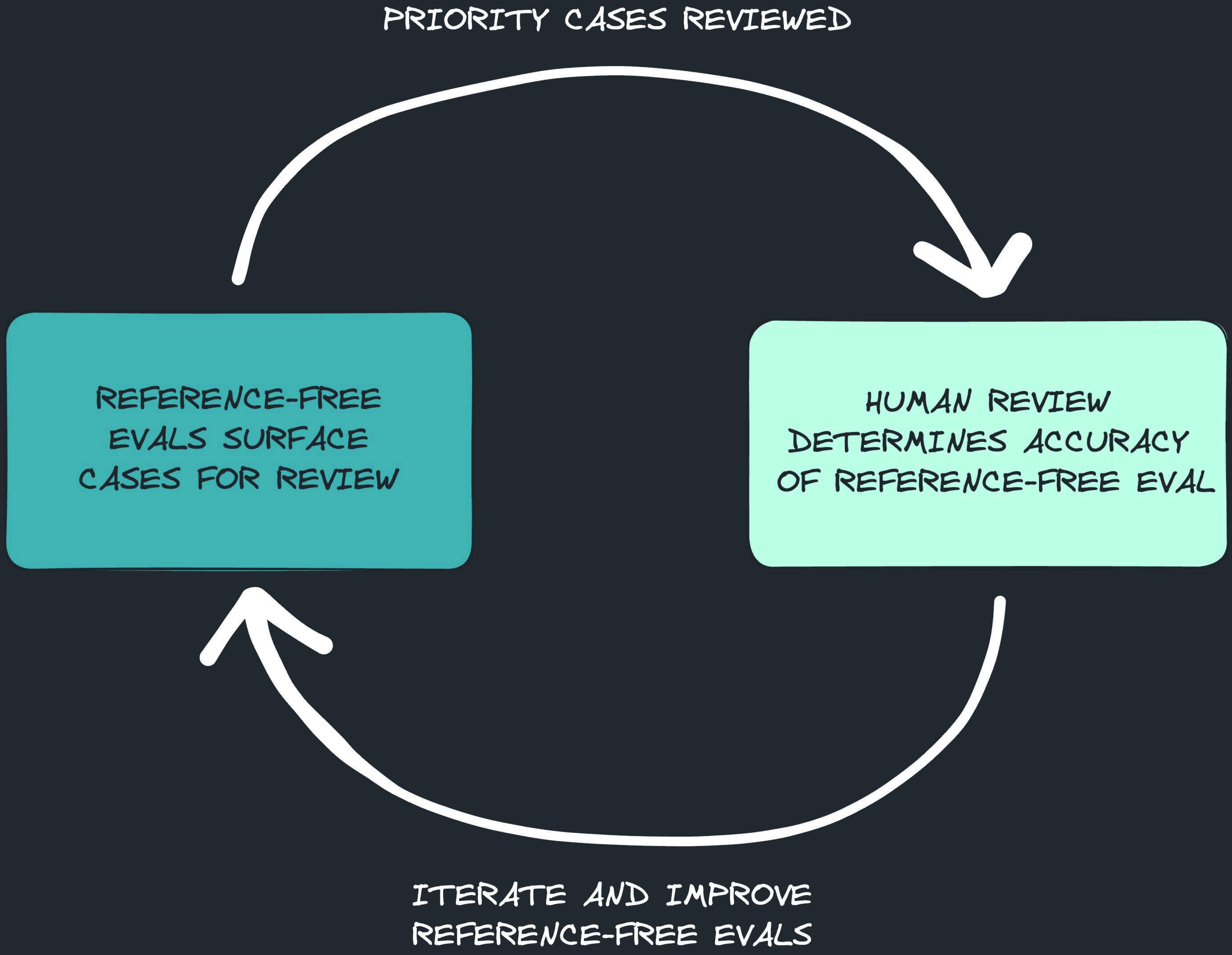
PREDICTED CORRECT OUTPUT

Predicted output gives us estimated performance metrics… and alignment scores



Confidence grading can be used to prioritise reviews





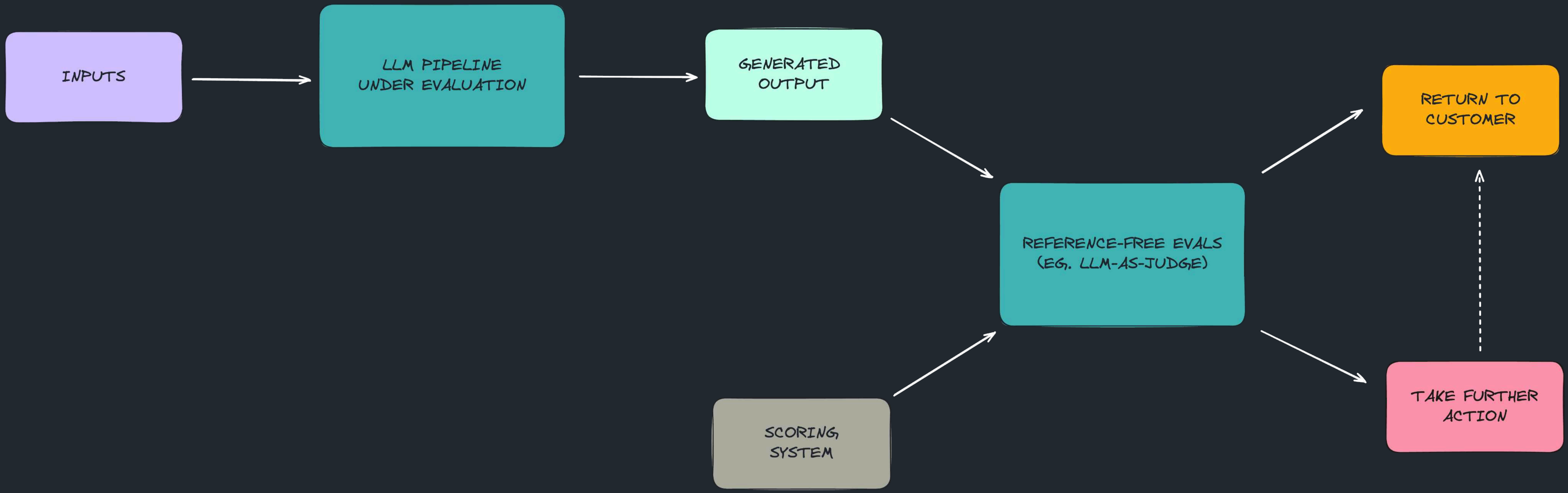
PRIORITY CASES REVIEWED

REFERENCE-FREE EVALS SURFACE CASES FOR REVIEW

HUMAN REVIEW DETERMINES ACCURACY OF REFERENCE-FREE EVAL

ITERATE AND IMPROVE REFERENCE-FREE EVALS

Once the reference-free eval system works, it can become part of the pipeline



Approach described in detail here:

Mission-Critical Evals at Scale *- Dr Chris Lovejoy, AI Engineer Summit (NY, Feb 2025)*
https://www.youtube.com/watch?v=cZ5ZJy19KMo

To build customer trust in your verticalized agent you need to:

1. Give them confidence in the AI performance
2. Handle data securely
3. Protect your application against LLM-specific attack vectors

Lay the foundations early for secure data handling

- Map out data usage strategy early (e.g., for performance monitoring, for fine-tuning)
- Be ready to offer customers isolated single-tenant environments (including dedicated model instances)
- Consider investing in synthetic data generation to reduce dependence on customer data

Keep on top of LLM-specific security considerations and how to mitigate them (because best practices are continually evolving)

- Prompt injections
- Sensitive information disclosure
- Data and model poisoning

TOP 10 FOR GEN AI

2025 Top 10 Risk & Mitigations for LLMs and Gen AI Apps

Expore the latest Top 10 risks, vulnerabilities and mitigations for developing and securing generative AI and large language model applications across the development, deployment and management lifecycle.

OWASP Top 10 for LLM Applications 2025

Version 2025
November 18, 2024

More detail here:

2025 Top 10 Risks and Mitigations for LLM Applications & Generative AI *- OWASP*
https://genai.owasp.org/llm-top-10/

LESSONS FROM BUILDING VERTICALIZED AI AGENTS

1. How domain experts can supercharge AI development - and why to build them a custom UI
2. Why prompting beats finetuning for verticalized agents - and how to go beyond ‘prompt engineering’
3. How to get and maintain customer trust - through intelligent performance monitoring and a secure LLM-native architecture

BONUS LESSON: Which domain experts to hire and how to use them

---

Described in more detail here: ***Why you need a principal domain expert for building vertical AI - and how to find them*** *- Dr Chris Lovejoy*
https://chrislovejoy.me/domain-expert-vertical-ai

Consider hiring a principal domain expert

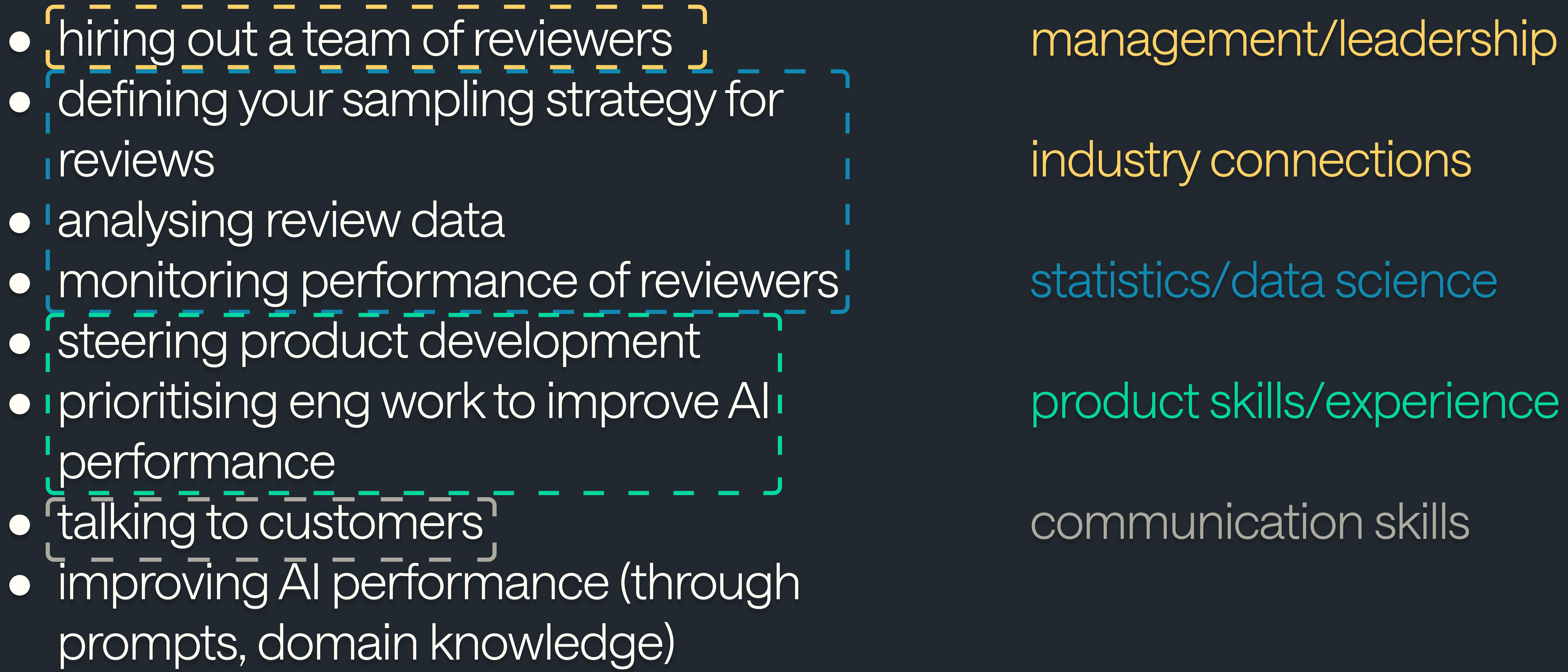
- Having a **directly-responsible individual (DRI)** helps you move faster
- Hire them as early as possible and given them ownership
- They can build the best intuition for how your AI system performs
- Your expert should do more than just look at data - they should help design and create your system

Hire more than “just” a domain expert

Your principal domain expert can also help with:

- hiring out a team of reviewers
- defining your sampling strategy for reviews
- analysing review data
- monitoring performance of reviewers
- steering product development
- prioritising eng work to improve AI performance
- talking to customers
- improving AI performance (through prompts, domain knowledge)

Hire more than “just” a domain expert



Your principal domain expert can also help with:

- hiring out a team of reviewers
- defining your sampling strategy for reviews
- analysing review data
- monitoring performance of reviewers
- steering product development
- prioritising eng work to improve AI performance
- talking to customers
- improving AI performance (through prompts, domain knowledge)

So it can be helpful if they have the following skills and experiences:

management/leadership

industry connections

statistics/data science

product skills/experience

communication skills

Thank you

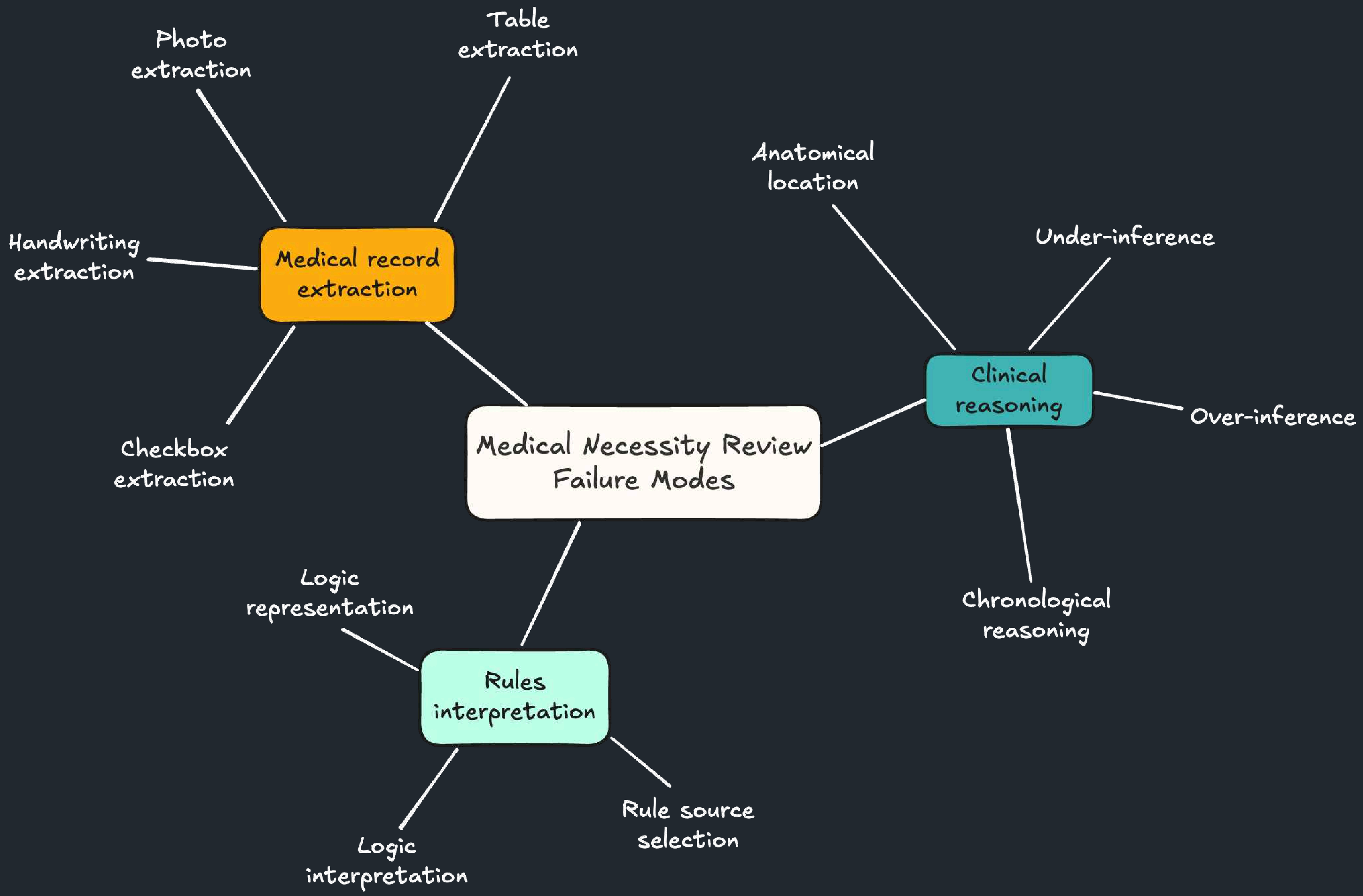
Dr Christopher Lovejoy, MD

www.chrislovejoy.me/email

hi@chrislovejoy.me

Appendix

Empower domain experts to define and maintain a failure mode taxonomy



What defines a “specialized vertical”?

Could a lay person develop “top tier” judgment
after working on this for 2 months?