

Week 6:

Reasoning Against Ourselves

Context *The Confident Fool: How Expertise Breeds Error*

Engineering *Building Systems to Make Reliable High Stakes Decisions*

LiGHT

Laboratory for intelligent
Global Health & Humanitarian
Response Technologies.

EPFL

ARIADNE LABS

ASHOKA
UNIVERSITY

**WORLD
ECONOMIC
FORUM**

Rwanda
Centre for the
Fourth Industrial
Revolution

The *Confident* Fool: How Expertise Breeds Error

Lecturer: Yusuf Kesmen



Light

Laboratory for intelligent
Global Health & Humanitarian
Response Technologies.

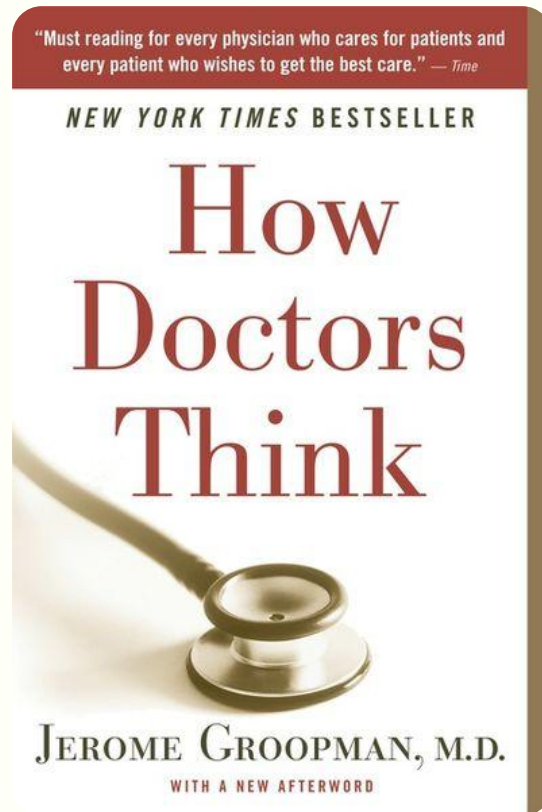
Why was Anne Dodge still ill?

Anne Dodge

- **15 years** of illness.
- Prior labels: **eating disorders**.
- Weight loss despite reported high food intake.

What does the label explain?

What remains unexplained?



This is not just one unusual case

≈1 in 20

US adults experience an outpatient
diagnostic error each year, by a 2014
estimate.



Autopsies reveal major missed diagnoses

23.5%

Median error rate

53 autopsy series, 1966–2002



How does a diagnosis take shape?

Before explaining errors, follow the ordinary reasoning process:

Represent

the
patient's
problem



Generate

possible
explanations



Investigate

ask, compare,
update



Decide

act now
or defer

First, turn the story into a clinical problem

I twisted my right knee yesterday
while playing football.



First, turn the story into a clinical problem

I **twisted** my **right knee** yesterday while playing **football**.



Acute

Yesterday

Right

Right knee

Traumatic

Sports-related

Twisting

Mechanism

Semantic qualifiers

- Organized as **structured knowledge in semantic axis**
- **Abstract representation** of presenting patient state
- Can be lossy

Time Course	Acute	Chronic
Laterality	Unilateral	Bilateral
Cause	Traumatic	Atraumatic

Semantic qualifiers

- Organized as **structured knowledge in semantic axis**
- **Abstract representation** of presenting patient state
- Can be lossy

Time Course	Acute	Chronic
Laterality	Unilateral	Bilateral
Cause	Traumatic	Atraumatic

The representation cues disease knowledge

- Doctors store the clinical knowledge in a structured database in the brain
 - **Context:** Who is at risk, and in what setting?
 - **Mechanism:** What process could explain the findings?
 - **Predictions:** What else should we expect?
- Semantic qualifiers activate the scripts for illnesses

Illness script

twisted, acute, right knee,
traumatic, yesterday, football



ACL

*Pivoting sports
Ligament Tear
Rapid Swelling*

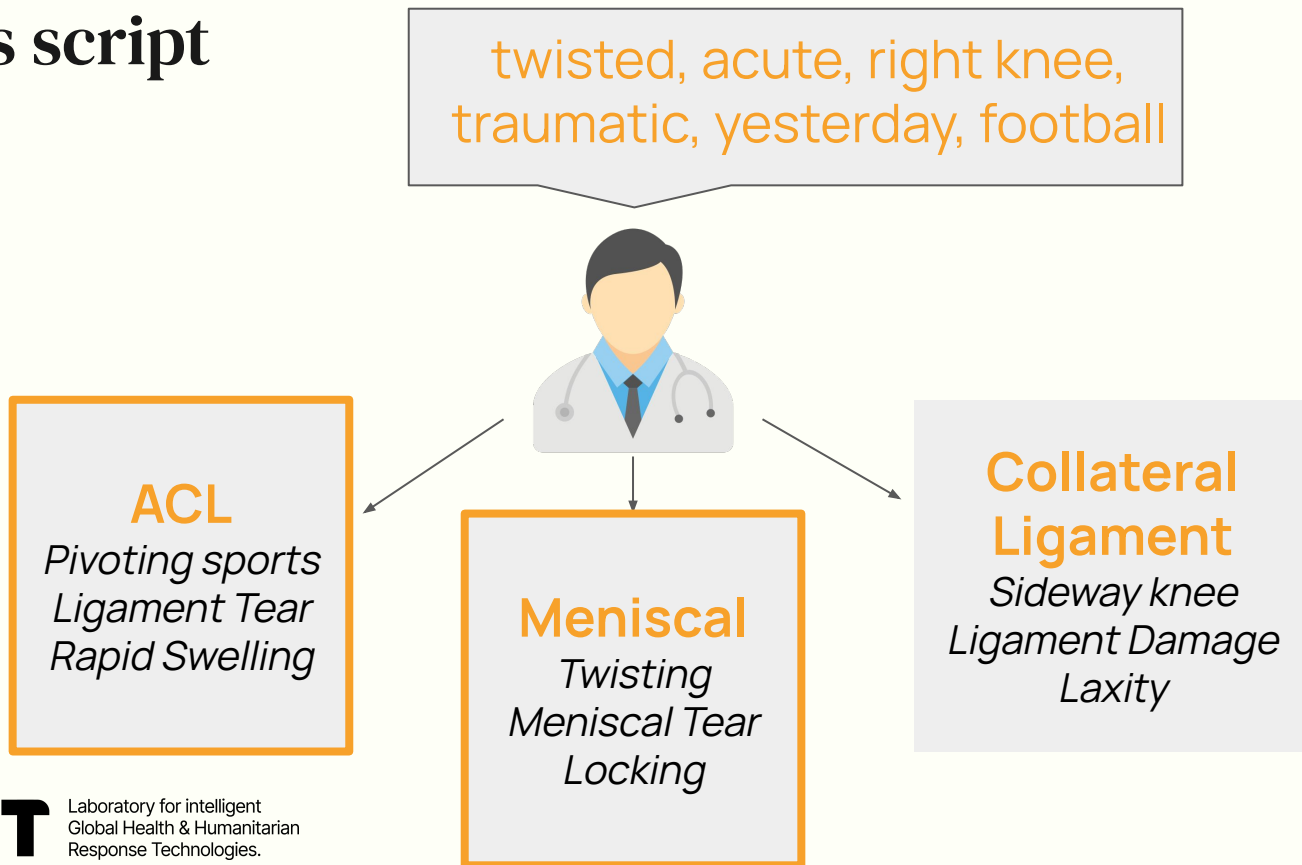
Meniscal

*Twisting
Meniscal Tear
Locking*

Collateral Ligament

*Sideway knee
Ligament Damage
Laxity*

Illness script



Same patient, different knowledge structures

I **twisted** my
right knee
yesterday while
playing **football**.



A twisting injury during
football? Did you hear a
pop or experience
immediate swelling?



Expert:
Illness Scripts

Twisting puts rotational
stress on the knee. That
could damage the
meniscus or ligaments.



Novice:
Biomedical Causal Networks

Why are some clinicians better?

- Even if they have the same knowledge:
 - They **structure the information** differently
 - They **reason on the information** differently

How do clinicians reason?

System 1

Recognition

Fast, effortless, intuitive

“This resembles pneumonia.”

System 2

Comparison

Slow, effortful, analytical

“What supports it?”

What conflicts? What else fits?”

This is called as **Dual-Process** theory

Daniel Kahneman (1934–2024)

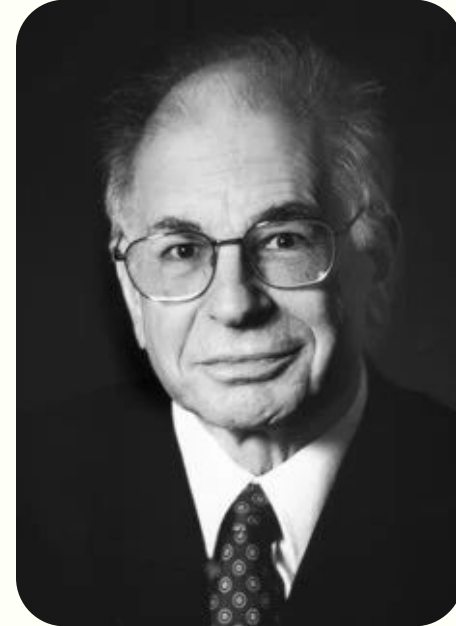
1974

Heuristics and
biases

Human use *shortcuts* in reasoning
which leads certain *biases*

2002

Nobel Prize for
Economic Sciences



Where can judgement fail?

Different failures can interrupt the same diagnostic process.

- **Framing** Wording changes our judgment.
- **Availability** Easy recall changes what comes to mind.
- **Anchoring** The first explanation resists revision.
- **Search satisfaction** The search stops too soon.
- **Overconfidence** Confidence exceeds accuracy.
- (...)

Which report feels safer?

REPORT A

90 / 100

patients survive

A This feels safer

REPORT B

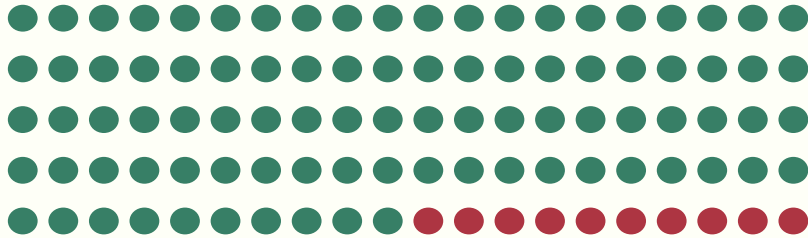
10 / 100

patients die

B This feels safer

Which report feels safer?

Same Distribution



• 90 survive • 10 die

Judgement can shift when equivalent information is expressed as a gain or a loss.

Framing: the numbers did not change

HANDOFF VERSION A

EMERGENCY DEPARTMENT NOTE

A patient with an **anxiety history** presenting with new breathlessness and tachycardia.

HANDOFF VERSION B

EMERGENCY DEPARTMENT NOTE

New **breathlessness and tachycardia** in a patient with an anxiety history.

Quick vote: where does R occur more often?

R

Try to recall words

A **R is the first letter**
Rain, river, road...

B **R is the third letter**
Try recalling these.

Availability: recall is not probability

B: the third letter is more frequent.

- Words beginning with R are easier to retrieve.
- We mistake **ease of recall** for frequency.

Easy to recall $\neq$ more likely

Clinicians tend to remember yesterday

YESTERDAY

MEMORABLE CASE

A dramatic pulmonary embolism in a breathless patient.



TODAY

NEW PATIENT

Breathlessness again.
Pulmonary embolism comes to mind immediately.

Availability in the clinic

Residents encountered new cases with **similar findings but different diagnoses** after reviewing earlier cases.

1.55

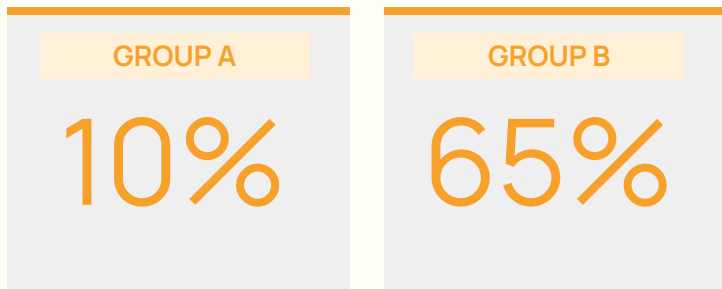
Similar findings

2.19

Other cases

Diagnostic accuracy score / 4.

Can a number move your estimate?



What percentage of UN member states were African in 1974?



Anchoring in a diagnostic workup

People tend to anchor to their initial starting points (e.g. chief complaints)

25%

Median UN estimate
Anchor: 10

45%

Median UN estimate
Anchor: 65

Clinicians tend to stuck to first hypothesis

Referral Note

Prior Assessment

“Migraine”
An earlier hypothesis



New Findings

Current Visit

New focal neurological
deficit and a sudden,
severe headache.

One correct diagnosis may still be incomplete.

Patient Presentation

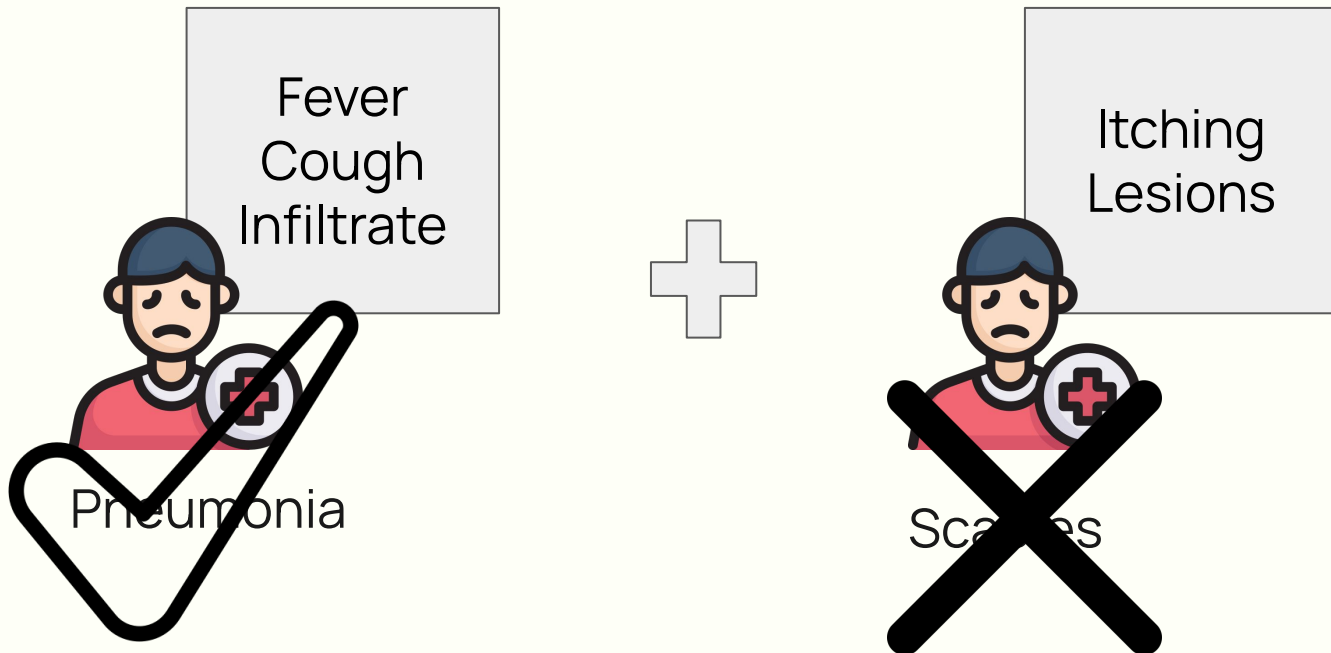
Fever, cough, and a focal chest infiltrate.

Also: intense nocturnal itch and lesions between the fingers.

“Pneumonia explains the respiratory findings.”

- A Stop: the diagnosis is found
- B Continue: an important finding remains

You found one diagnosis. Are you finished?



How much did confidence fall?

- **Easy cases:** 55.3% correct.
Confidence: 7.2 / 10.
- **Hard cases:** 5.8% correct.

Confidence on the hard cases?

A $\approx 1 / 10$

B $\approx 3 / 10$

C $\approx 6 / 10$



Overconfidence: confidence barely moved

Diagnostic Accuracy (%)

Reported Confidence (0–10)

Easy  55.3

Easy  7.2

Hard  5.8

Hard  6.4

Feeling **certain** is not the same as being correct.

At which handoff does “maybe” become “known”?



Repeated claims are not independent evidence.

Observed Evidence

1

Original assessment

≠

Copies of the claim

3

Confident records

Other ways judgment can fail

- **Attribution**

Patient characteristics or labels distort the interpretation of clinical findings.

- **Base-rate neglect**

Ignoring prior probabilities when evaluating new evidence.

- **Confirmation Bias**

Favoring evidence that supports an existing hypothesis while overlooking contradictory findings.

Cognitive and system factors in diagnostic error

Review of **100 diagnostic error cases**

74%

Cognitive contributors

65%

System contributors

The same case can involve **both** types of contributor.

How to fix this?

We know clinicians are vulnerable to cognitive biases.

- Can we simply teach them to think more carefully?
- Is **analytical reasoning** the solution?
- Is **knowing about biases** enough?
- What actually helps?

Can't we just think more analytically?

204 assistant doctor solves 20 clinical case

44.5%

Think Fast

45.0%

Think Slow

Thinking longer does not
necessarily mean thinking better



Can we teach clinicians to avoid bias?

191 medical students in a controlled trial

45%

Bias training

45%

Control group

Teaching cognitive forcing did not improve performance in this trial.



What about structured reflection?

1.55

Initial non analytical reasoning

2.03

After structured reflection

Reflection checks **supporting evidence, conflicts and alternatives**.

Second-year residents, same cases. Score / 4.

What about better knowledge?

91 residents in a controlled trial

40%

Training on
test set

24%

No training

Better knowledge can make
reasoning more reliable.



When clinical intuition deserves trust

01 Stable patterns

Clinical cues predict outcomes reliably.

02 Repeated experience

Similar cases offer opportunities to learn.

03 Useful feedback

Follow-up reveals correct and missed diagnoses.

Back to Anne Dodge



- 15 years, nearly 30 doctors. Repeatedly diagnosed with anorexia, bulimia, and psychological problems. Down to 37 kg
- Her boyfriend insisted she see another specialist
- Dr. Myron Falchuk questioned previous assumptions, listened to her story from the beginning

Back to Anne Dodge

- **Anchoring** Keeping the “eating disorder” label despite new evidence.
- **Confirmation bias** Reading continued weight loss as support for the existing label.
- **Diagnostic momentum** Treating a diagnosis passed between clinicians as an established fact.
- **Search Satisfaction** Stopping the search once a plausible explanation is available.

What should a safer system do?

- Represent the patient's problem.
- Keep plausible diagnoses in play.
- Ask questions that distinguish them.
- Update with new evidence.
- Act or defer with uncertainty visible.
- Probabilistically grounded



End

CONTEXT

Questions?

Light

Laboratory for intelligent
Global Health & Humanitarian
Response Technologies.

Building Systems to Make *Reliable* High Stakes Decisions

Lecturer: Yusuf Kesmen

LiGHT

Laboratory for intelligent
Global Health & Humanitarian
Response Technologies.

Outline

- 01 Introduction
- 02 What is the problem?
- 03 Why does it happen?
- 04 What is there?
- 05 MoBayes

Building **Systems** to Make **Reliable** High Stakes Decisions

- **System**
- **Reliable**
- **High Stakes**
- **Decision**

This is conversational clinical decision support.

Clinical decision support before LLMs

1959

Ledley & Lusted

Probability and logic

1976

MYCIN

Expert-written rules

1982

INTERNIST-I

A curated disease-finding KB



Medical Decision with Language Models

MedGemma



MEDITRON

Open medical training



Meditron-3

LiGHT

Laboratory for intelligent
Global Health & Humanitarian
Response Technologies.

Logos and illustrations: Google Research / Google DeepMind (MedGemma, Med-PaLM 2)

Medical LLMs: strong exam performance

91.1%

Med-Gemini on MedQA

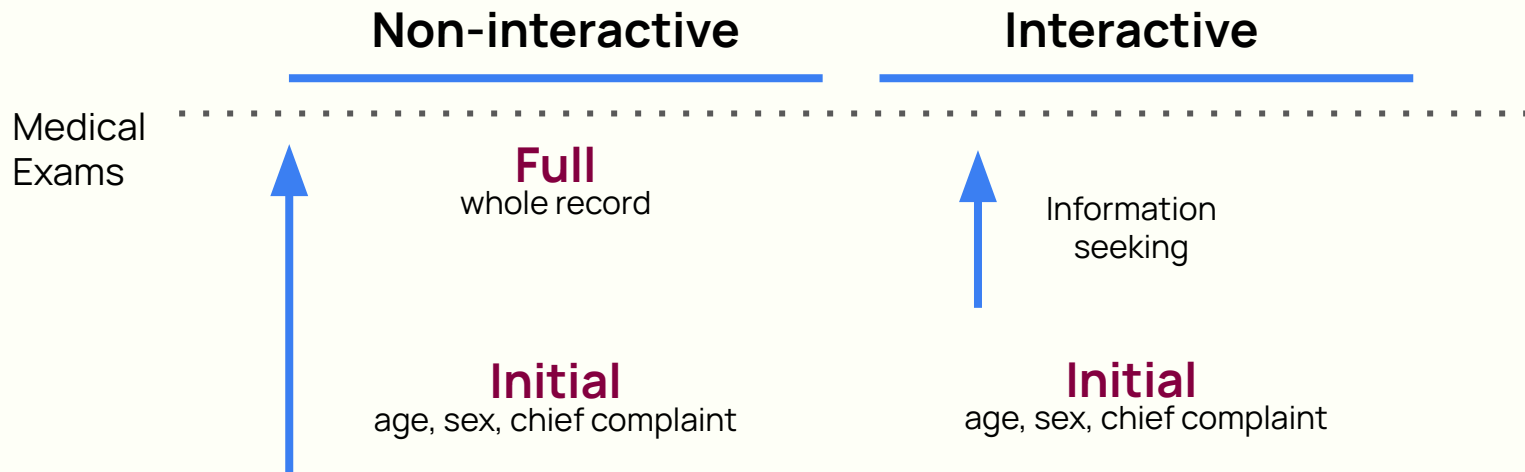
What does an exam score leave **untested**?

Why are they **not** used in **real practice**?

- Regulations
- Real life is noisy

High Stakes

Clinical decision making is reasoning under uncertainty



The target: a realistic clinical consultation

Incomplete patient answers

A limited question budget

Several possible conditions

Abstention when unresolved
(...)



Multi-Turn,
Multi-Step,
Abductive
Conversations

Single-turn
MedQA

LLMs also include medical decision failures

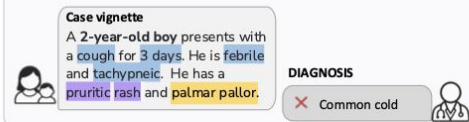
It is empirically shown that LLMs have problem with

- Anchors
- Search Satisfaction
- Framing
- Overconfidence
- etc

You can test it yourself.

An example: LLMs are single hypothesis trackers

(a) Single-turn, Single-label



(b) Interactive, single-label

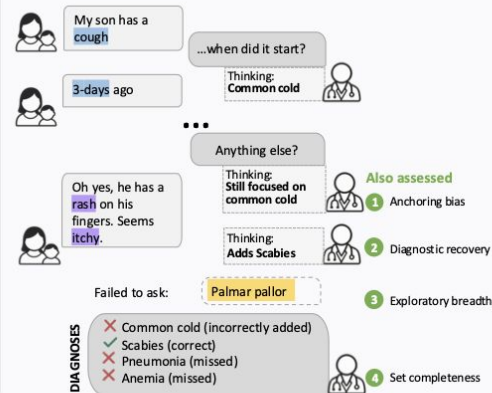


Scored against single reference diagnosis

Reference: Pneumonia (hidden)

Exact Score: 0/1

(c) Multi-turn, Multi-label

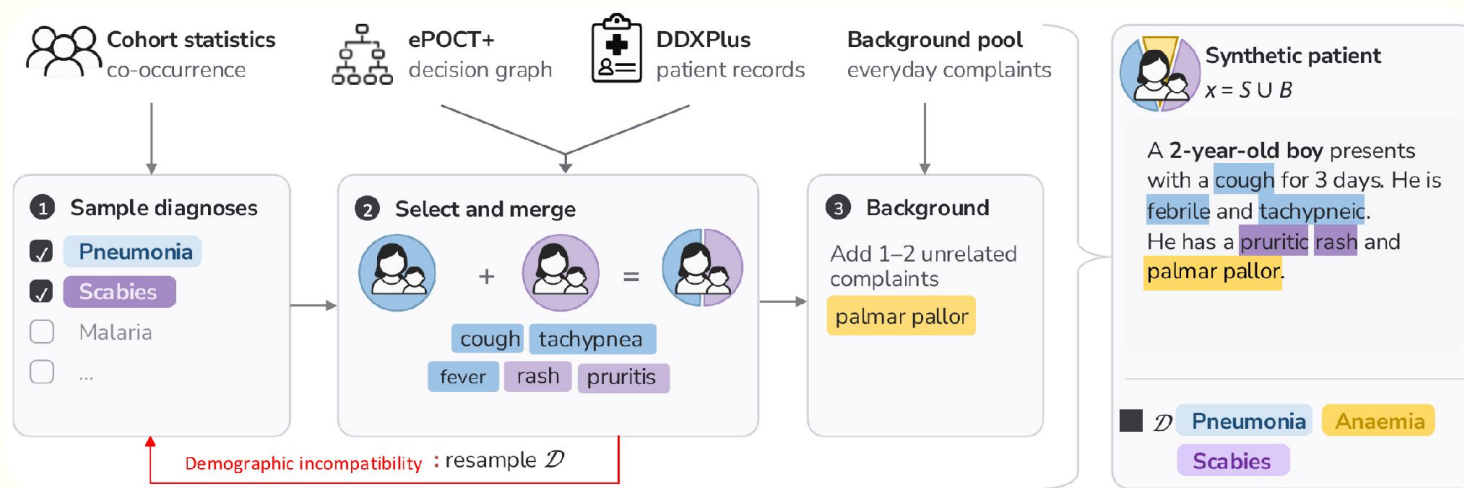


Scored against K reference diagnoses

Reference: Pneumonia, Scabies, Anemia (hidden, K also hidden)

Answer: Jaccard Score: 1/4

CLIMB: constructing a patient with several diagnoses

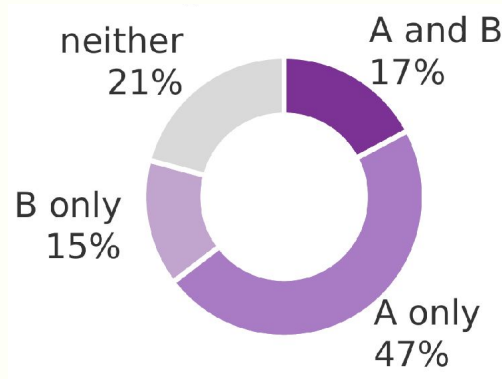


Recover the **complete diagnosis set**, with the true count hidden.

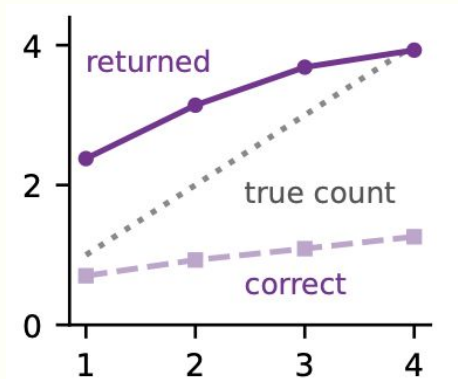
Why does it happen?

CLIMB: the opening diagnosis dominates

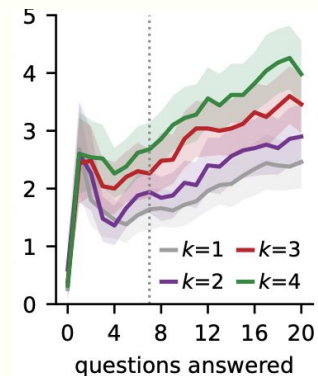
Models often retain the opening diagnosis. Asking for more diagnoses can add unsupported alternatives.



Opening diagnosis dominates



Correct count plateaus



Wrong diagnoses accumulate

Why do these failures happen?

- Inconsistent probabilities

Estimates keep changing

- Coarse estimates

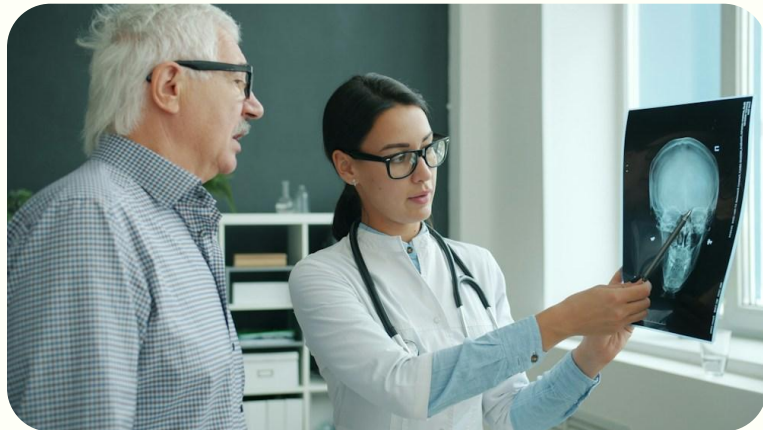
Subtle differences disappear

- Overconfidence

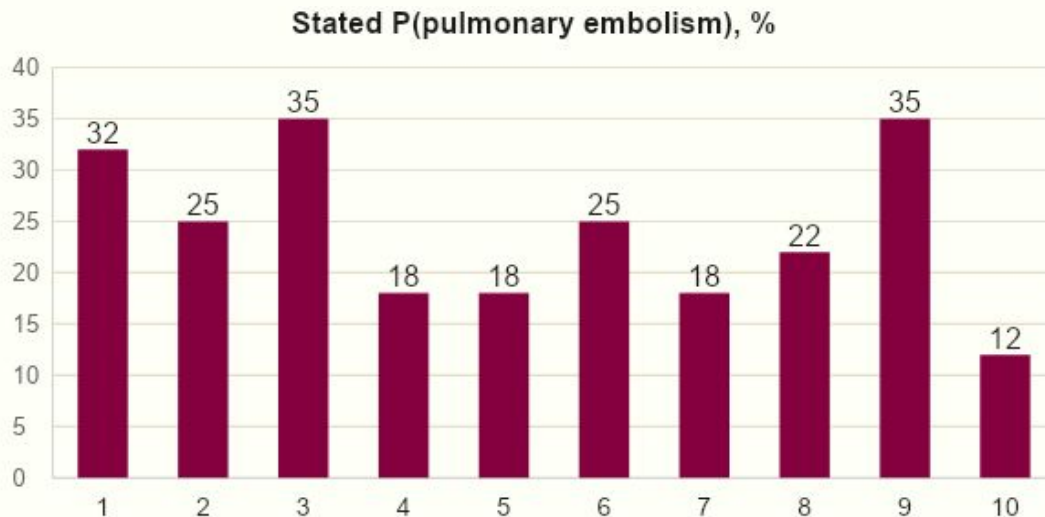
Confidence exceeds accuracy

- Hallucination

Unsupported claims appear



For example: Probability is not a stable belief



12–35%

Same prompt and complaint.

It is also very coarse grained.

Because: what does the training objective reward?

Pretraining

Predict the next token.

Learn data

$$\mathcal{L}(\theta) = - \sum_t \log P_{\theta}(x_t \mid x_{<t})$$

Post-training

Imitate demonstrations.

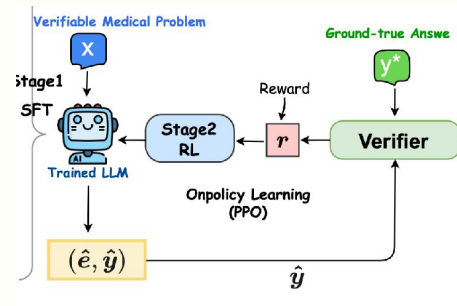
Learn preference

However, data and reward choices can carry **human biases**.

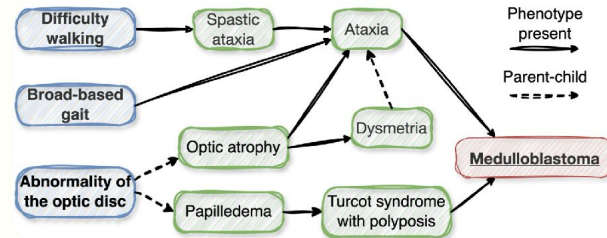
Reasoning models and “System 2”

- Consume some **more tokens** before giving the answer
- Isn't it like **system 2** thinking? We spend some time rather than giving answer directly
- But, reasoning models also have the **similar problems**.

HuatuogPT-o1

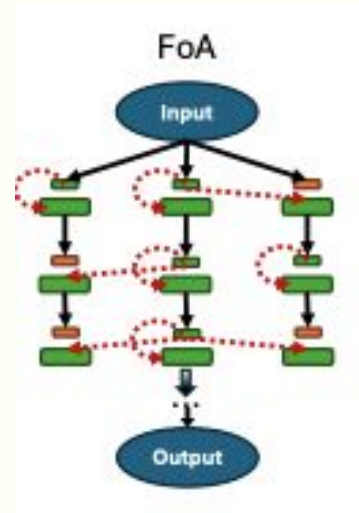


MedReason-8B



Structuring the Reasoning Process

- Chain of Thought: Explicit sequence of steps
 - Are the premises correct?
- Self-reflection: Critique and revision of an answer
 - Did it repeat the same assumptions?
- Self-consistency: Branch and do majority voting
- Fleet of Agents: Evolutionary search in solution space
- (...)



Decision models

Decision models are gaining attention for fast choices among defined options. They return structured scores that software can act on.



Jev · TypeSafe



Microsoft-Decision-1

OpenAI Developers



API Docs
Decisions

OpenAI Decisions API

Diagnosis is abductive reasoning

- **Deduction** predicts findings from a hypothesis.
- **Induction** learns patterns from examples.
- **Abduction** compares explanations of observed findings.

We need proper probabilistic structures.

- BedLLM, BIRD, DeLLMa, ...

BED-LLM: which question is informative?

Current history: $h_2 = (("Born in the 20th century?", "Yes"), ("Is this person male?", "Yes"))$.

(A) **Extract beliefs (§3.1)**. Construct hypothesis set $\Theta^{\text{cand}} = \{\theta_n\}_{n=1}^N$ by sampling candidate hypotheses, $\theta \sim p_{\text{LLM}}(\theta; h_2)$, and rejecting any θ that is inconsistent with h_2 . Use joint model $p(\theta, y; h_2, x) = p_f(\theta; h_2)p_{\text{LLM}}(y; [\theta, x])$, where $p_f(\theta; h_2)$ is uniform over Θ^{cand} .

Barack Obama Steve Irwin Hugh Laurie Banksy Elvis Presley ...

(B) **Generate candidate questions (§3.2)**. Sample $\tilde{x}_{1:M} \sim p_{\text{LLM}}(\tilde{x}_{1:M}; [h_{t-1}, \Theta^{\text{cand}}])$.

$\tilde{x}_1$: "Was this person born in Antarctica?"

$\tilde{x}_2$: "Was this person born in the 19th century?"

$\tilde{x}_3$: "Is this person an artist?"

$\tilde{x}_4$: "Is this person European?"

$\tilde{x}_5$: "Does this person prefer thrash metal over death metal?"

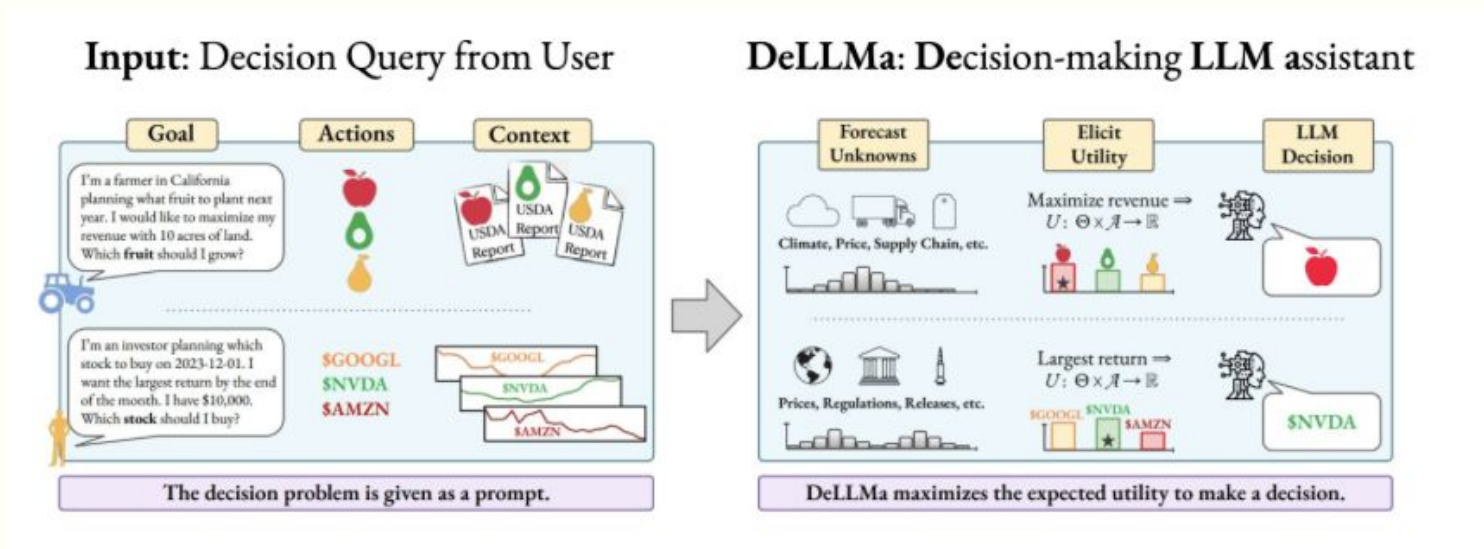
(C) **Estimate EIG (§3.3)**. Compute $\underbrace{H[\tilde{p}(y; [h_2, x])]}_{\text{marginal entropy}} - \underbrace{\frac{1}{N} \sum_{n=1}^N H[p_{\text{LLM}}(y; [\theta_n, x])]}_{\text{expected conditional entropy}} \approx \text{EIG}_{\theta}(x; h_2)$.

Question	Marg.	Exp. cond.	EIG	Intuition
$\tilde{x}_1$ Born in Antarctica?	≈ 0	≈ 0	≈ 0	Answer "No" for all θ_n
$\tilde{x}_2$ Born in 19th century?	≈ 0	≈ 0	≈ 0	Redundant given h_2
$\tilde{x}_3$ An artist?	0.56	0.41	0.15	Uneven split
$\tilde{x}_4$ European?	0.97	0.17	0.80	Balanced split; crisp answers
$\tilde{x}_5$ Thrash vs. death metal?	0.89	0.88	0.01	Uncertain even given θ

(D) **Select and ask the best question**. Choose $x_3 = \tilde{x}_4$ by maximizing EIG. Output this to the user.

(E) **Update the history**. Observe answer, y_3 . Set $h_3 = (h_2, (x_3, y_3))$.

DeLLMa: Decision Making Under Uncertainty with LLMs

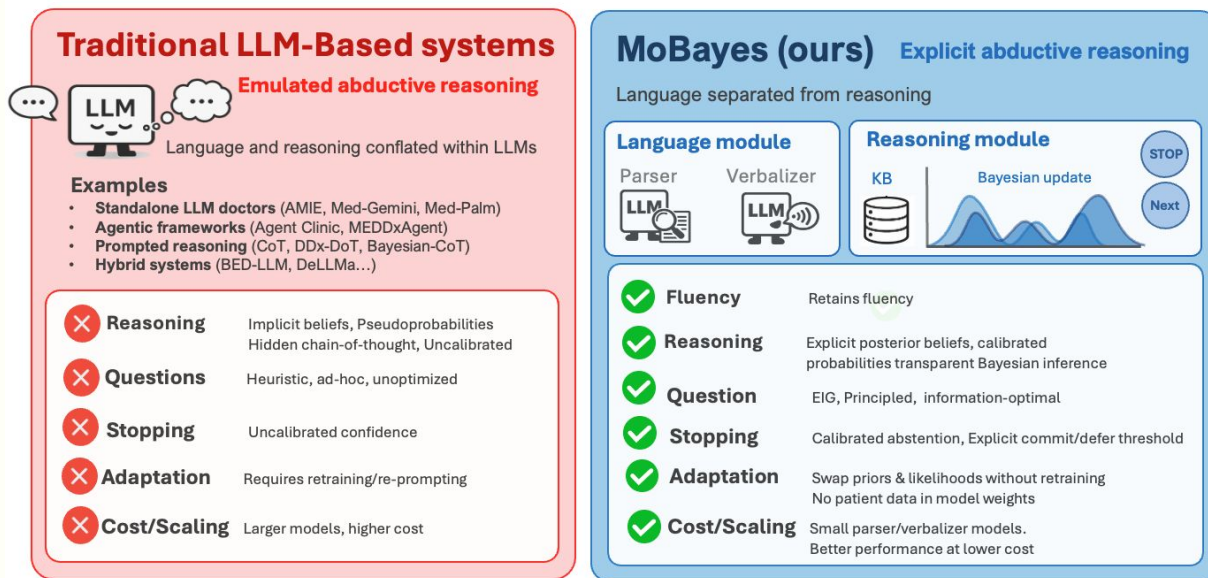


Clinical reasoning needs

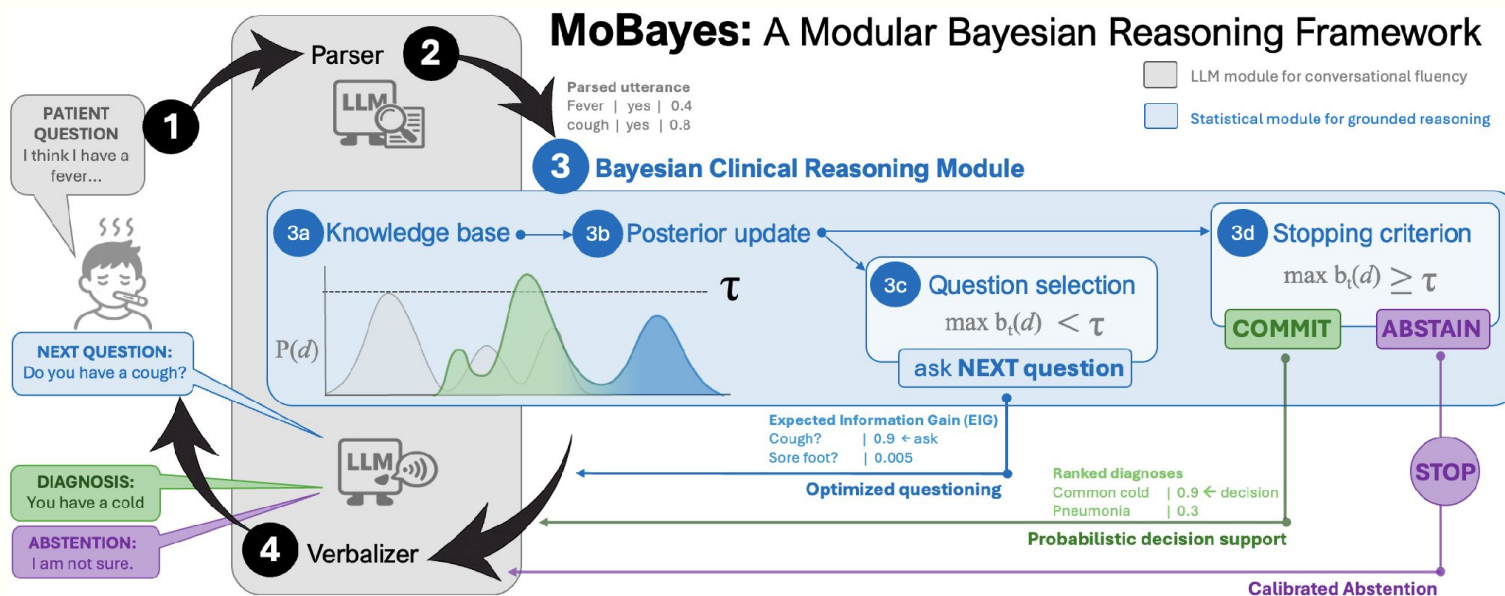
- A **belief distribution** over diseases
- A numerical objective for the next question
- A policy for **answering or deferring**
- An output that represents uncertainty



MoBayes separates language from inference



MoBayes: the full consultation loop



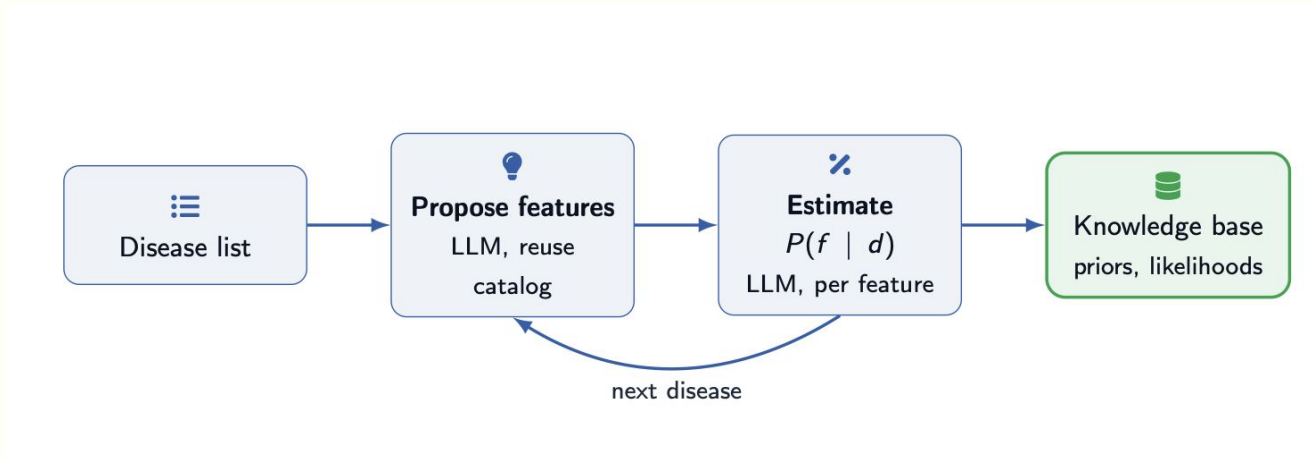
Building a knowledge base from records

Hospital records should be used:

- To generate a probability table
- It defines a closed universe. When evidence is insufficient, refer to clinician

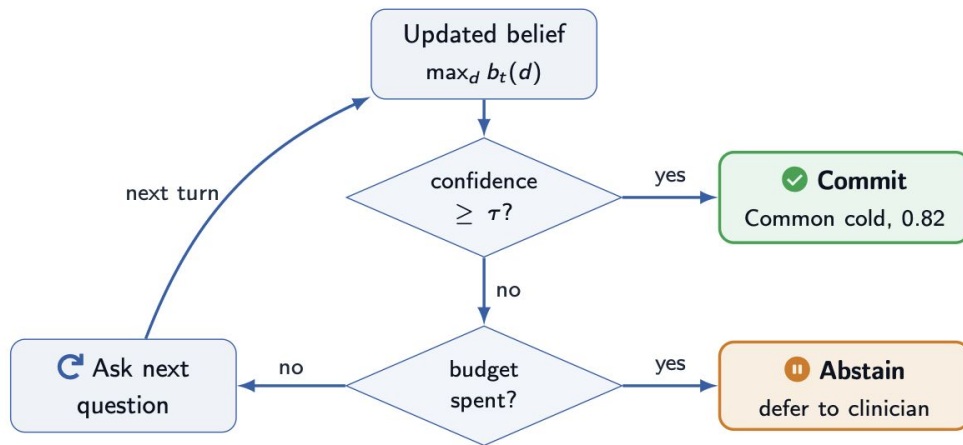
Disease	Fever	Cough
Pneumonia	0.80	0.85
Bronchitis	0.35	0.90
Asthma	0.10	0.40

Building an external KB from an LLM



Elicit probabilities once, then **inspect and validate** the table.

Abstention is a decision policy



A warm-up phase blocks early commits; τ tunes triage versus safety.

Belief update and question selection

BELIEF

Track the current disease probabilities.

UPDATE

Use new evidence to revise those probabilities.

NEXT QUESTION

Ask what is expected to reduce uncertainty most.

Uncertain evidence has less influence

“I think I lost weight.”

Parser confidence: $0 \leq c \leq 1$

$$\text{Adjusted likelihood} = c \times P(v \mid d) + (1 - c)$$

$$c = 1$$

Usual likelihood update

$$c = 0$$

Belief stays unchanged

Accuracy and coverage must be scored together

We need new metrics to define the trade-off between selective accuracy. We want AI doctors that are never wrong!

90%

Selective accuracy: 54 / 60

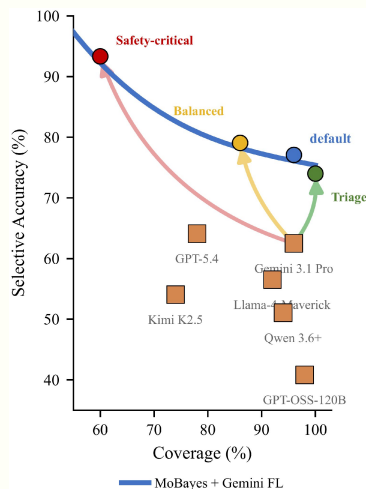
60%

Coverage: 60 / 100

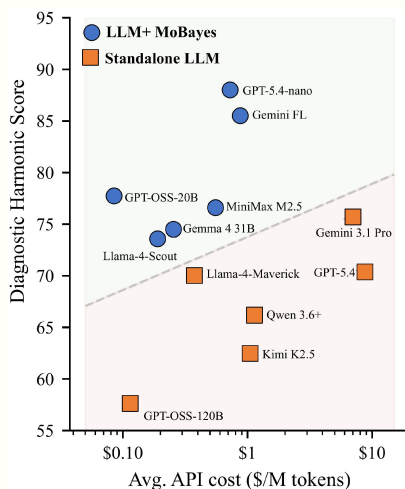
$$\text{DHS} = \frac{2AC}{A + C} = \frac{2 \times 0.90 \times 0.60}{0.90 + 0.60} = 0.72$$

MoBayes: three results to examine

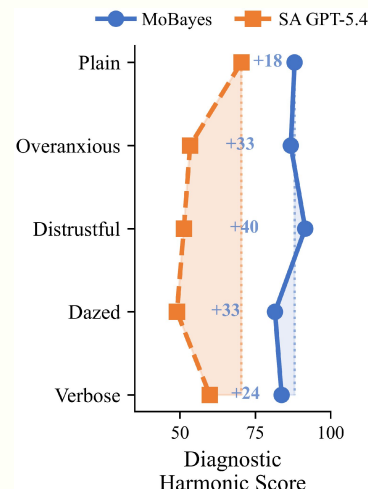
Controllable



Cost Efficient

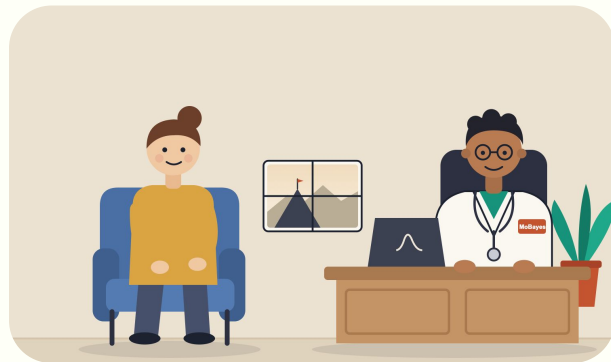


Robust



CLINIVERSE: scaling the knowledge base

- **Extract** medical knowledge from LLMs.
- **Build** a structured knowledge base.
- **Make** it inspectable and reusable.



Questions?

Thank You

Except where otherwise indicated, original content in this presentation is © 2026 Hartley, Keitel, Klein, Sasu, Suttels, Arni, Kesmen, Nemo and licensed under CC BY 4.0: <https://creativecommons.org/licenses/by/4.0/>.

Third-party materials are excluded from this license and remain subject to their indicated rights.

