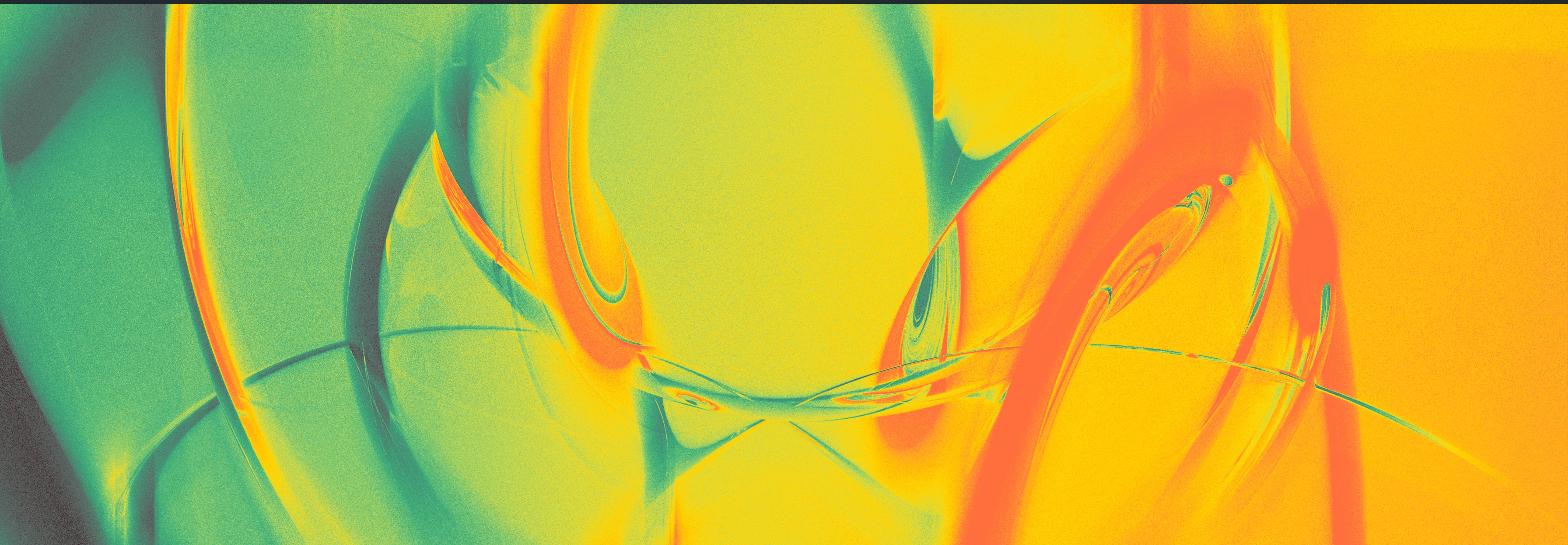


AI Eval in Vertical Industries

Guest lecture for “AI Eval 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?



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?



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
<i>Retrospective reasoning</i>	Incorrectly using evidence that became available after the decision point to justify whether the patient should have had the procedure in the first place.
<i>Under-inference</i>	Not making valid inference, e.g saying something ‘needs to be explicitly stated’ when it could be reasonably inferred
<i>Over-inference</i>	Drawing conclusions that go beyond what the evidence supports. Making assumptions without sufficient basis.
<i>Misunderstanding chronology</i>	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

The screenshot shows a trace visualization interface for a classification task. The main panel displays the trace details for a specific trace ID: `main:ba27e7b1-e23e-4f50-87de-420cf038190f`. The top right shows the trace was recorded on `2025-03-31 18:12:57.041` with `Latency: 1.24s`, `Total Cost: $0.000763`, and `650 → 113 (Σ 763)`.

The left sidebar lists the trace structure:

- `main`: `1.24s $0.000763`
- `classify_feedback`:
 - `OpenAI-generation`: `1.02s 163 → 31 (Σ 194) $0.000194`
 - `classify_feedback`:
 - `OpenAI-generation`: `1.22s 163 → 27 (Σ 190) $0.00019`
 - `classify_feedback`:
 - `OpenAI-generation`: `1.01s 160 → 27 (Σ 187) $0.000187`
 - `classify_feedback`:
 - `OpenAI-generation`: `0.95s $0.000192`
 - `classify_feedback`:
 - `OpenAI-generation`: `0.94s 164 → 28 (Σ 192) $0.000192`

The right panel shows the `Input` and `Output` data.

Input:

```
{  
  args: [  
    0: [  
      0: "The chat bot on your website does not work."  
      1: "Your customer service is exceptional!"  
      2: "Could you add more features to your app?"  
      3: "I have a question about my recent order."  
    ]  
  ]  
  kwargs: {}  
}
```

Output:

```
[  
  0: {  
    feedback: "The chat bot on your website does not work."  
    classification: [  
      0: "BUG"  
    ]  
    relevance_score: 0.9  
  }  
  1: {  
    feedback: "Could you add more features to your app?"  
    classification: [  
      0: "SUGGESTION"  
    ]  
    relevance_score: 0.8  
  }  
  2: {  
    feedback: "Your customer service is exceptional!"  
    classification: [  
      0: "PRAISE"  
    ]  
    relevance_score: 0.9  
  },
```

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

Example 2: A custom view within the product

The screenshot shows the Avila mobile application interface. At the top left is the Avila logo with a notification badge '1'. To the right is the date 'Tuesday, June 3, 2025'. On the far left, under 'Patient Details', it shows 'NAME: Karlie' and 'PHONE NUMBER: +15559991236'. Below this is the 'Blueprint' section with 'ONBOARDING' status. In the center, the 'Messages' screen displays a conversation. A purple message from 'ava' at 6/3/2025, 10:25:24 AM with 'CONFIDENCE: 100%' contains the text: 'Great, you're ready to start your treatment. Last thing - can you tell me what time it is where you are, right now? We won't use this information in any way other than to help keep you on your treatment schedule.' Below this is a 'RATIONALE' section: 'Patient confirmed they have their medication from the clinic. Following the onboarding blueprint Message 0.7 to determine their local time/timezone so we can schedule future messages correctly. This is required before proceeding to the Step 1 + Step 2 treatment module.' A blue message from 'patient' at 6/3/2025, 10:24:21 AM with 'CONFIDENCE: 100%' contains the text: 'Yes I got it from the clinic when they sent me home'. A purple message from 'ava' at 6/3/2025, 10:26:15 AM with 'CONFIDENCE: 100%' contains the text: 'Sure it's 7:26'. To the right, there is a 'Show Rationale' toggle switch. The 'Approval Needed' section shows a confidence level of 70% with a red 'CONFIDENCE: 70%' indicator. It lists two items: 'Timezone Offset updated to -420' and 'Treatment phase updated to step1step2'. Below this is a 'Current Messages' section with a message from 'ava' at 6/3/2025, 10:25:24 AM with 'CONFIDENCE: 100%' containing the text: 'Thanks, let's get started. Have you already taken the first pill?' followed by a 'SEND NOW' button. A large blue 'a' icon is positioned next to the message. At the bottom right, there is a 'RATIONALE' section for the previous message from 'ava' with a detailed explanation of the treatment sequence and medication regimen.

Example 3: A custom review dashboard

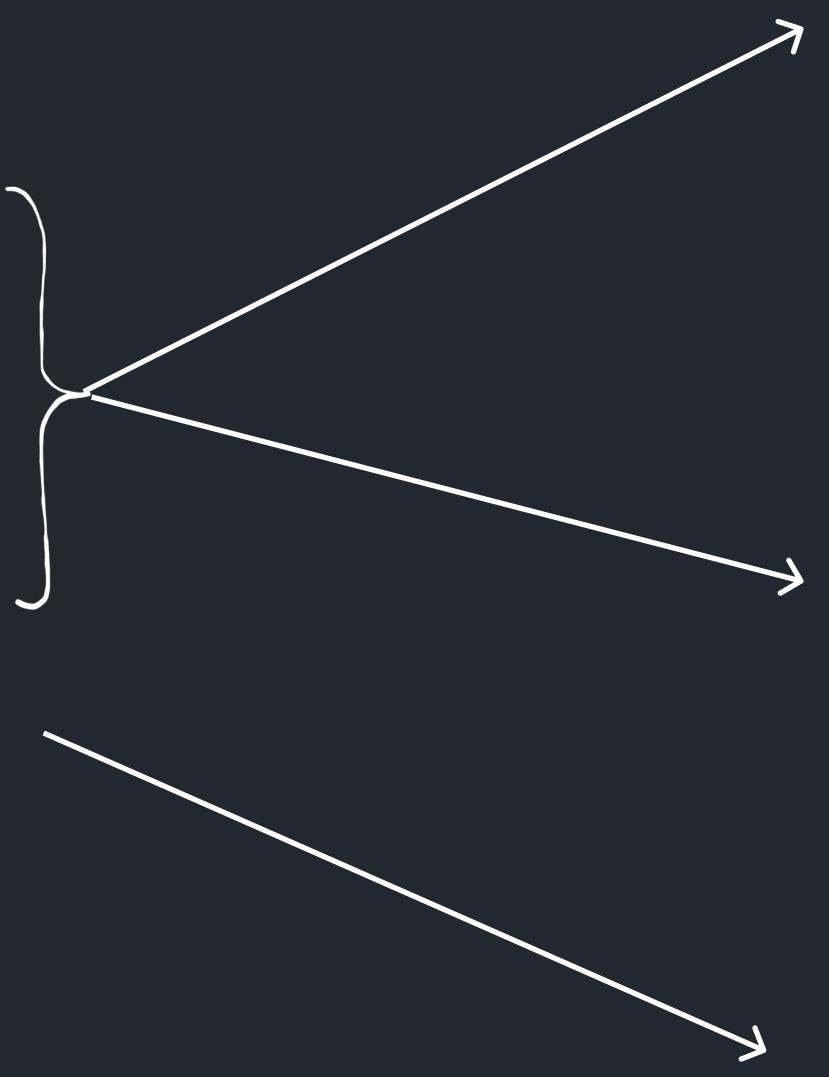
Principles for building an effective domain expert review dashboard

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

MRI Cervical Spine

Activity Outcomes Review (0) ←

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

Scalpel

Evidence.pdf Guidelines for CPT Code 95782.pdf

Page 1/10 1.0x ↗

83/12/2024 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

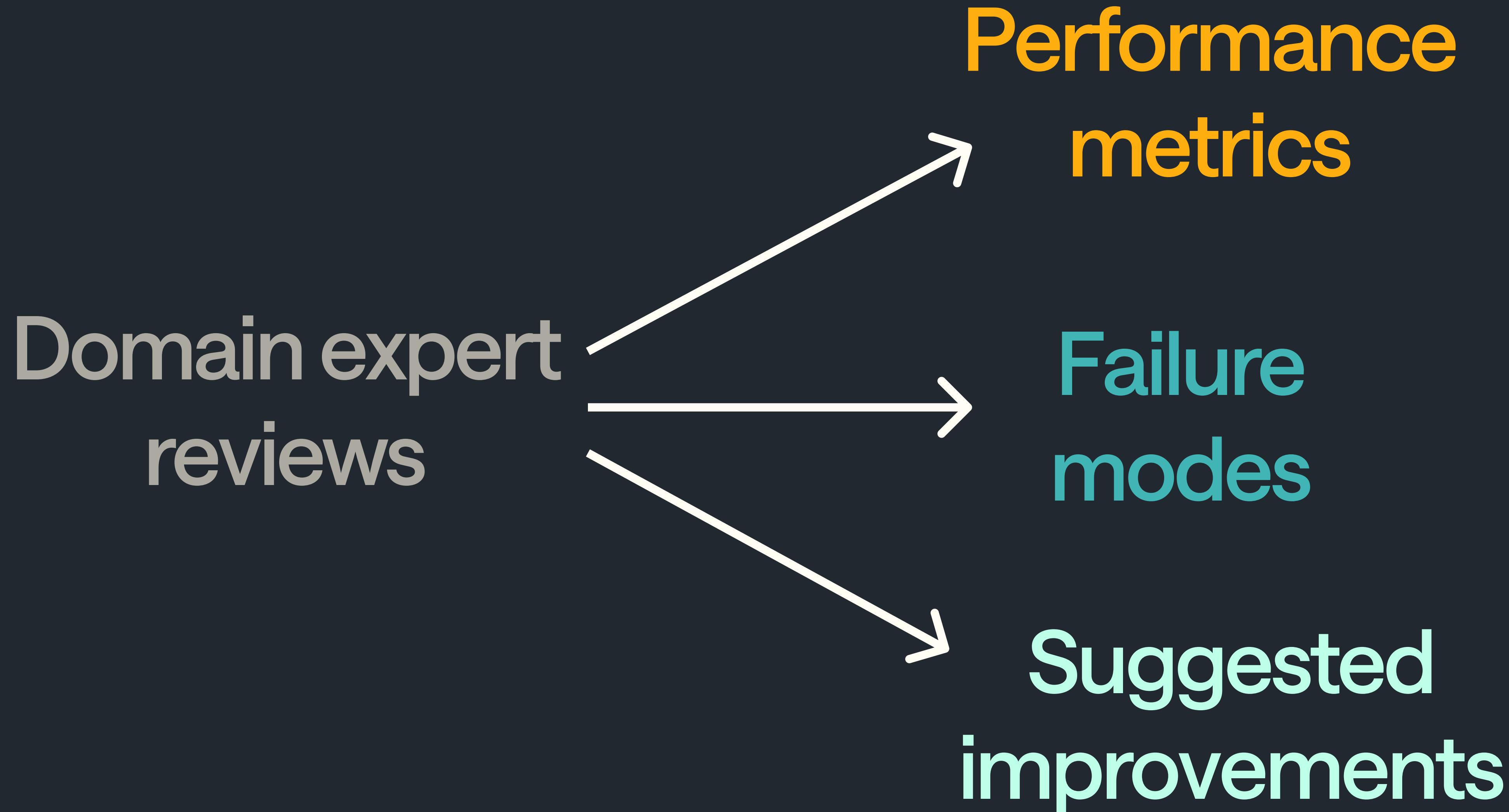
Understand current
decision point

Review the relevant
medical evidence

Appraise the AI
output



Principle 3: Design reviews that
give the data you need



Mary Seacole

Elephant

Inbox (3)

WorkFlos

FloNotes

Policies

Apps

Workspace Data +

Cases

Members

Scalpel

Admin

Help Center

Anterior

< Clinical Tasks

MRI Cervical Spine

Activity Outcomes Review (0) ←

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

Scalpel

Evidence.pdf Guidelines for CPT Code 95782.pdf

Page 1/10 1.0x ↗

83/12/2024 12:56:34 Toontown Health → FAX Toontown Health Page 001

ToonTown Health ToonTown Medical Center

Fax

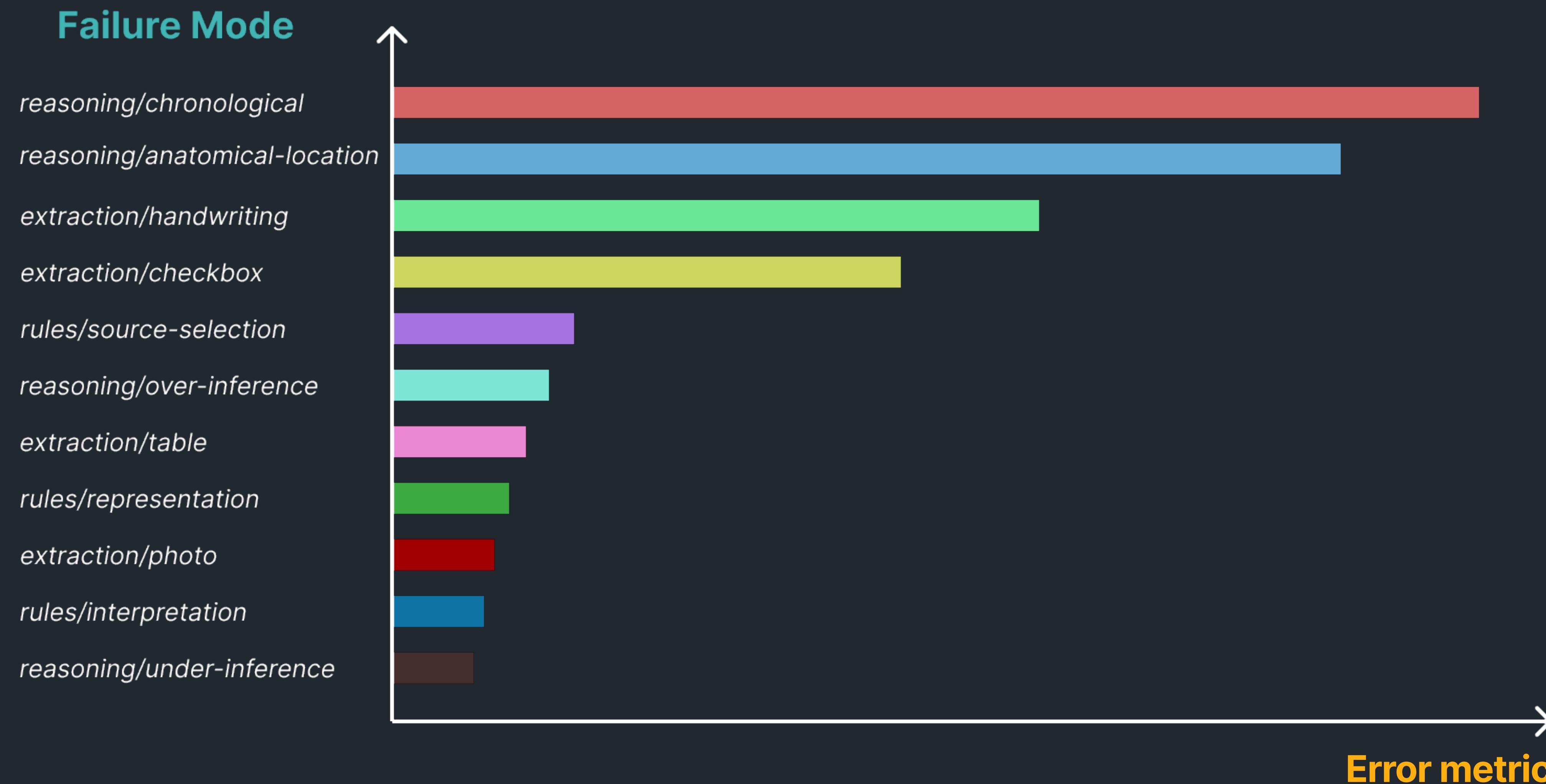
TO: -
FAX: 321-654-987

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: -

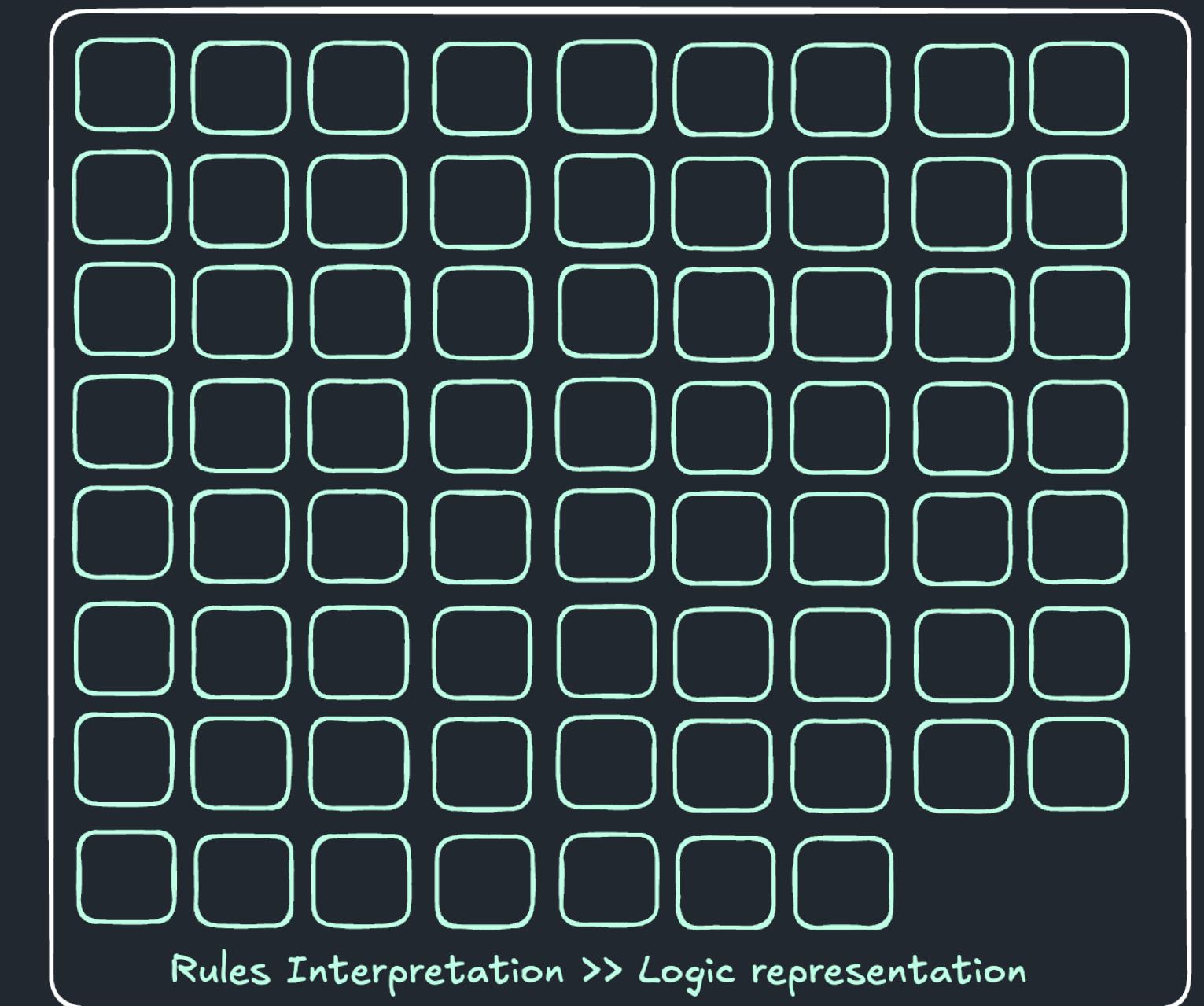
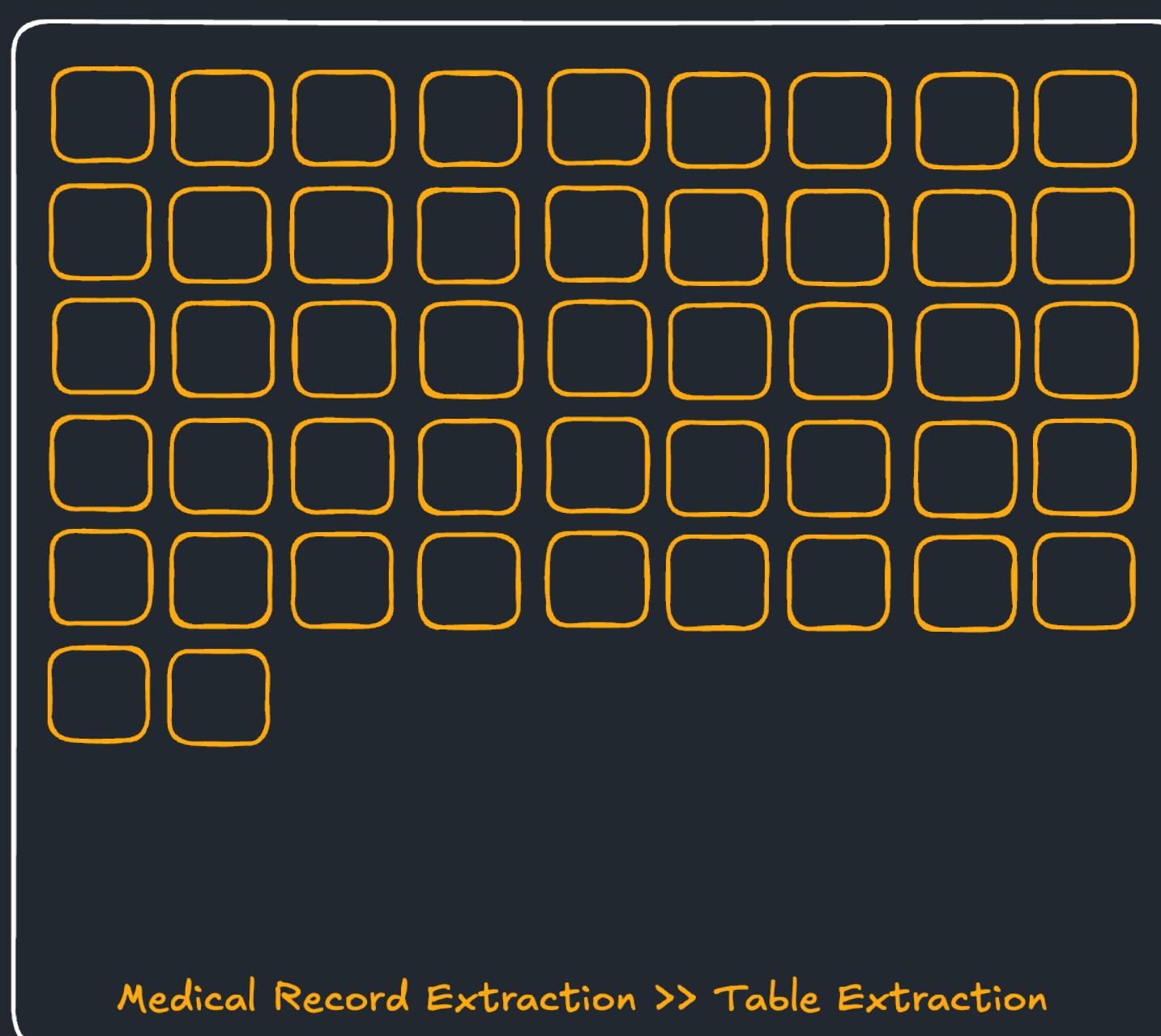
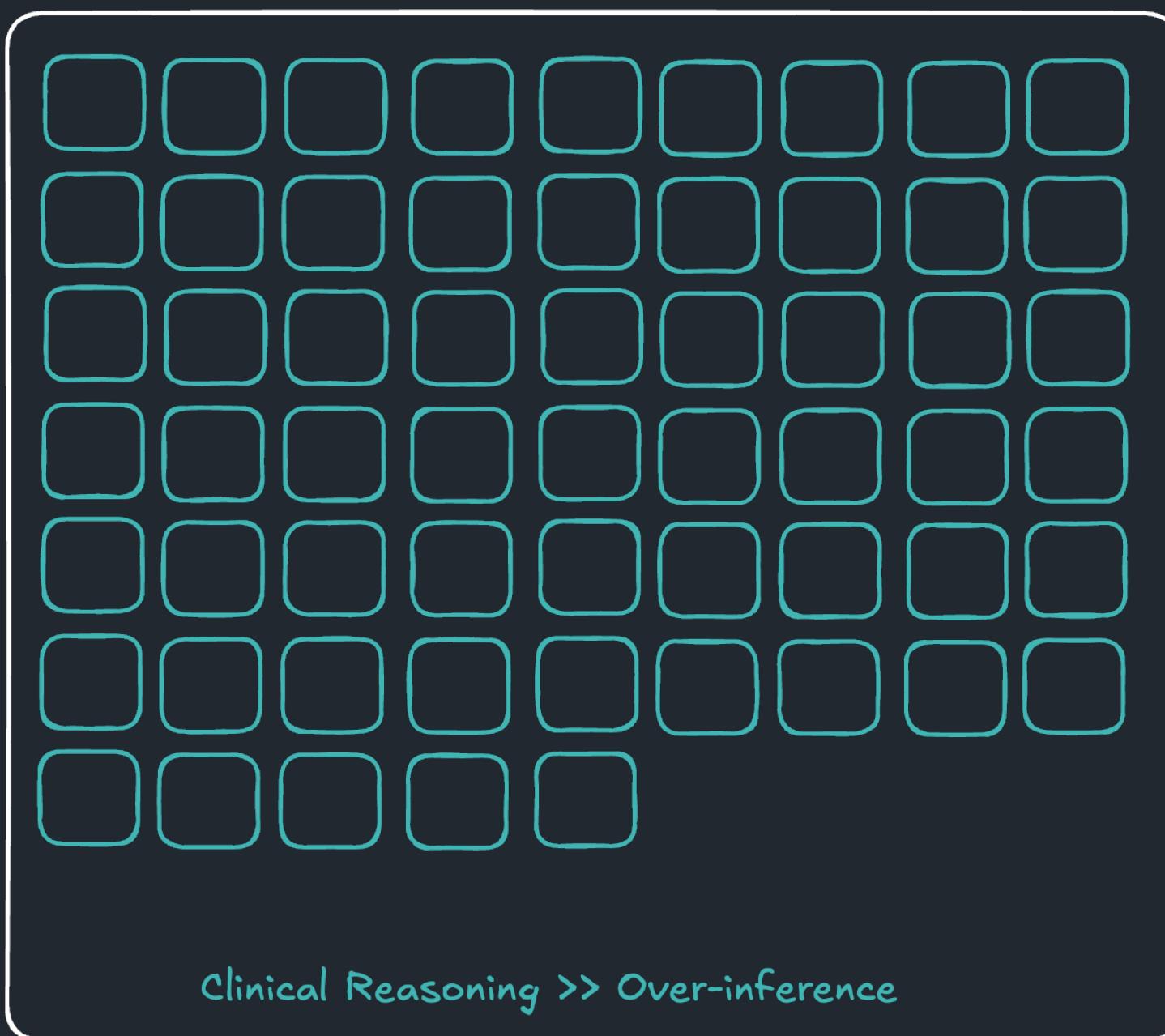
Confidentiality Notice – This fax contains confidential medical information protected by law. It is intended only for the named recipient. Any unauthorized disclosure, copying, or distribution is prohibited. If received in error, please notify sender immediately and destroy this document.

Using review data to
power your system

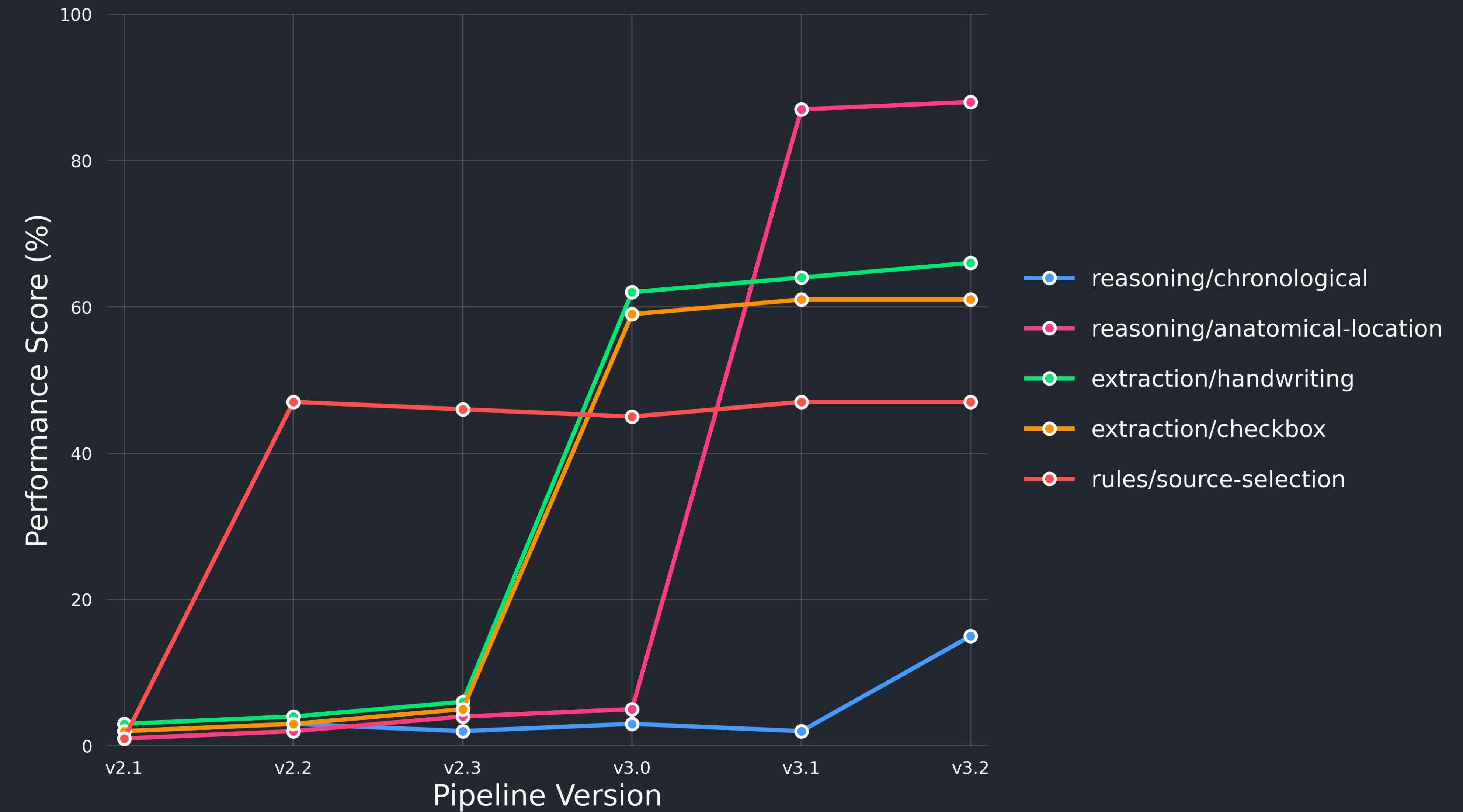
(1) Use metrics and failure modes from production data to prioritise work



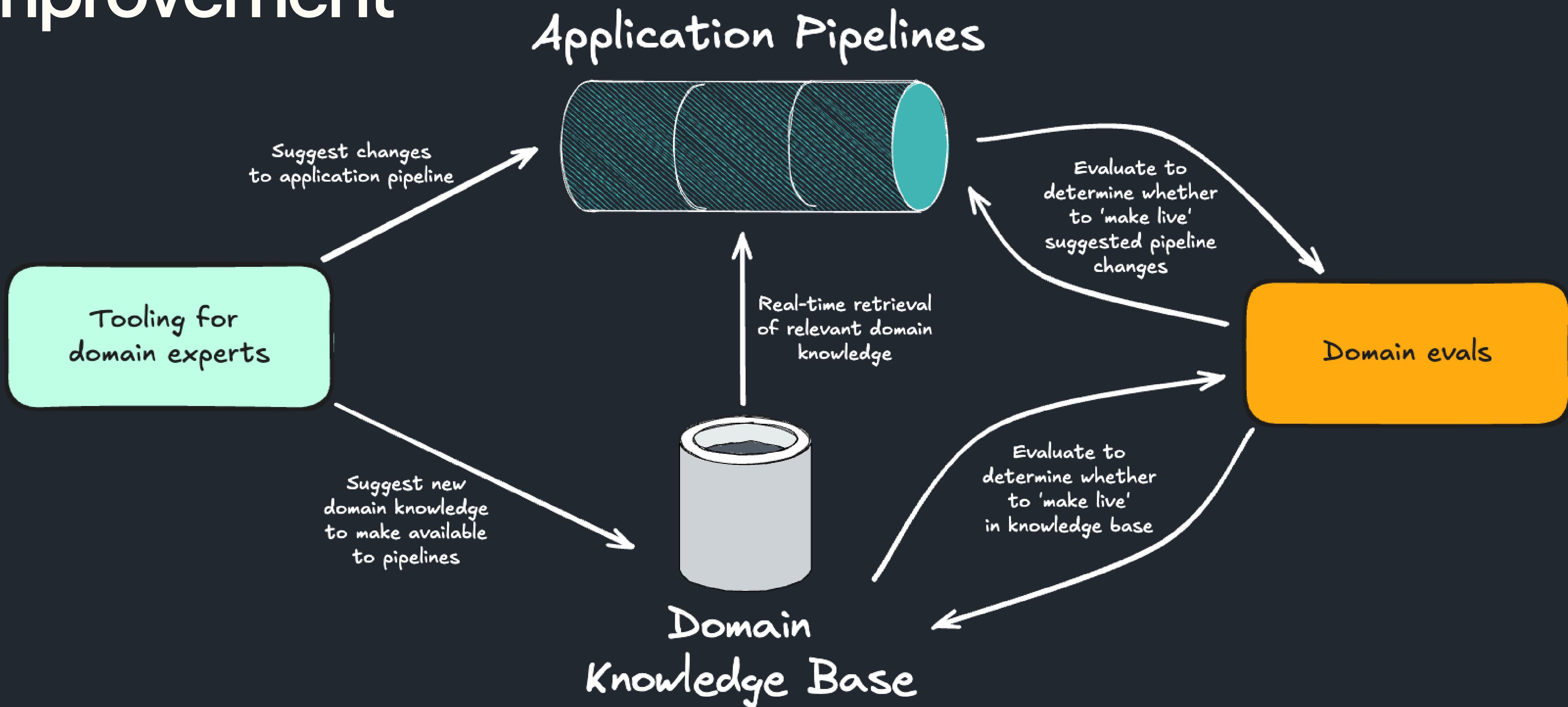
(2) Use failure mode datasets to test the impact of changes



(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 give 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

