

TO TRUST AI, BE CLEAR AND VERIFY

FASTER & BETTER FIELD DATA ANALYSIS FOR REGULATED MEDICAL PRODUCTS

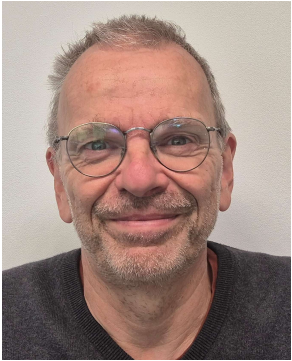
Ivo Canjels, Philips IGT
Dennis Dams, TNO-ESI

Powered by industry, academia and TNO



TNO ESI

TEAM DELPHI



Ivo Canjels



Dennis Dams



Richard Doornbos



Emile van Gerwen



Christos Kitsanelis



Anca Lichiardopol



Micha Lipplaa



JeanLuc Oudshoorn



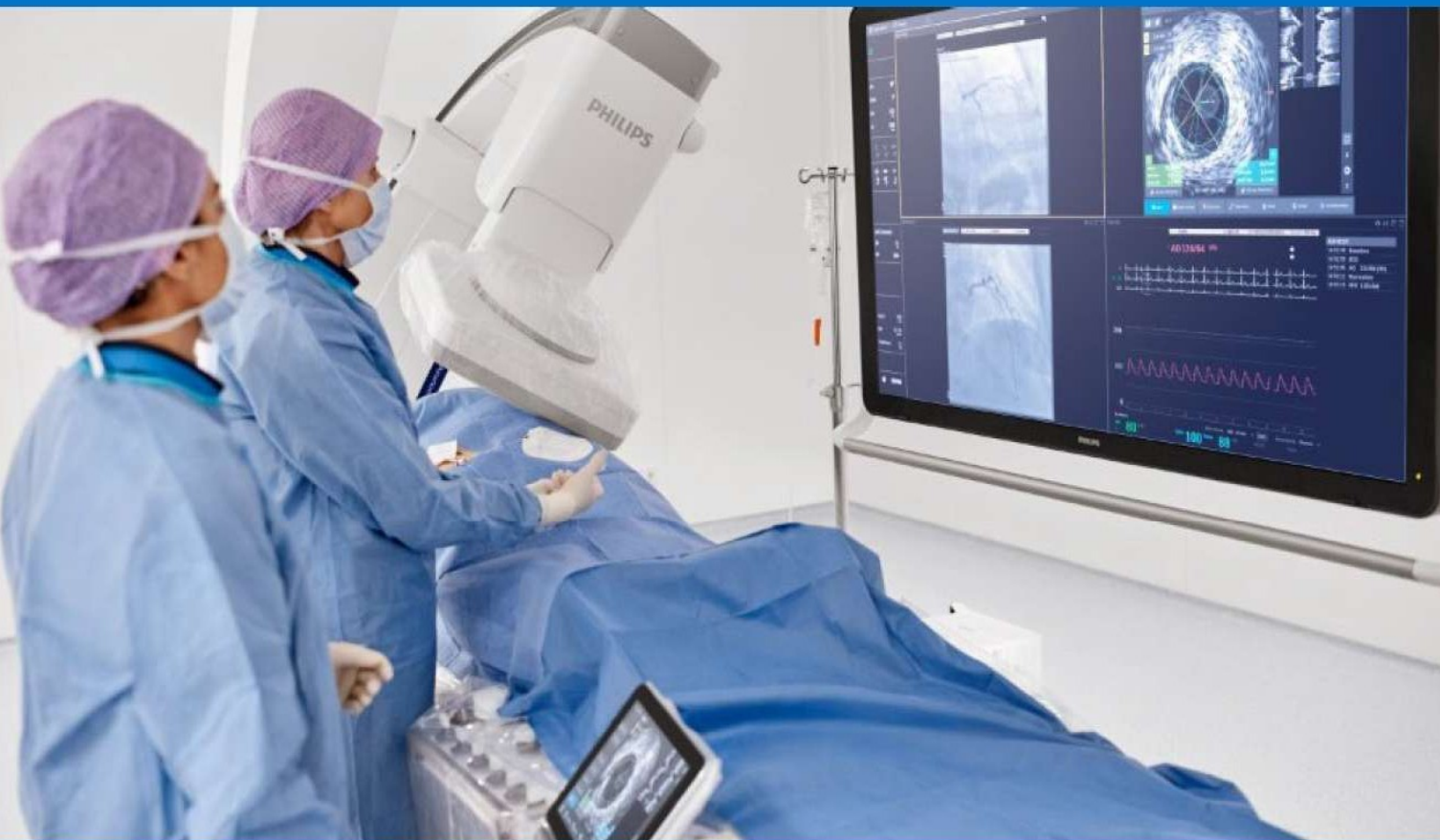
Monika Srivastava



Erik Teesink

Philips Image Guided Therapy

At Philips, our purpose is to improve people's health and well-being through meaningful innovation. We aim to improve 2.5 billion lives per year by 2030, including 400 million in underserved communities.



PHILIPS – MARKET LEADER IN INTERVENTIONAL X-RAY SYSTEMS

- Philips IGT (Image Guided Therapy) is worldwide market leader in high-end interventional X-ray systems
- Used in patient treatment in minimally invasive procedures:
 - Cardiology
 - Neurology
 - Oncology
 - Other clinical domains
- Installed base > 30,000 systems



Azurion System

QUALITY CONTROL UNDER REGULATORY OVERSIGHT IS EXPENSIVE

The *current reality* in managing customer feedback:

- Installed base generates >200,000 user feedback reports (yearly)
 - *Large teams needed* for analysis → Millions €
- Analysis must be *systematic* to ensure correction and prevention
 - *Handle information overload*: large amounts of unstructured text
 - *Manage human interpretation*: analysis tends to be subjective

Examples of quality processes that are subject to *regulatory oversight*:

- Customer feedback and complaints
- Enhancement Requests
- CAPA (Corrective and Preventive Actions)
- PPMRR (Production/Post-Market Risk Reviews)



CAN AI HELP?

Opportunity: *Improve quality while reducing cost*

- Can save > 1 Million per year
- By AI support during analysis of feedback, complaint and CAPA data

Challenge: *How to ensure trustworthiness of AI's answers?*

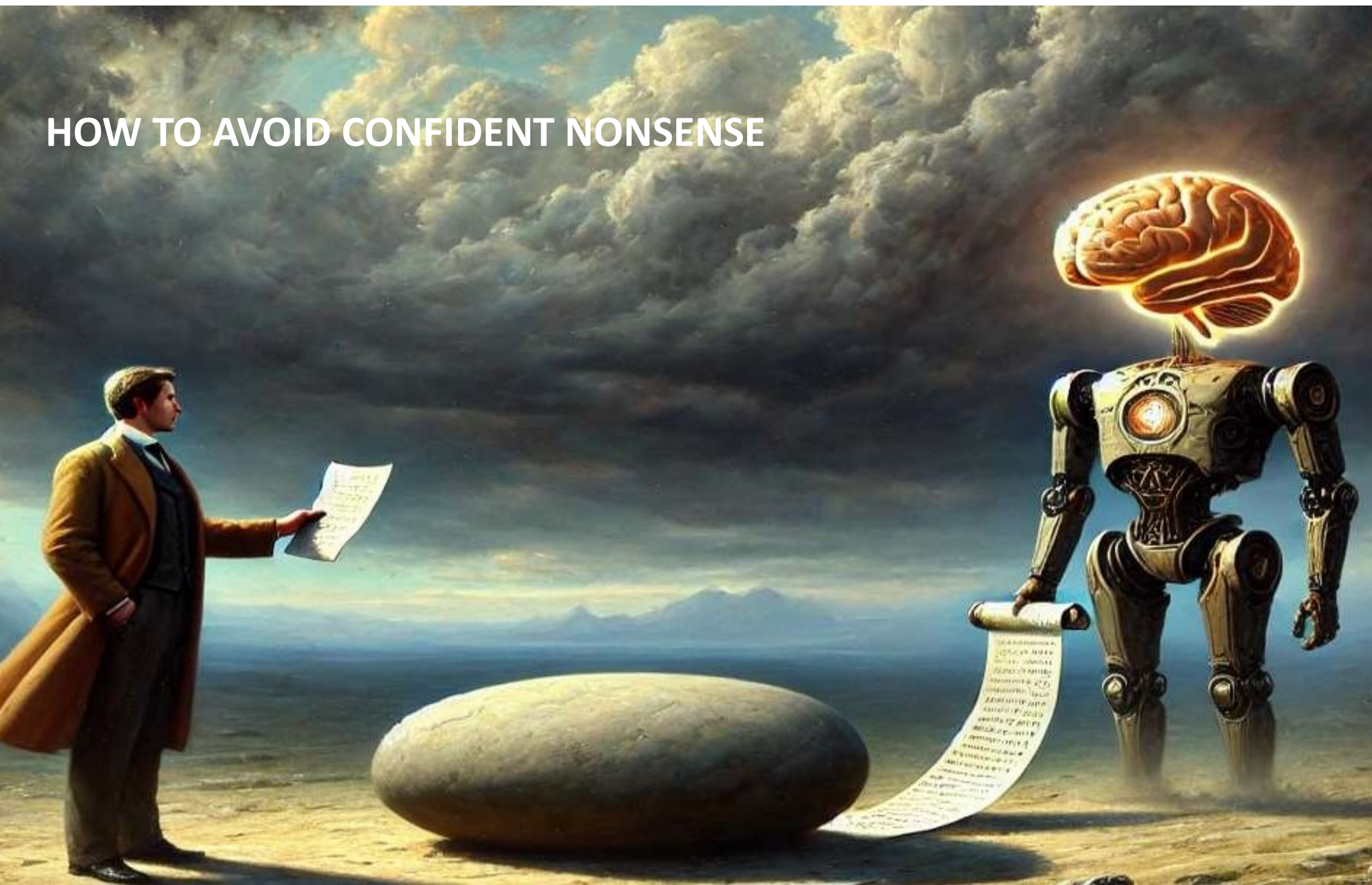
- Correctness?
- Reproducibility?
- Validation?



Project DELPHI – Joint initiative Philips IGT & TNO-ESI

- *Experiment* – e.g. which aspects of the quality process can be supported by AI?
- *Measure* – e.g. what is the ground truth to compare to?
- *Validate* – e.g. how do we convince ourselves and the authorities?

HOW TO AVOID CONFIDENT NONSENSE

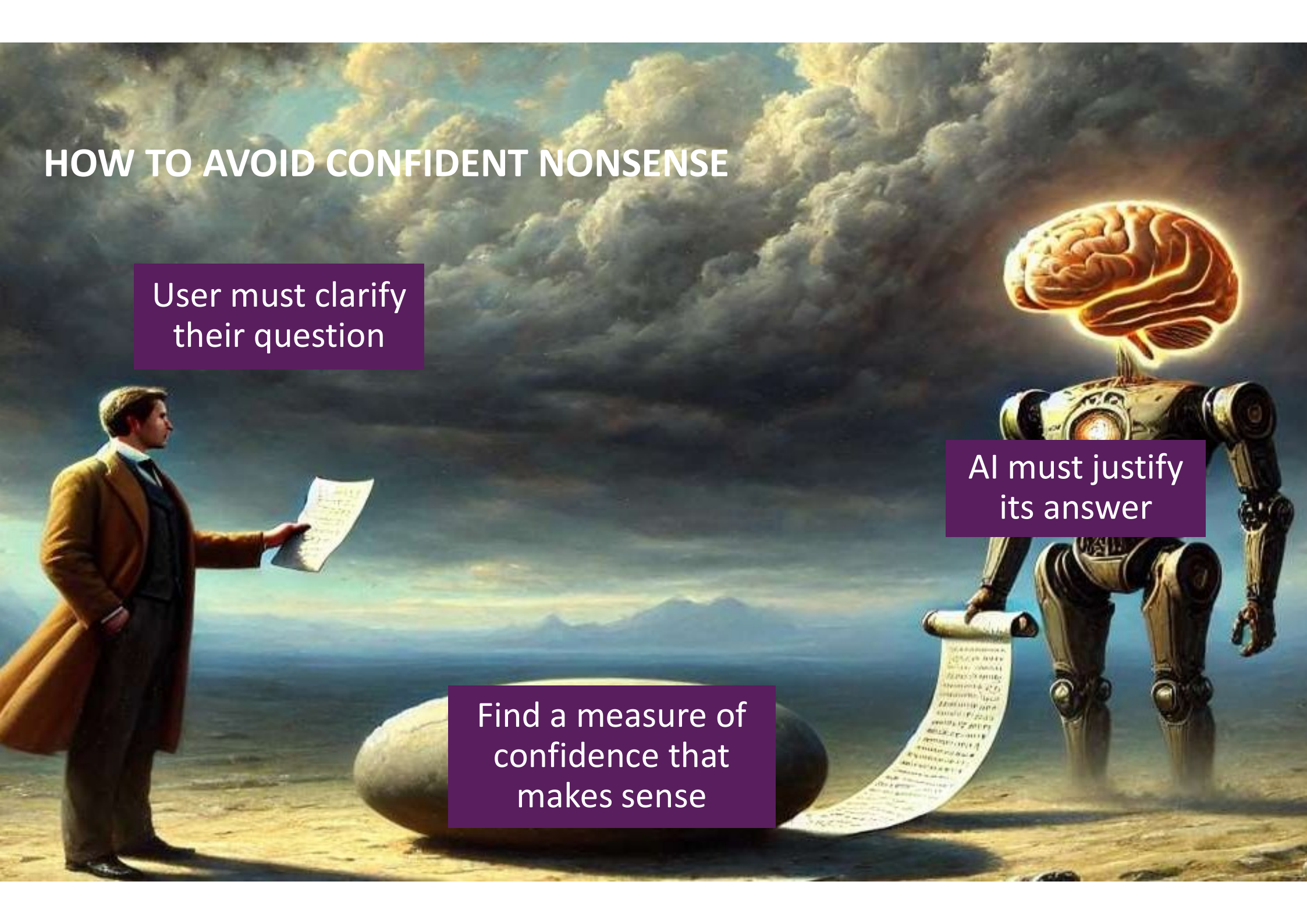


HOW TO AVOID CONFIDENT NONSENSE

User must clarify
their question

AI must justify
its answer

Find a measure of
confidence that
makes sense

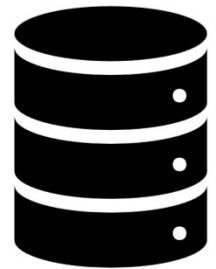


LET'S LOOK AT A RECENT USE CASE



Geometry Module

“How many feedback reports are about a problem with the **geometry module** that result in a **loss of functionality**?”



report
database

WE LET AN LLM LOOK AT EACH REPORT

Prompt:

Tell me whether the following report is about a problem with the geometry module that resulted in a loss of functionality:

feedback report



YES /

“Complex classification”

THE LLM NEEDS TO ACCOMPANY ITS ANSWER BY A RATIONALE

Prompt:

Tell me whether the following report is about a problem with the geometry module that resulted in a loss of functionality:

feedback
report
4879421



YES, *because:*

Malfunctioning of
shutter button

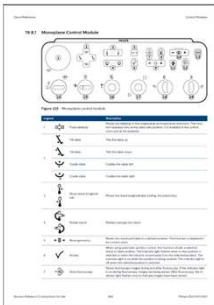
YOU NEED TO TEACH THE LLM SOME DOMAIN-SPECIFIC KNOWLEDGE

Prompt:

Tell me whether the following report is about a problem with the geometry module that resulted in a loss of functionality:

feedback
report
4879421

Documentation of the module:



YOU NEED TO EXPLAIN TO THE LLM HOW TO READ THE REPORT

Work Order #: WO-11739790

12-Nov-2024

The system meets the specification for the performed service and is returned to use

Resolution : Patient entry option not available.

Resolution : Reload DFI and Reset SCPU. Reconnect subsystem.

Performed calibration, tested and handed over machine to user in working condition

Problem Description by Engineer : Problem description by engineer :

system is not booting

Tools & Equipment : MULTIMETER,2025-02-19 00:00:00

Work Order #: WO-11739412

The system does not meet the specification for the performed service and is not returned to use

Problem Description by Engineer : Problem description by engineer :

system is not booting

Diagnostic performed by Engineer : Checked and found DFI is not booting.

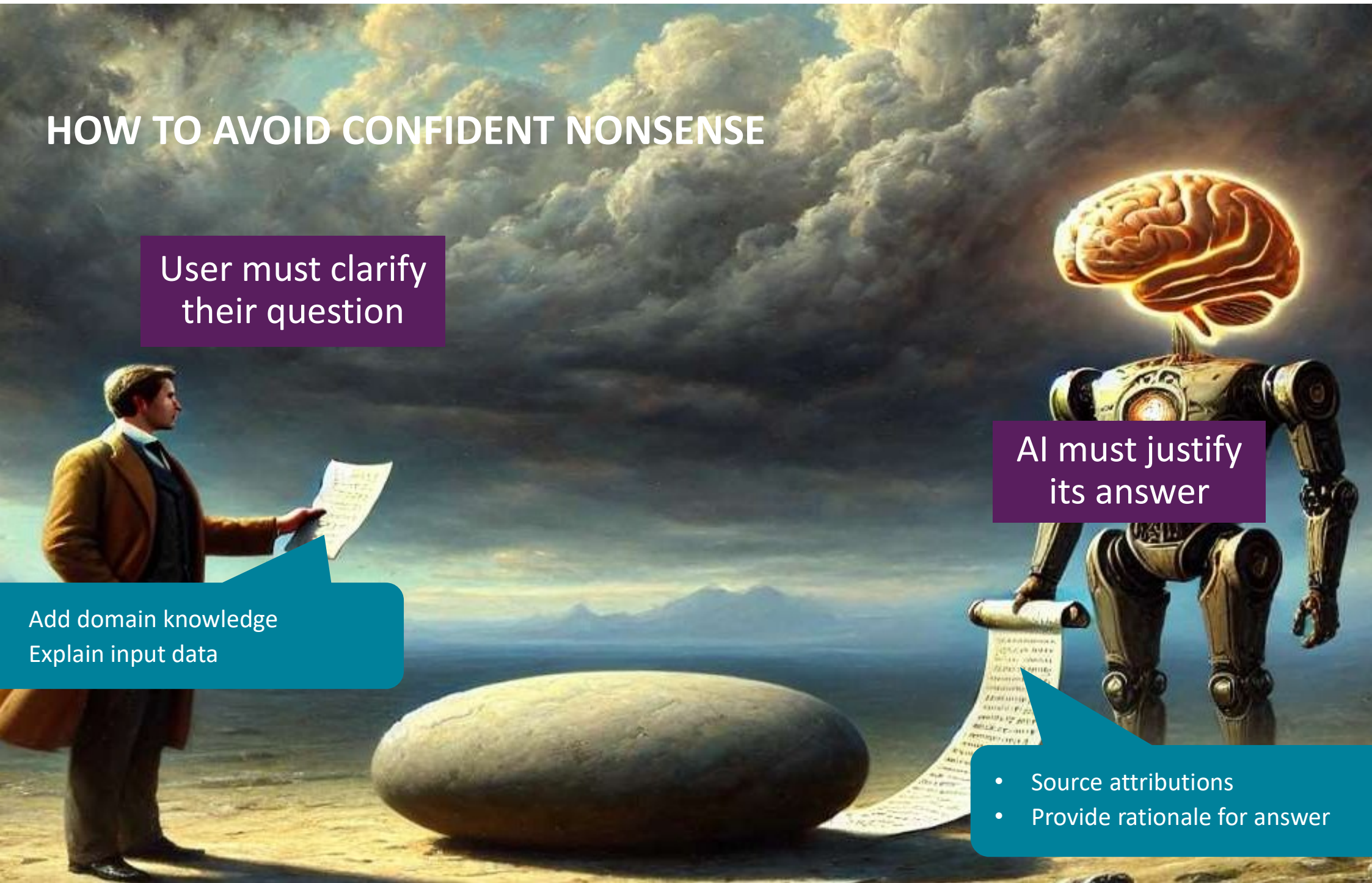
HOW TO AVOID CONFIDENT NONSENSE

User must clarify
their question

Add domain knowledge
Explain input data

AI must justify
its answer

- Source attributions
- Provide rationale for answer





User must clarify
their question

BUILDING TRUST

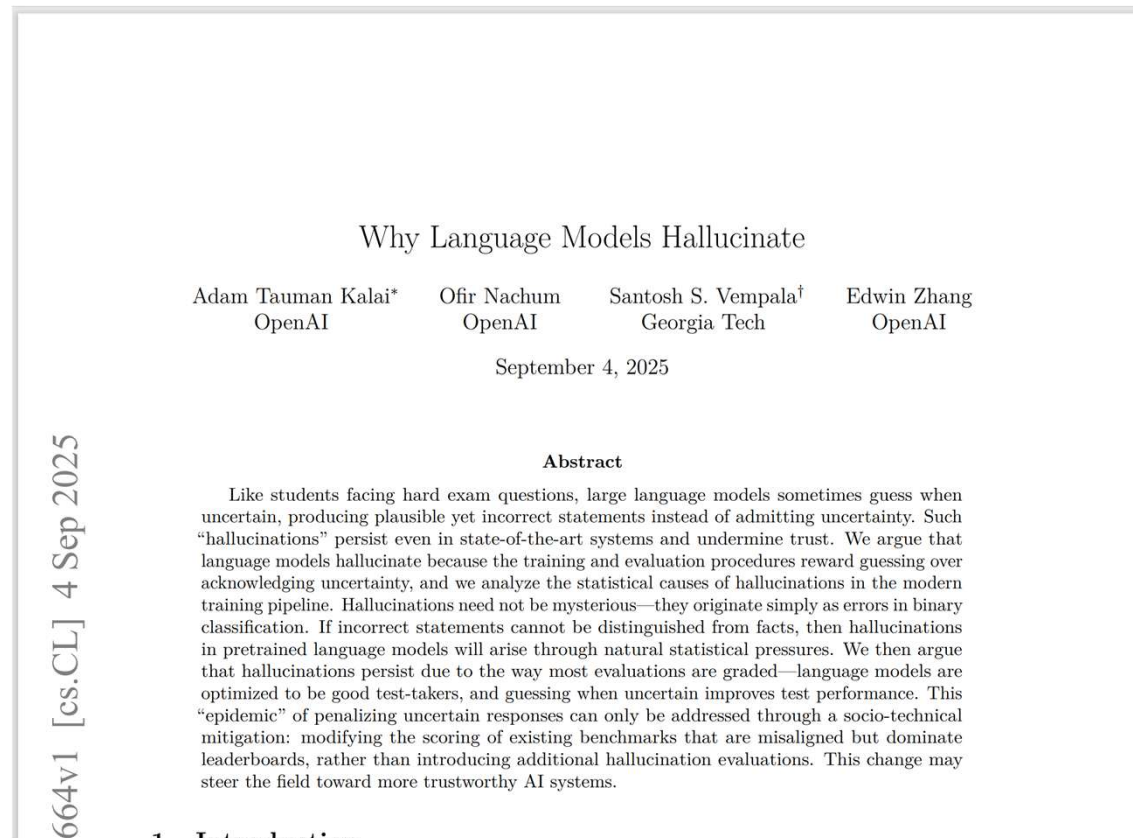
AI must justify
its answer

LOOKING FOR A MEASURE OF CONFIDENCE THAT MAKES SENSE

- Idea 1: Let the LLM to answer “DONTKNOW” in case it’s not really sure
- Idea 2: Ask the LLM how sure it is about its answer (low/medium/high confidence)

THE LLM WAS TAUGHT TO GUESS RATHER THAN TO ADMIT UNCERTAINTY

- Idea 1: Let the LLM to answer “UNKNOWN” in case it’s not really sure
- Idea 2: Ask the LLM how sure it is about its answer (low/medium/high confidence)



SO WE HOOKED UP THE LLM TO A POLYGRAPH

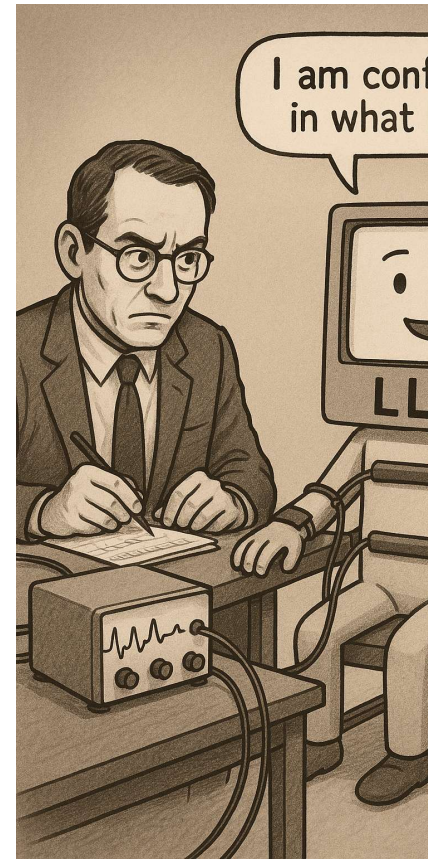
Idea 3: Measure the token-level probability of each answer

1.0 —
0.9 —
0.8 —
0.7 —
0.6 —
0.5 —
0.4 —
0.3 —
0.2 —
0.1 —
0.0 —

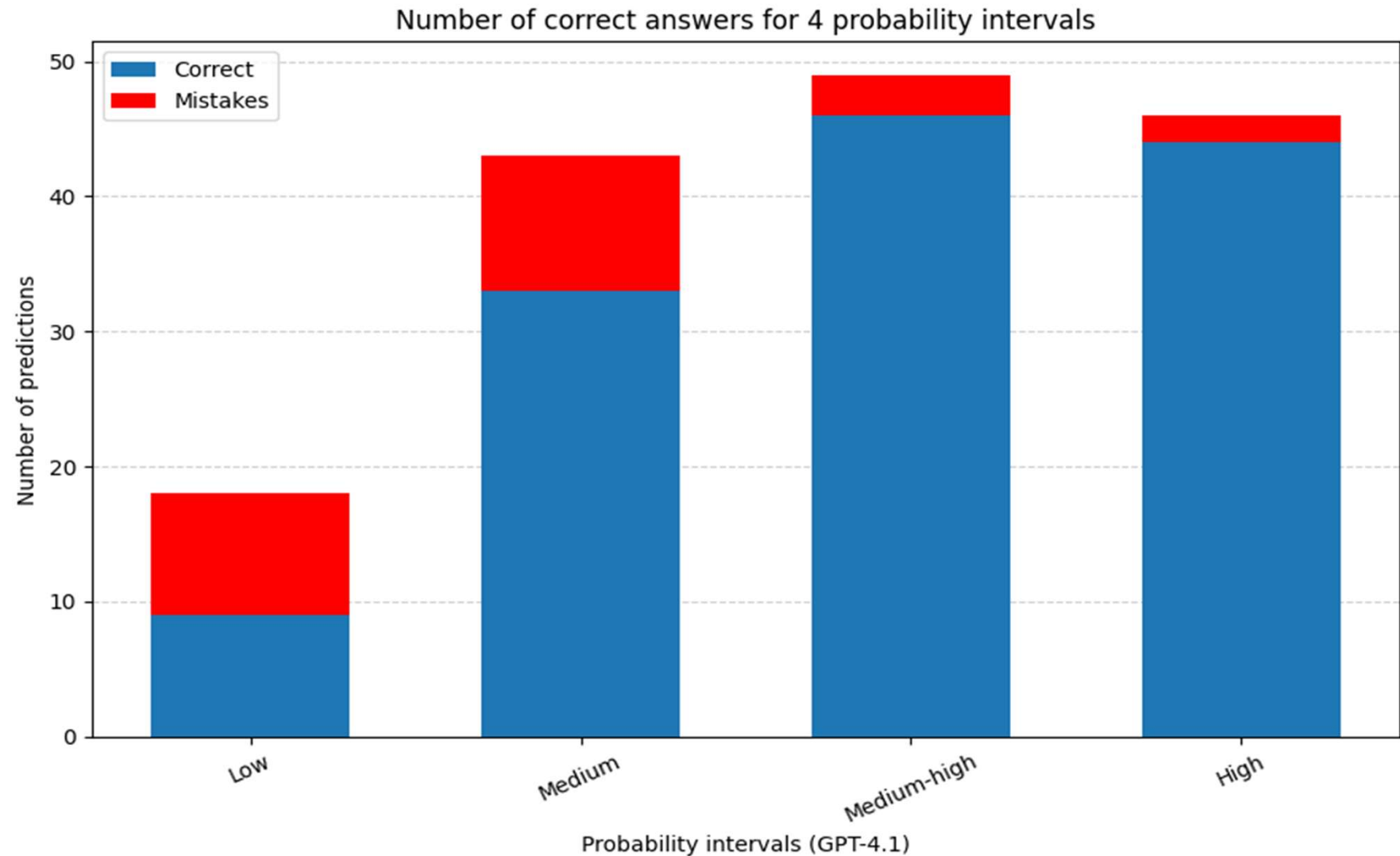
Accept answers that have these probabilities?
Which percentage of those is correct then?

What about these?

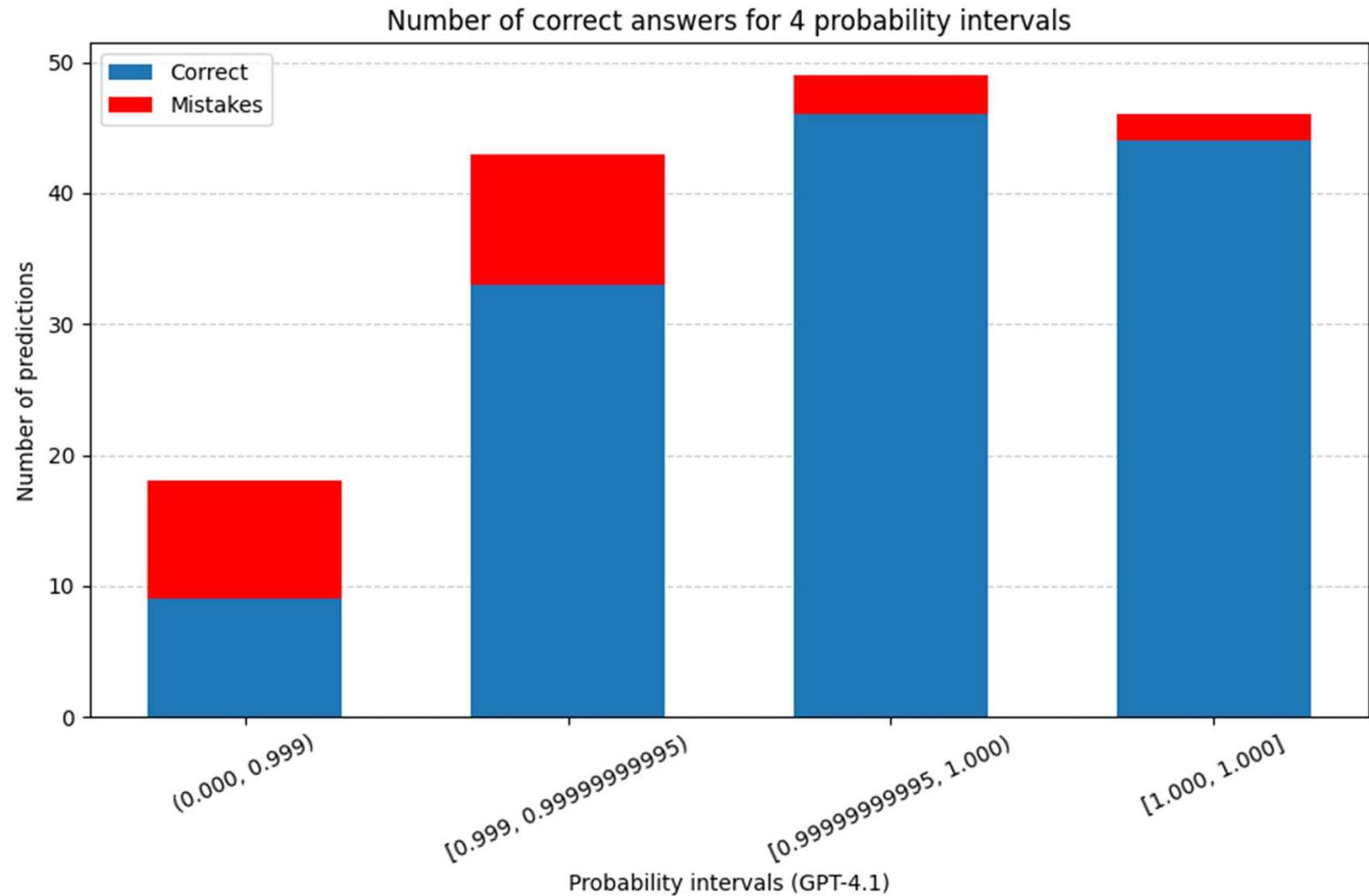
Reject these?
How many answers do we lose then?



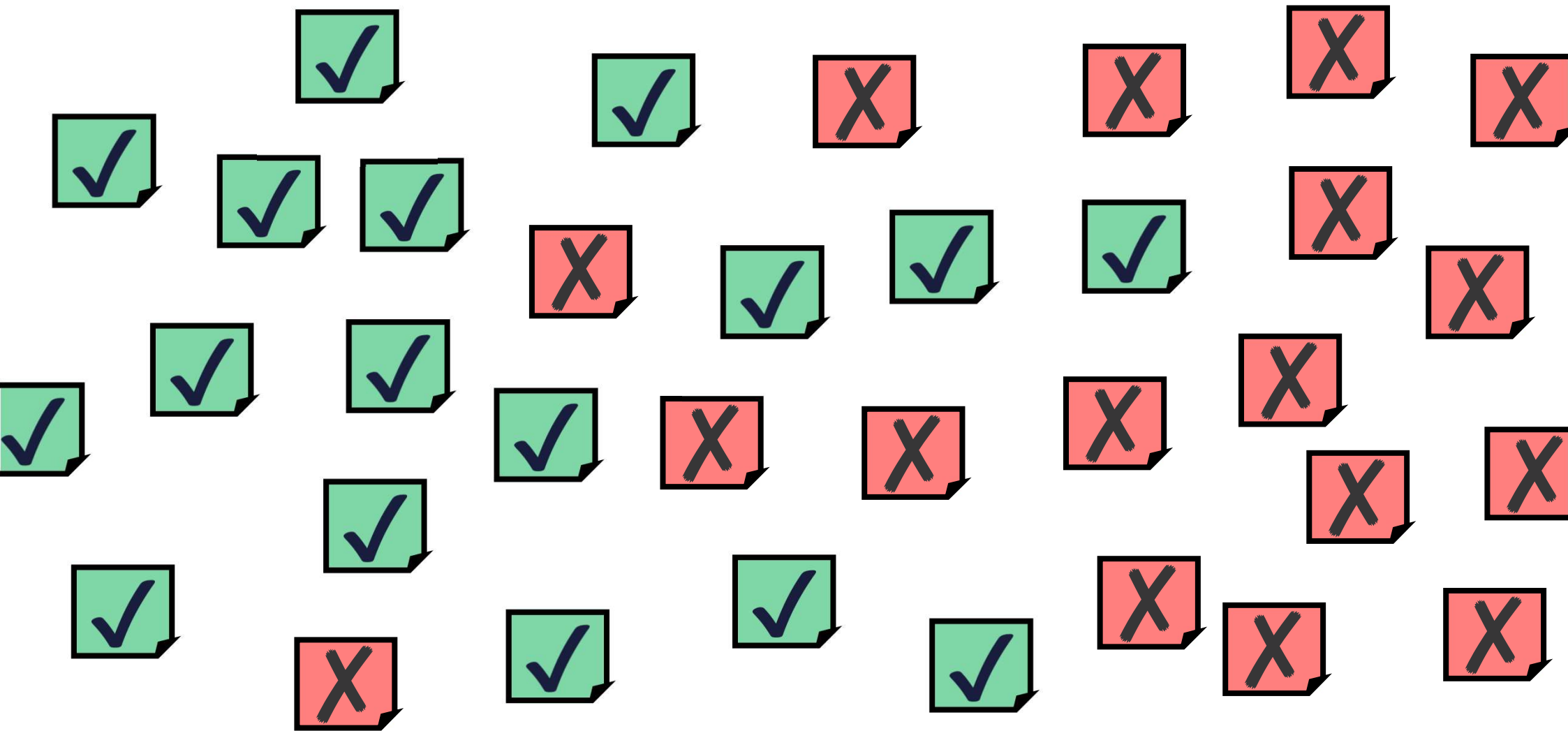
AT WHICH PROBABILITY VALUES CAN WE TRUST THE ANSWER?



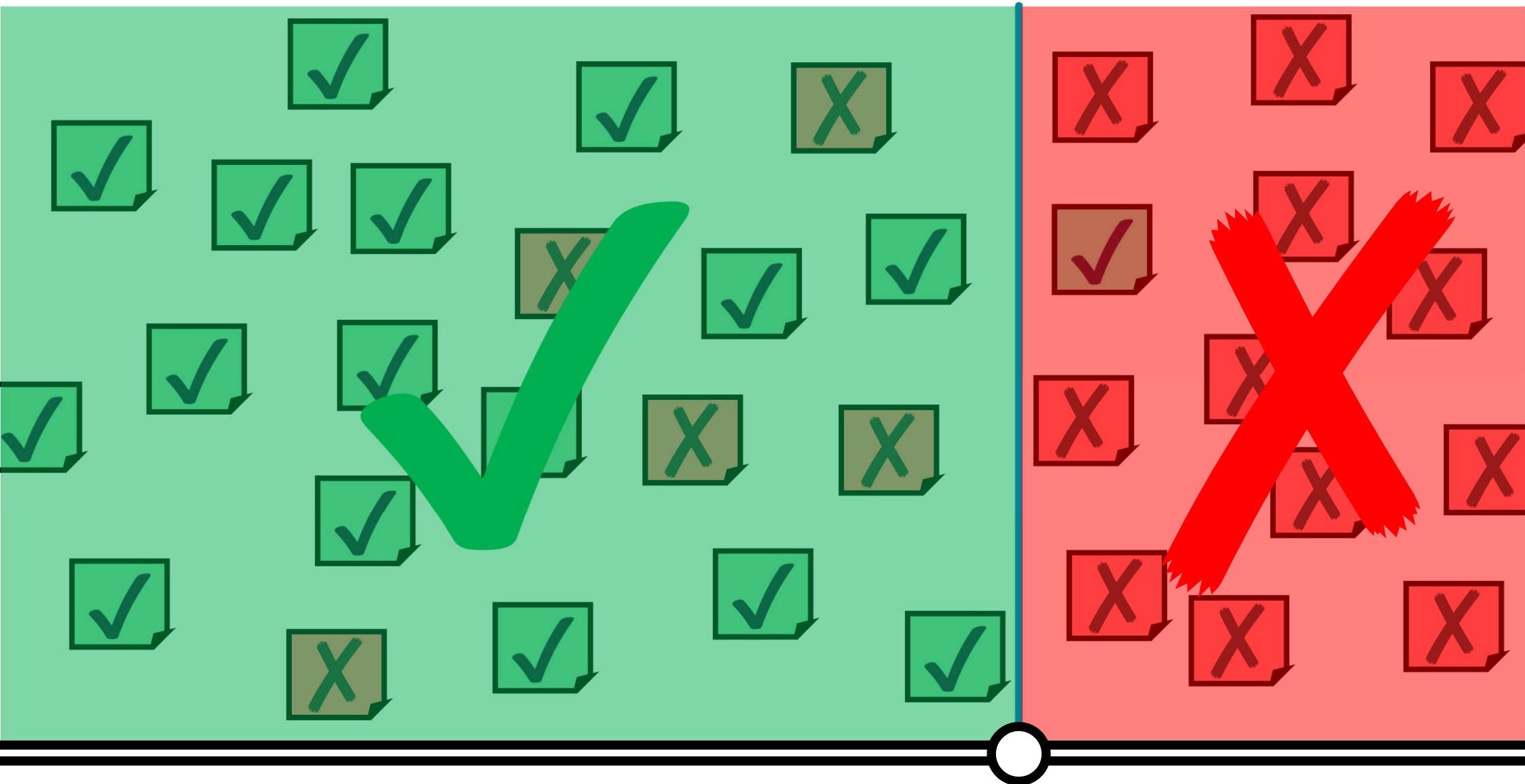
AT WHICH PROBABILITY VALUES CAN WE TRUST THE ANSWER?



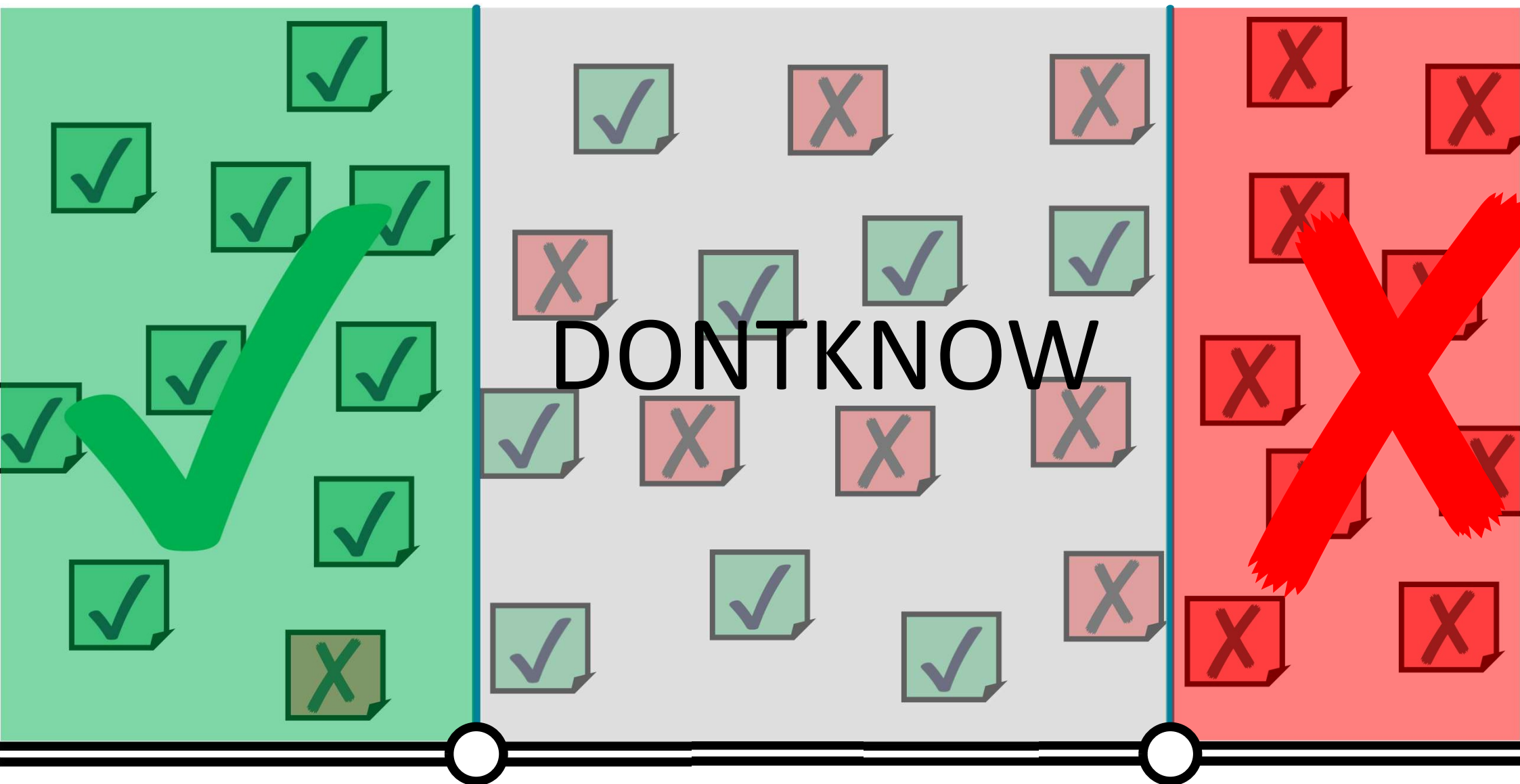
THE TEST SET (✓ = YES, X = NO)



YES/NO ASSESSMENT BY A SINGLE PROMPT



IDEA 4: CONFIDENCE LEVELS BY USING 2 CONSERVATIVE PROMPTS



HOW TO AVOID CONFIDENT NONSENSE

User must clarify
their question

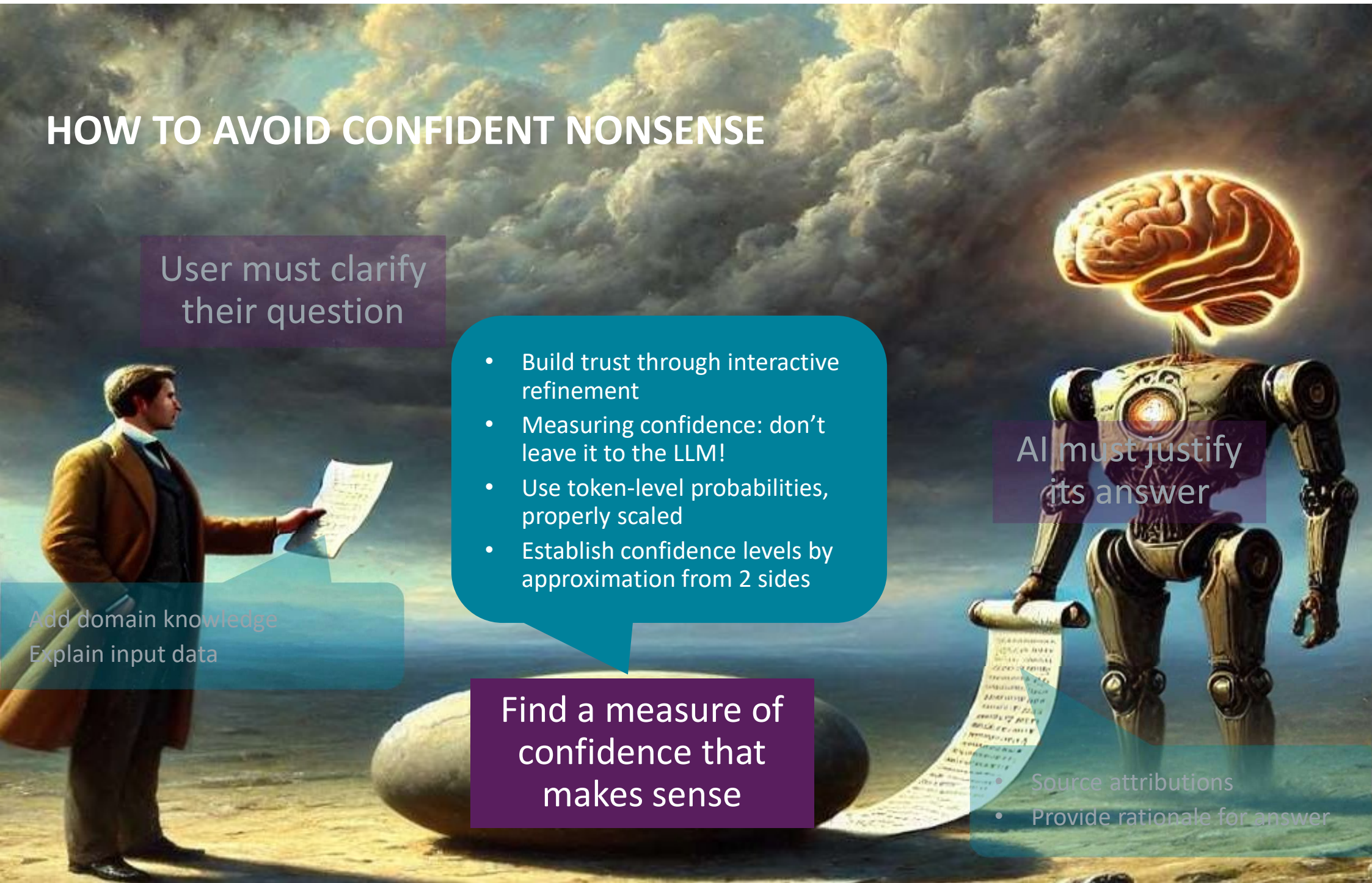
Add domain knowledge
Explain input data

- Build trust through interactive refinement
- Measuring confidence: don't leave it to the LLM!
- Use token-level probabilities, properly scaled
- Establish confidence levels by approximation from 2 sides

Find a measure of
confidence that
makes sense

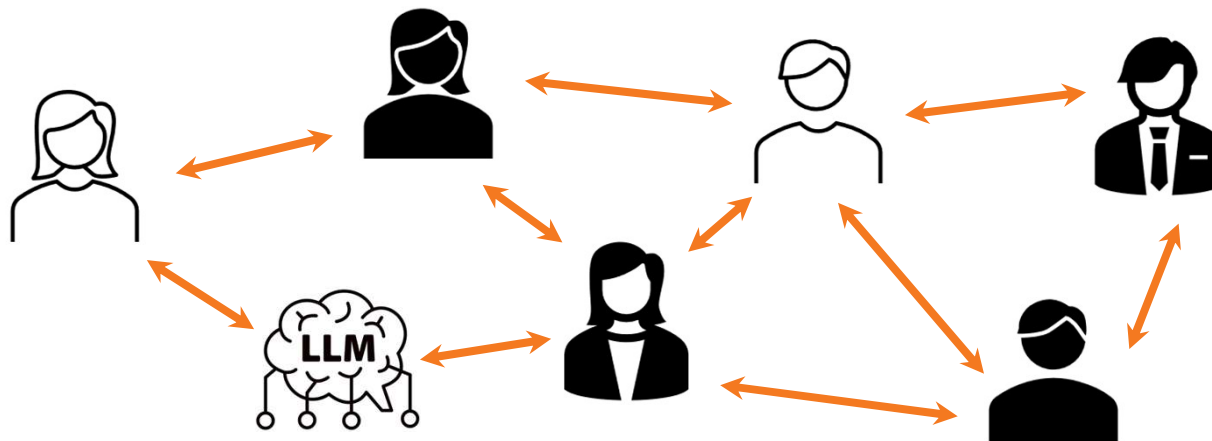
AI must justify
its answer

- Source attributions
- Provide rationale for answer



SO WHAT DID WE LEARN?

- LLM-based analysis of customer feedback reports
 - *reduces cost* – because fewer human specialists need to be called in
 - *increases consistency* – because an LLM can tirelessly repeat the same analysis over 1000s of reports
- Trustworthiness of LLM output can be brought to an acceptable level
 - by applying the suggestions from this presentation...
 - ... and comparing the LLM to what human experts can realistically achieve



AND WHAT'S NEXT FOR US?

- Support trust building via conversational Prompt Refinement Interface (demo in the innovation market!)
- Bring the current prototype to a larger user group for further feedback
- Turn our experimental platform to the level of a validated tool
- And add device logs into the mix!

Modify prompt template

Upload a document containing your initial prompt details and creation guidelines. Once uploaded, you will be able to customize it according to your needs.

Upload file with prompt details (txt, md, yaml) Prompt Difference

prompt.yaml 2.4 KB

Assessment selection

- ☒ No movement user command
- ☒ Problem with emergency stop
- ☒ Procedure aborted

+ Add Assessment Clear all

Reference document upload

Rewrite prompt with AI

Basic Instruction

I will give you the text of a complaint report regarding the Philips Allura Interventional X Ray system. You must carefully read the full report and answer several questions about it, which are listed below. For each of these questions, your answer must be either "NO" or "YES", accompanied with a rationale how you got to your answer. This is a very important task! Really take your time to think about your answers. Reason step by step.

Information about the system

The system has several kinds of control modules. I am concerned with 2 of them: the geometry module and the imaging module (aka 'iso gen', 'iso imaging').

One control on the geometry module deserves special attention: the 'float tabletop' control.

Some controls on the control modules (both the geometry module and the imaging module) consist of plastic covers that are attached onto a metal pin. If the plastic cover is damaged, broken or removed, it may be that the metal pin can still be used to operate the system. This is an important point to keep in mind.

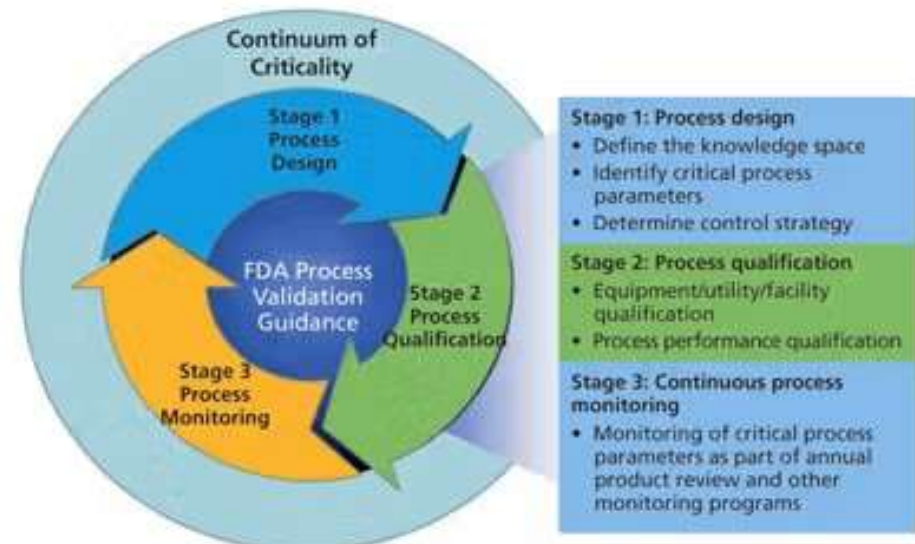
Questions to be answered

Q1: Is it the case that the user issues a command that is supposed to result in the movement of some part of the system, but yet no movement is observed?
Answer this question with "NO" or "YES".

Q2: Does the reported problem involve an emergency stop (also called E stop or e stop)?
Answer this question with "NO" or "YES".

Q3: In case there was a clinical procedure ongoing, did it have to be aborted, or was the patient moved to a different system in order to complete the procedure?

Export prompt



An X-ray image of two hands positioned to form a heart shape. The fingers are spread out, and the thumbs are pressed together at the base, creating a silhouette of a heart. The bones of the hands and wrists are clearly visible against a dark background.

THANK YOU!

