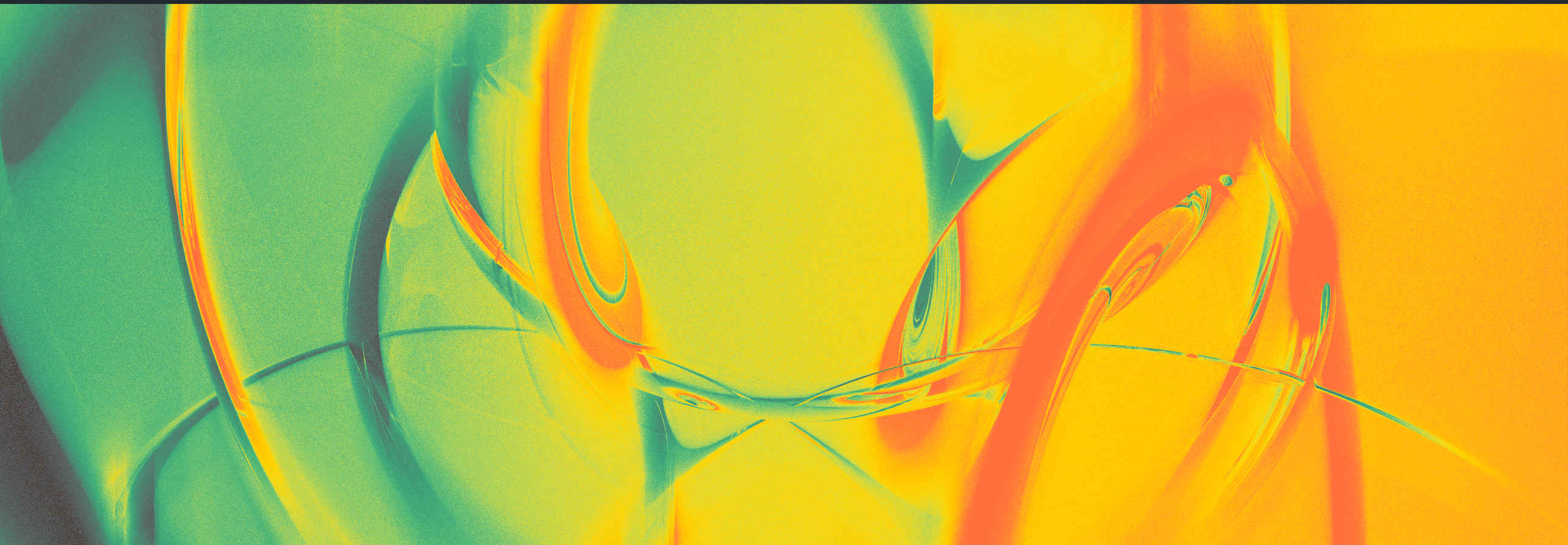
# AI Evals in Vertical Industries

Guest lecture for “AI Evals For Engineers & PMs” course

Dr Christopher Lovejoy, MD
Head of Clinical AI

2025-08-07
Anterior



# OUTLINE

- Why is it harder to do evals for specialized verticals?
- Empowering domain experts to look at your data
- Principles for building an effective domain expert review dashboard
- Using review data to power your system
- How to hire the right domain experts

# What defines a “specialized vertical”?

# What defines a “specialized vertical”?

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

# Why is it harder to do evals for specialized verticals?

# Why is it harder to do evals for specialized verticals?

1. It's harder to define quality / correctness

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

Florence:

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.

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

Florence:

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.

# Why is it harder to do evals for specialized verticals?

1. It's harder to define quality / correctness
2. It's harder to define failure modes

# Example failure modes: clinical reasoning

| Failure Mode | Description |
| --- | --- |
| *Retrospective reasoning* | Incorrectly using evidence that became available after the decision point to justify whether the patient should have had the procedure in the first place. |
| *Under-inference* | Not making valid inference, e.g saying something ‘needs to be explicitly stated’ when it could be reasonably inferred |
| *Over-inference* | Drawing conclusions that go beyond what the evidence supports. Making assumptions without sufficient basis. |
| *Misunderstanding chronology* | Incorrect interpretation or application of the sequence of events (reading events in wrong order) |
| ... | ... |

# Why is it harder to do evals for specialized verticals?

1. It's harder to define quality / correctness
2. It's harder to define failure modes
3. It's harder to write prompts
    a. Prompts in your main pipelines
    b. LLM-as-judge prompts to evaluate your pipelines

# Why is it harder to do evals for specialized verticals?

Error analysis

1. It's harder to define quality / correctness
2. It's harder to define failure modes
3. It's harder to write prompts
    a. Prompts in your main pipelines
    b. LLM-as-judge prompts to evaluate your pipelines

# Why is it harder to do evals for specialized verticals?

Error analysis

1. It's harder to define quality / correctness
2. It's harder to define failure modes
3. It's harder to write prompts
    a. Prompts in your main pipelines
    b. LLM-as-judge prompts to evaluate your pipelines

Implementing Automated Evaluators

# Why is it harder to do evals for specialized verticals?

Error analysis

1. It's harder to define quality / correctness
2. It's harder to define failure modes
3. It's harder to write prompts
   a. Prompts in your main pipelines
   b. LLM-as-judge prompts to evaluate your pipelines

Improvement

Implementing Automated Evaluators

# The solution: bring domain experts into the loop

# The solution: bring domain experts into the loop

But where? and how?

“look at your data”

# “look at your data”

*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.*

# “look at your data”

*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.*

# “but what does this mean?”

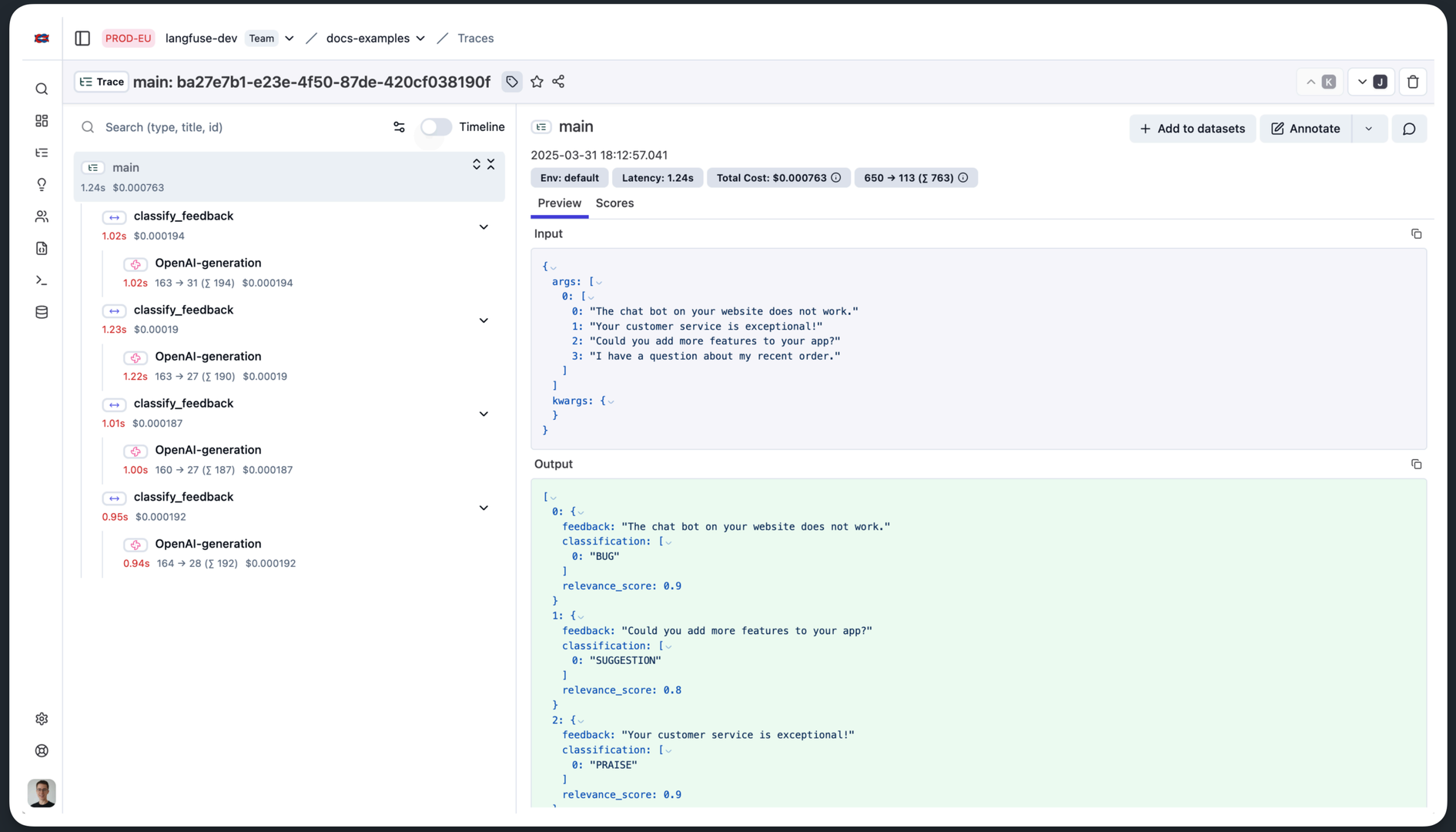
“look at your data” ⟶ “empower domain experts to look at (and translate) your data”

*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.*

“but what does this mean?”

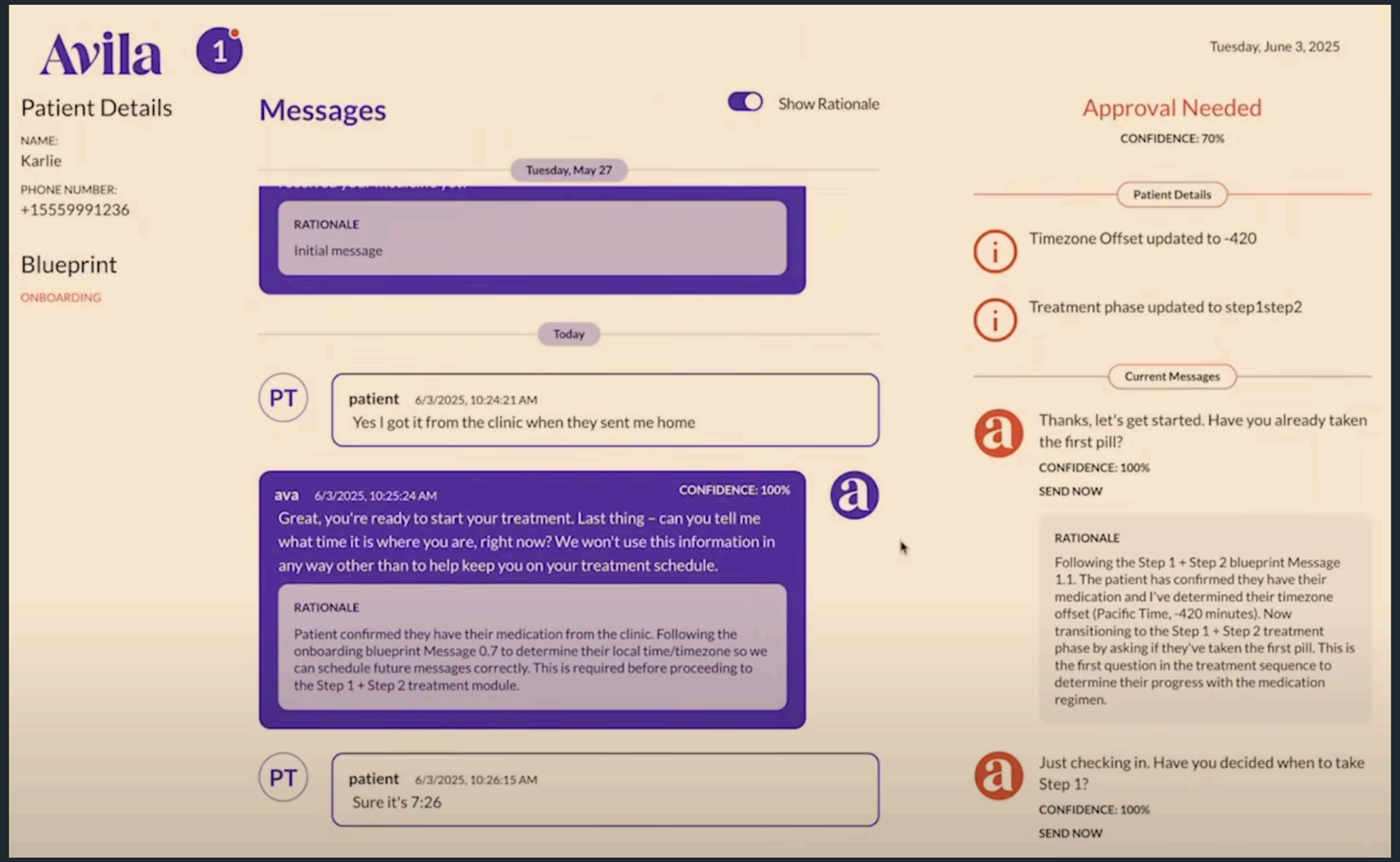
# Empowering domain experts to look at your data

# Example 1: Raw traces



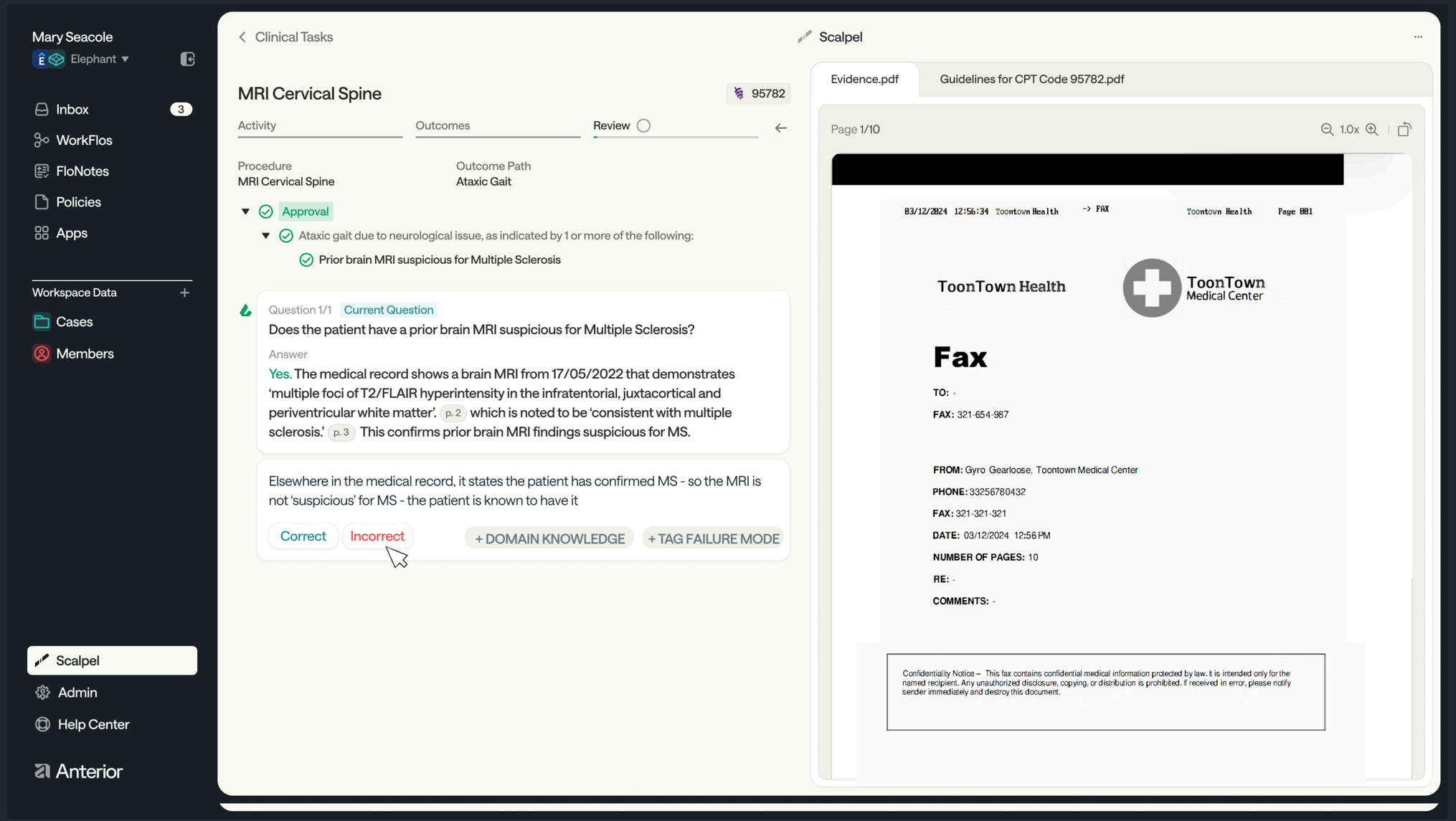

**Source:** *https://langfuse.com/images/docs/instructor-trace.png*

# Example 2: A custom view within the product




**Source:** *https://www.youtube.com/watch?v=sn79oS4MZFI*

# Example 3: A custom review dashboard




**Source:** *https://chrislovejoy.me/llm-native-expert-system*

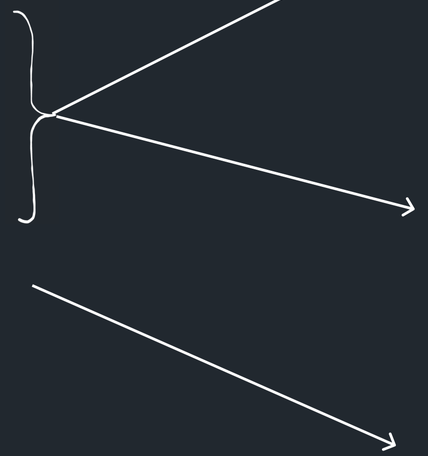
# Principles for building an effective domain expert review dashboard

# Optimise for 3 things:

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

# Optimise for 3 things:

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

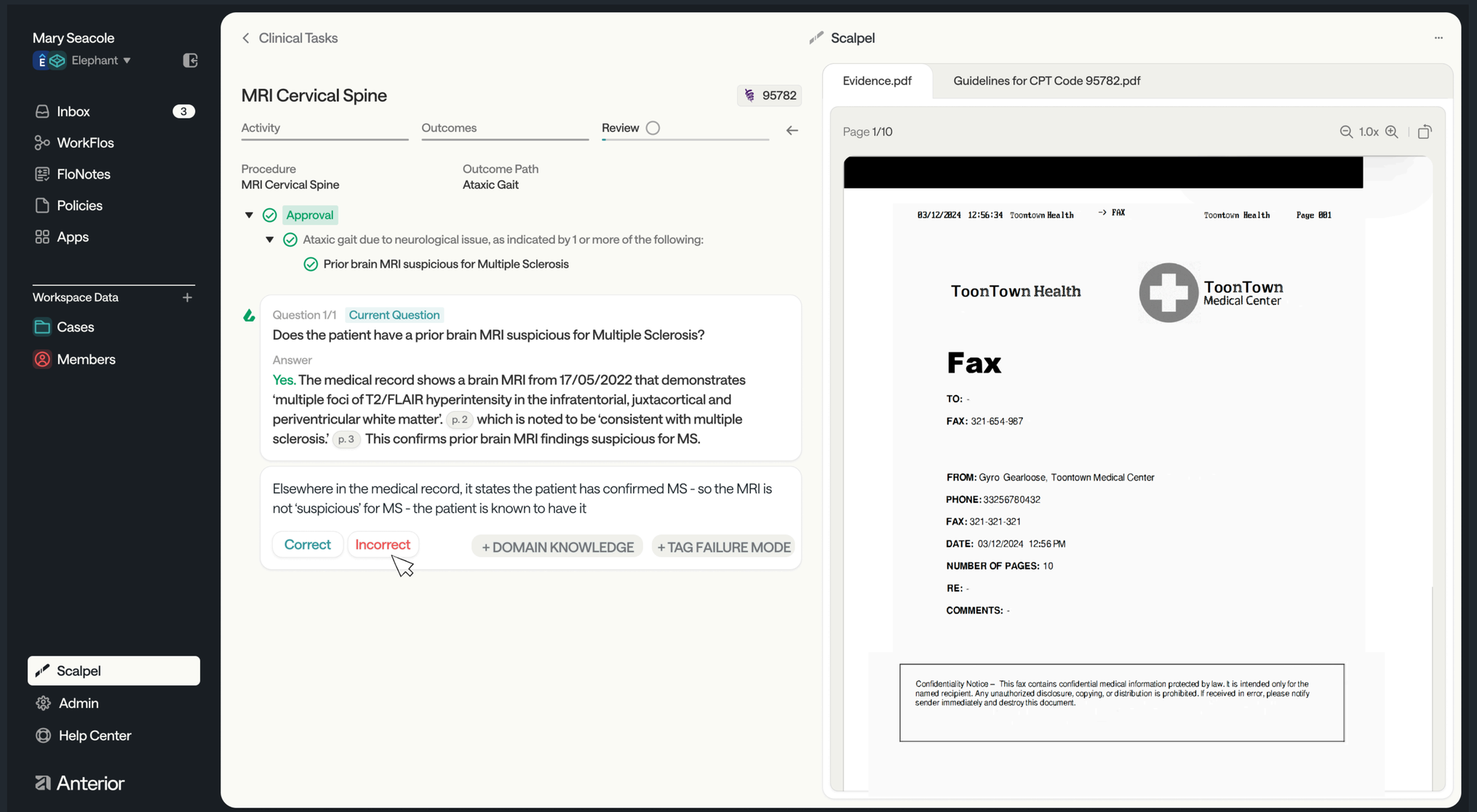


Principle 1: Optimise for clearly surfacing all required context

Principle 2: Optimise the review flow sequence

Principle 3: Design reviews that give the data you need

# Principle 1: Optimise for clearly surfacing all required context


Mary Seacole
Elephant
Inbox
3
WorkFlos
FloNotes
Policies
Apps
Workspace Data
Cases
Members
Scalpel
Admin
Help Center
Anterior
Clinical Tasks
Scalpel
MRI Cervical Spine
95782
Activity
Outcomes
Review
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
Correct
Incorrect
+ DOMAIN KNOWLEDGE
+ TAG FAILURE MODE
Evidence.pdf
Guidelines for CPT Code 95782.pdf
Page 1/10
1.0x
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.

# Principle 2: Optimise the review flow sequence

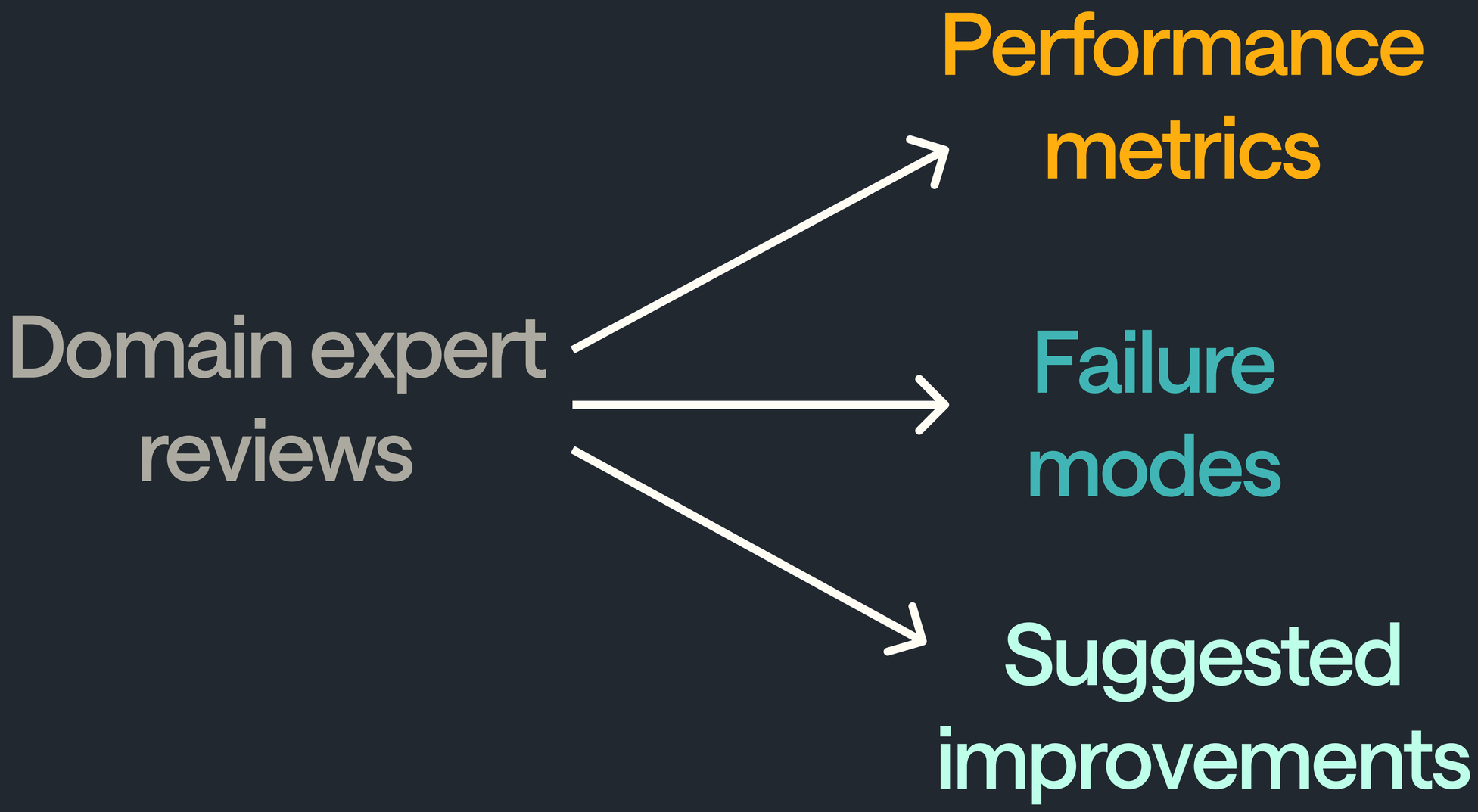
Review case summary

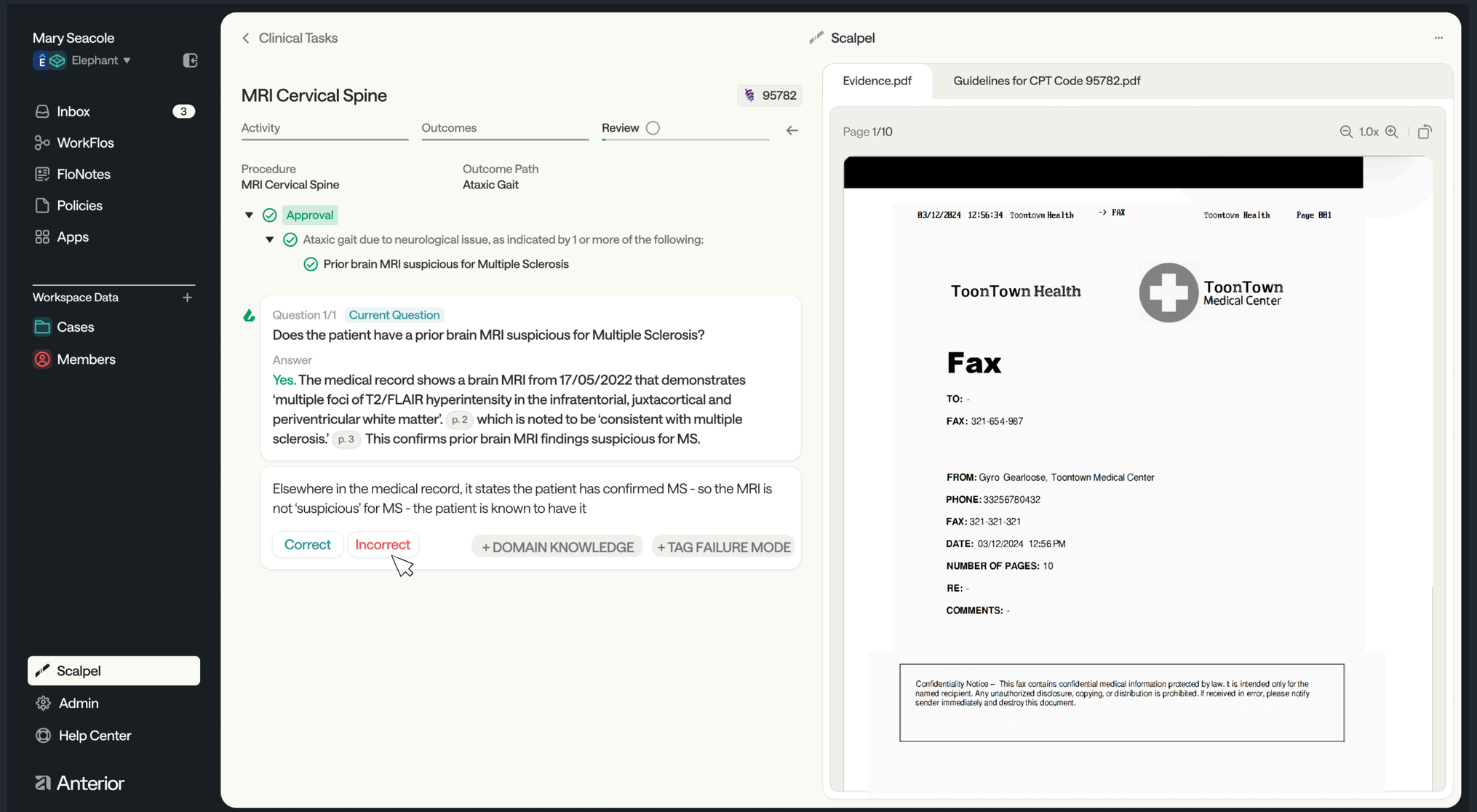
Understand current decision point

Review the relevant medical evidence

Appraise the AI output

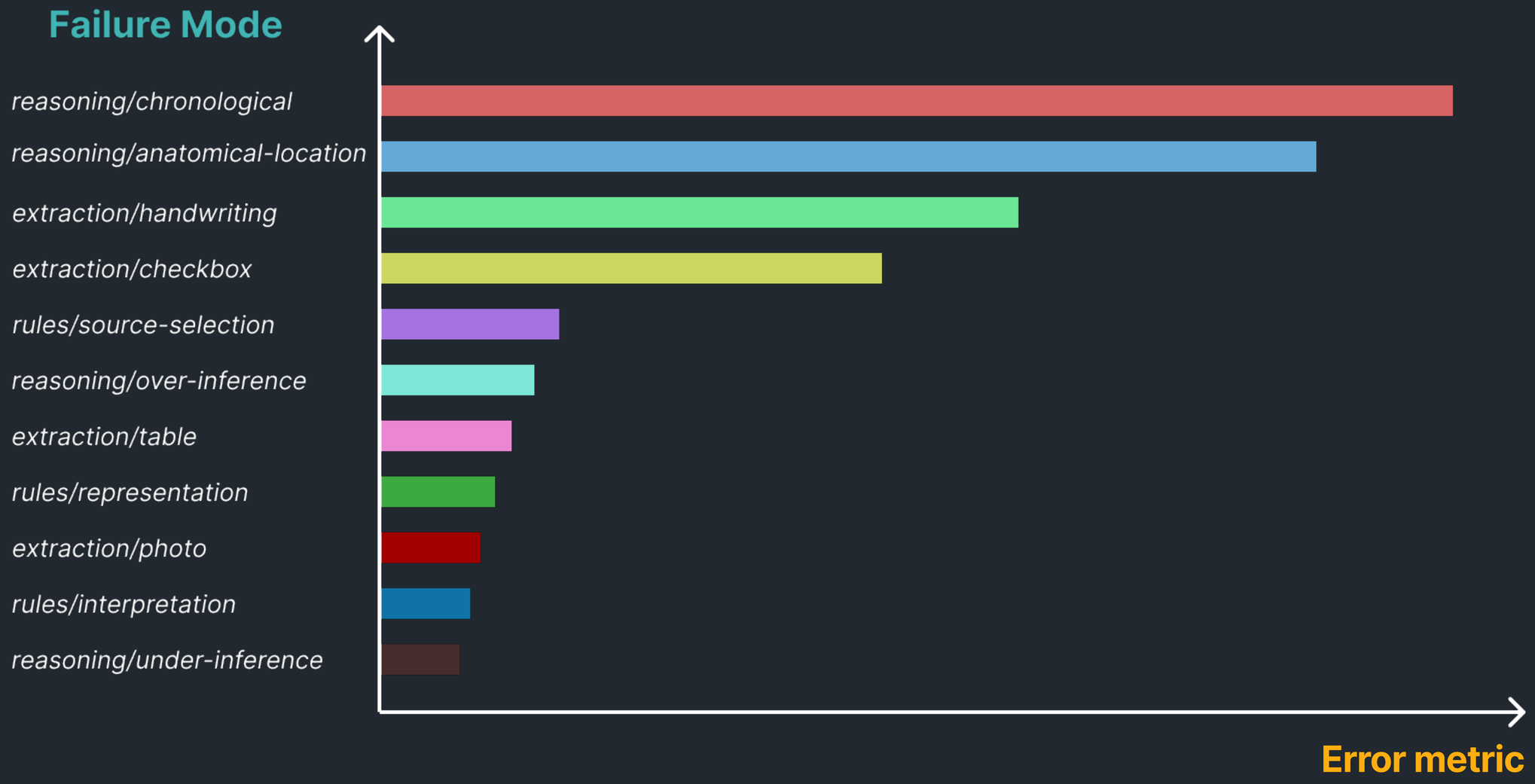
# Principle 3: Design reviews that give the data you need


Domain expert reviews
Performance metrics
Failure modes
Suggested improvements

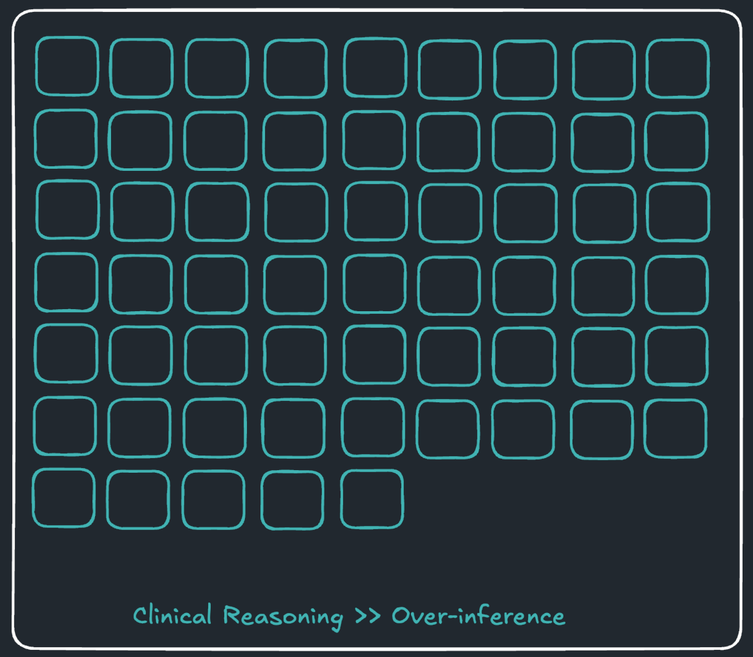

Mary Seacole
Elephant
Inbox
3
WorkFlos
FloNotes
Policies
Apps
Workspace Data
Cases
Members
Scalpel
Admin
Help Center
Anterior
Clinical Tasks
Scalpel
MRI Cervical Spine
95782
Activity
Outcomes
Review
Procedure
MRI Cervical Spine
Outcome Path
Ataxic Gait
Approval
Ataxic gait due to neurological issue, 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
Correct
Incorrect
+ DOMAIN KNOWLEDGE
+ TAG FAILURE MODE
Evidence.pdf
Guidelines for CPT Code 95782.pdf
Page 1/10
1.0x
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.

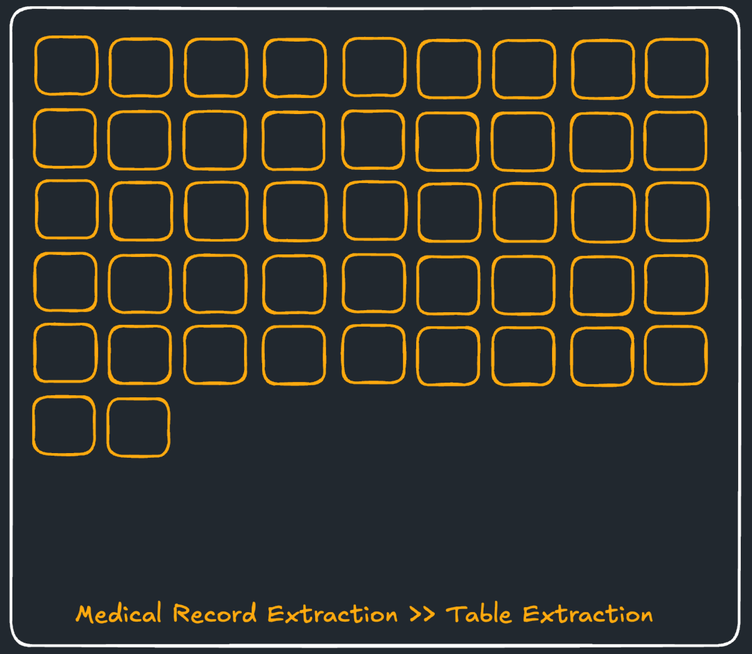
# Using review data to power your system

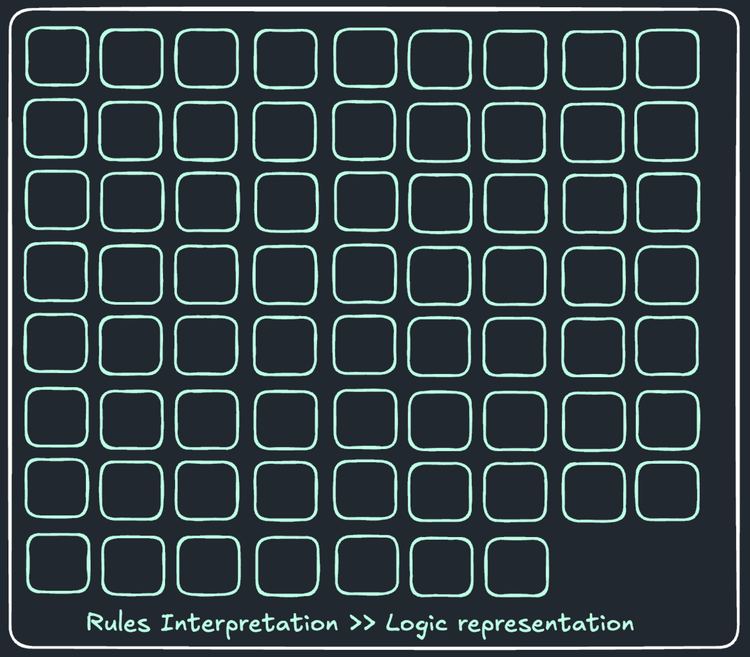
# (1) Use metrics and failure modes from production data to prioritise work



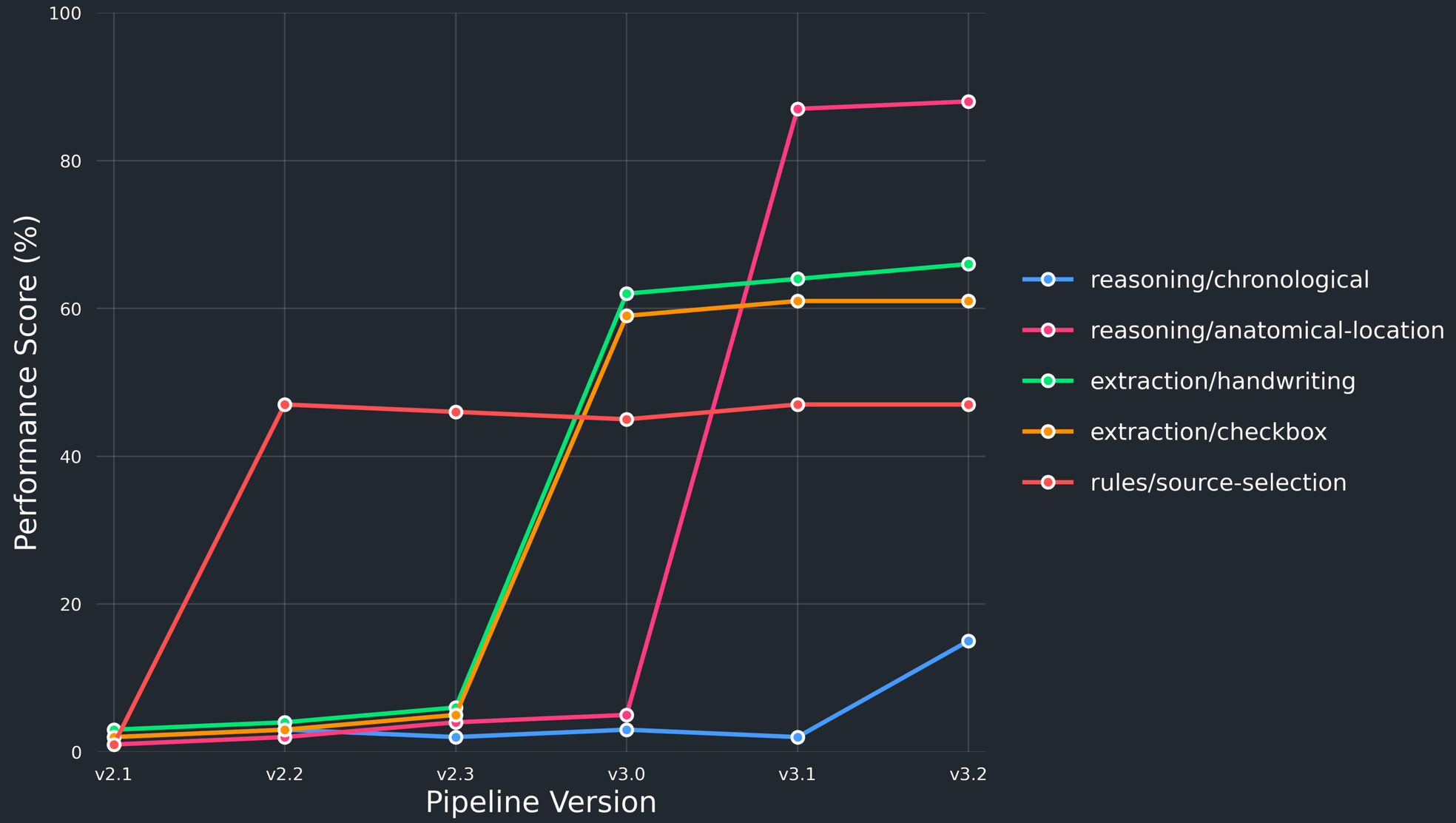
# (2) Use failure mode datasets to test the impact of changes



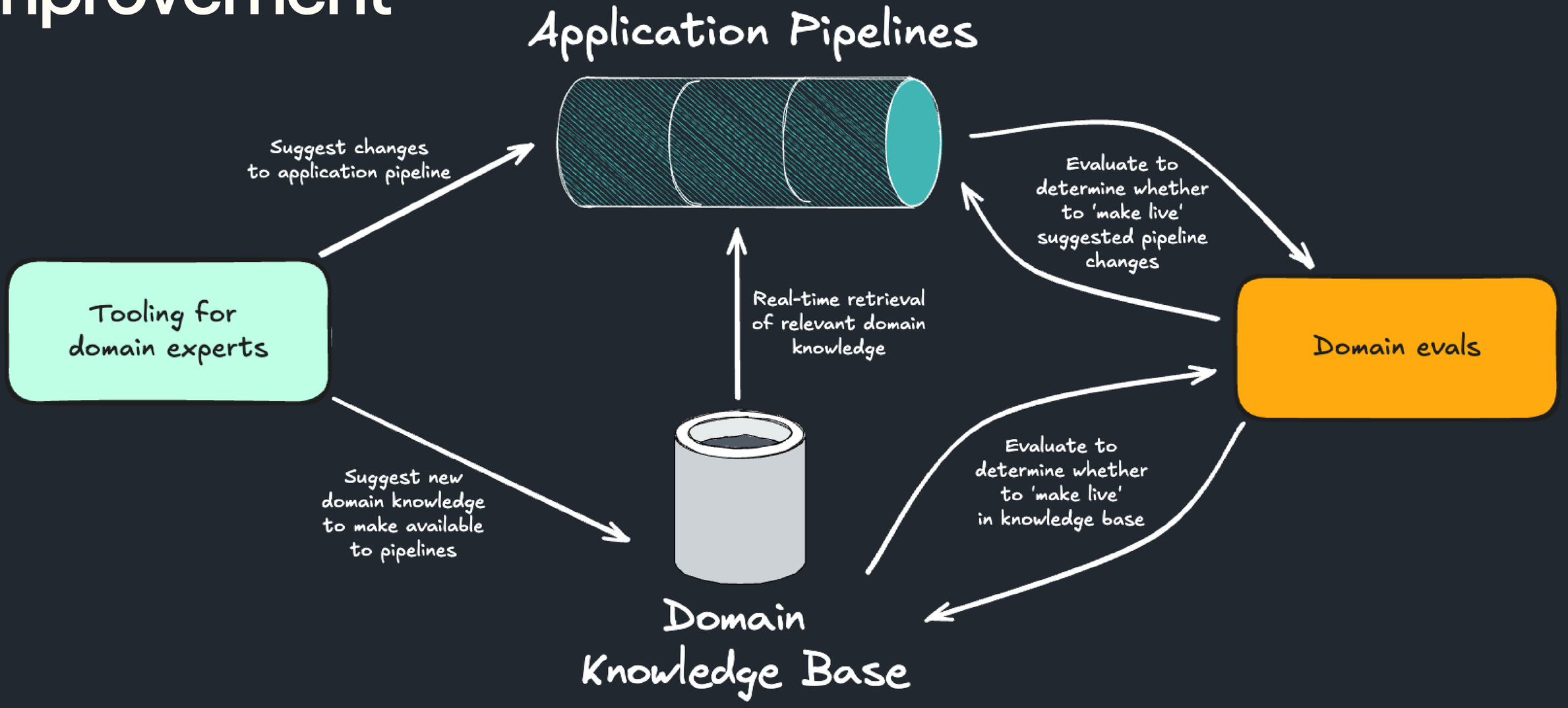



# (2) Use failure mode datasets to test the impact of changes



# (3) Create mechanisms for automated improvement



# Hiring domain experts

# It helps to have a **principal domain expert**

- Having a **directly-responsible individual (DRI)** helps you move faster
- They can build the best intuition for how your AI system performs
- Hire them as early as possible and given them ownership
- 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

# In Summary

- Most of the general eval principles apply to specialized verticals. There are added challenges with (i) defining quality and/or correctness, (ii) defining failure modes and (iii) writing prompts
- Domain experts perform a critical ‘translation’ step by looking at your data and converting it into actionable insights
- You can empower them through custom review dashboards which optimise for quality and speed of reviews and generate helpful review data
- That review data can prioritise work, facilitate AI iteration and even make automated improvements
- Hire a principal domain expert early, ideally with a breadth of skills beyond their domain expertise so they can drive the development of this system

# Thank you


*Dr Christopher Lovejoy, MD*

*Head of Clinical AI, Anterior*

*hi@chrislovejoy.me*

*www.chrislovejoy.me*