NIST INFORMATION TECHNOLOGY LABORATORY

Building Measurement Probes into Agentic AI Ecosystems

About the ITL AI Program

ITL's AI North Star:

To strengthen trust in AI, accelerate its adoption, and expand U.S. AI dominance by providing the vital measurement science, testing and evaluation, guidance, and standards.

Impact Areas

#1 ADVANCING TESTING, EVALUATION, VERIFICATION, AND VALIDATION (TEVV) FOR TRUSTWORTHY AI
Transforming the measurement of AI – especially system trustworthiness – via TEVV to ensure that AI is deployed and used responsibly, reliably, and efficiently

#2 EMPOWERING INFORMED DECISION-MAKING
Providing resources for managing AI benefits and risks, empowering industry, government to make informed decisions about AI trustworthiness and use

#3 GLOBAL STANDARDS LEADERSHIP
Positioning U.S. as preeminent in AI technical and governance standards – ensuring U.S. leads global AI innovation, combats adversaries' growing influence

#4 DOMAIN-SPECIFIC ADVANCES
Enabling U.S. to lead in applying AI to high-priority area – including manufacturing and cybersecurity for critical infrastructure – via innovative approaches to address measurement challenges

NIST Strategy for American Technology Leadership in the 21st Century

Aligning with president's science and technology agenda, NIST will focus on:

1. Accelerating the buildout and scale-up of the U.S. quantum industrial base
2. **Solidifying American dominance in AI innovation**
3. Harnessing the power of biotechnology
4. Growing U.S. leadership in semiconductors

NIST champions the U.S. industry-led, market-driven, and voluntary approach to international standards development.

AI Innovation for the 21st Century

NIST will catalyze American AI innovation by partnering with industry to accelerate development and adoption of AI systems and applications, including:

- AI-driven autonomous agents to **increase U.S. manufacturing productivity.**
- AI-based agents to **protect/secure U.S. critical infrastructure from cyberthreats.**
- Consistency in **measurement of AI system performance, reliability, and security**.
- U.S. capacity to **rapidly evaluate** AI system capabilities.

nist.gov/director/strategic-priorities

NIST Efforts to Advance Agentic AI

NIST is expanding its agentic AI efforts to meet U.S. needs and catalyze American innovation.

Examples of ongoing agentic AI-related efforts:

- Launching the NIST AI Agent Standards Initiative: NIST Center for AI Standards and Innovation (CAISI) and ITL
- Expanding and customizing **existing cybersecurity and other frameworks and guidance** – including profiles, taxonomies, overlays, and other resources – to specifically incorporate agentic AI considerations
- Exploring approaches to **identify, manage, and authorize access and actions** taken by software agents, including AI agents – and provide practical guidelines to securely implement AI agents
- Engaging in U.S. and international standards activities related to agentic AI
- Advancing research related to measurement and evaluation of agentic AI

Contact Us:

Scan the code to subscribe for AI-related updates from NIST's Information Technology Laboratory (ITL)

Or email us: itl-ai-program@nist.gov

Next webinar: Details Coming Soon!

Just released: Concept Note on Artificial Intelligence Risk Management Framework Profile on Trustworthy AI in Critical Infrastructure

Building Measurement Probes into Agentic AI Ecosystems

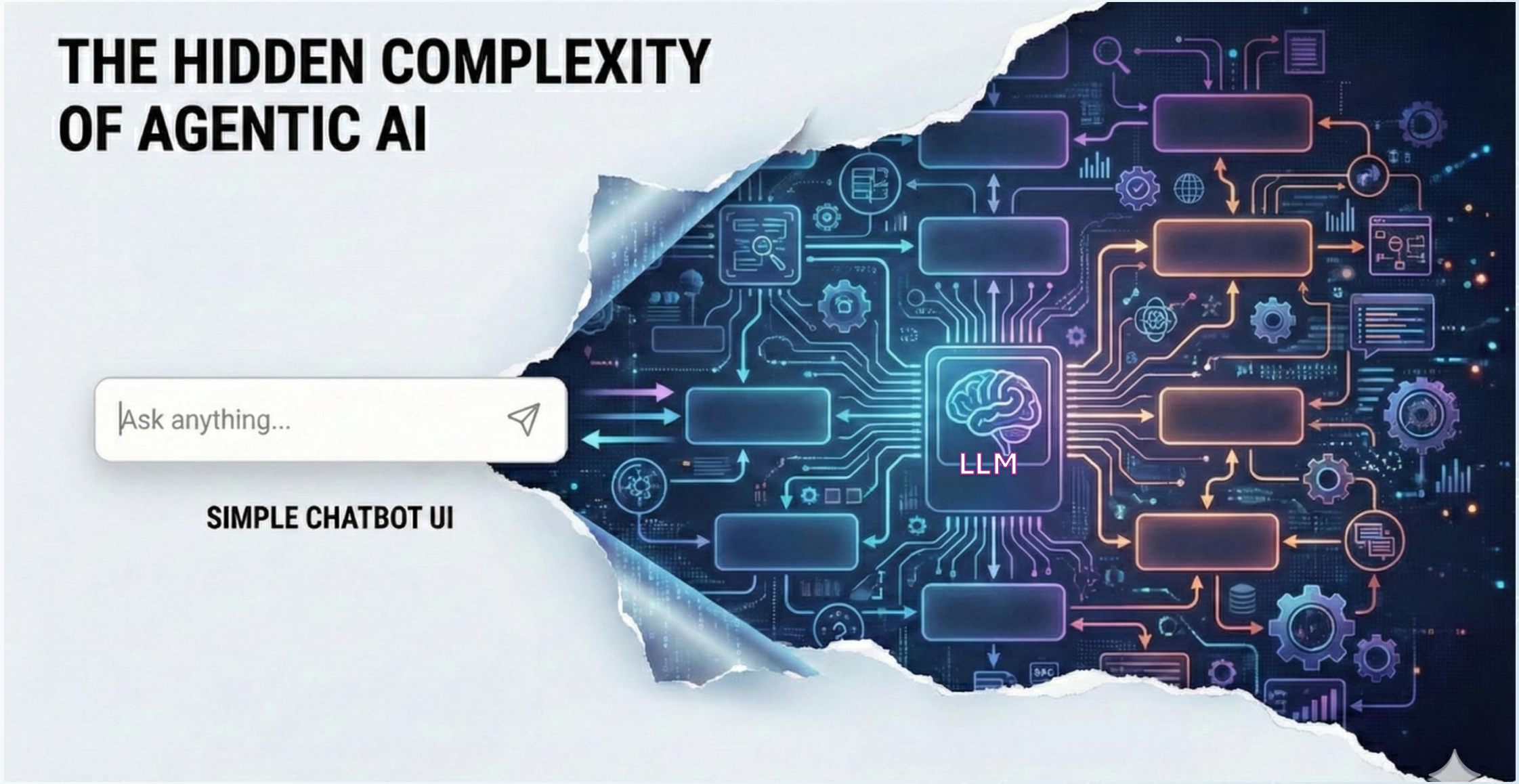
NIST INFORMATION TECHNOLOGY LABORATORY

NIST Disclaimer

- Certain commercial entities, equipment, or materials may be identified during this presentation to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

- All figures and images shown during this presentation were created and edited using Gemini (developed by Google) in accordance with author instructions. All content has been reviewed and verified.

THE HIDDEN COMPLEXITY OF AGENTIC AI



SIMPLE CHATBOT UI

Agenda

The Agentic AI Stack

Deep Research as a Case Study

Measurement Probes

Agentic Measurement Deep Dive

Q&A

Agenda

The Agentic AI Stack

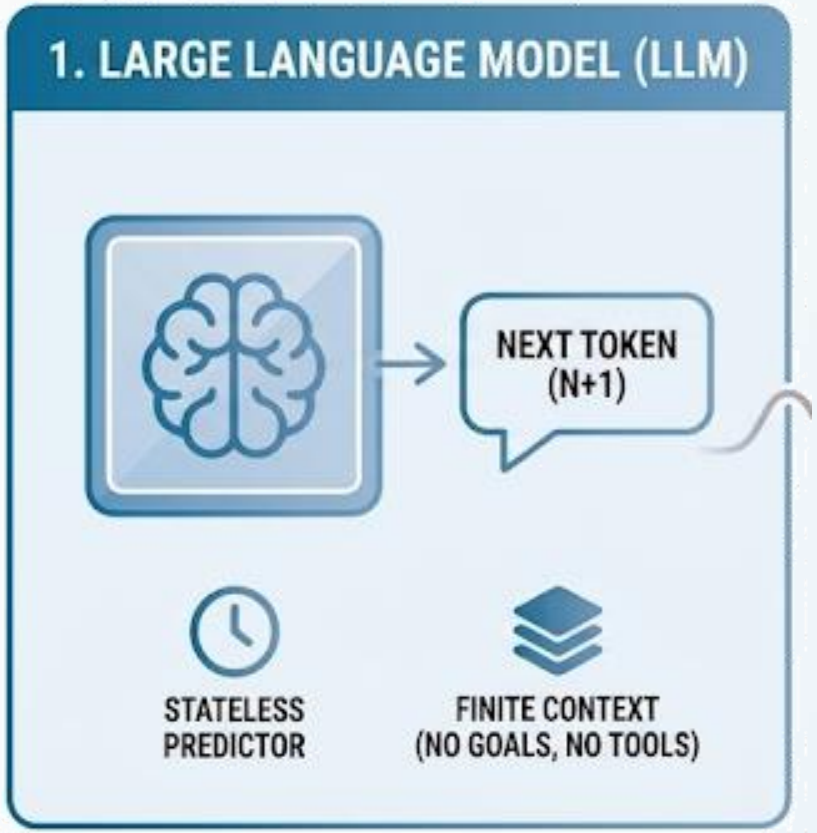
Deep Research as a Case Study

Measurement Probes

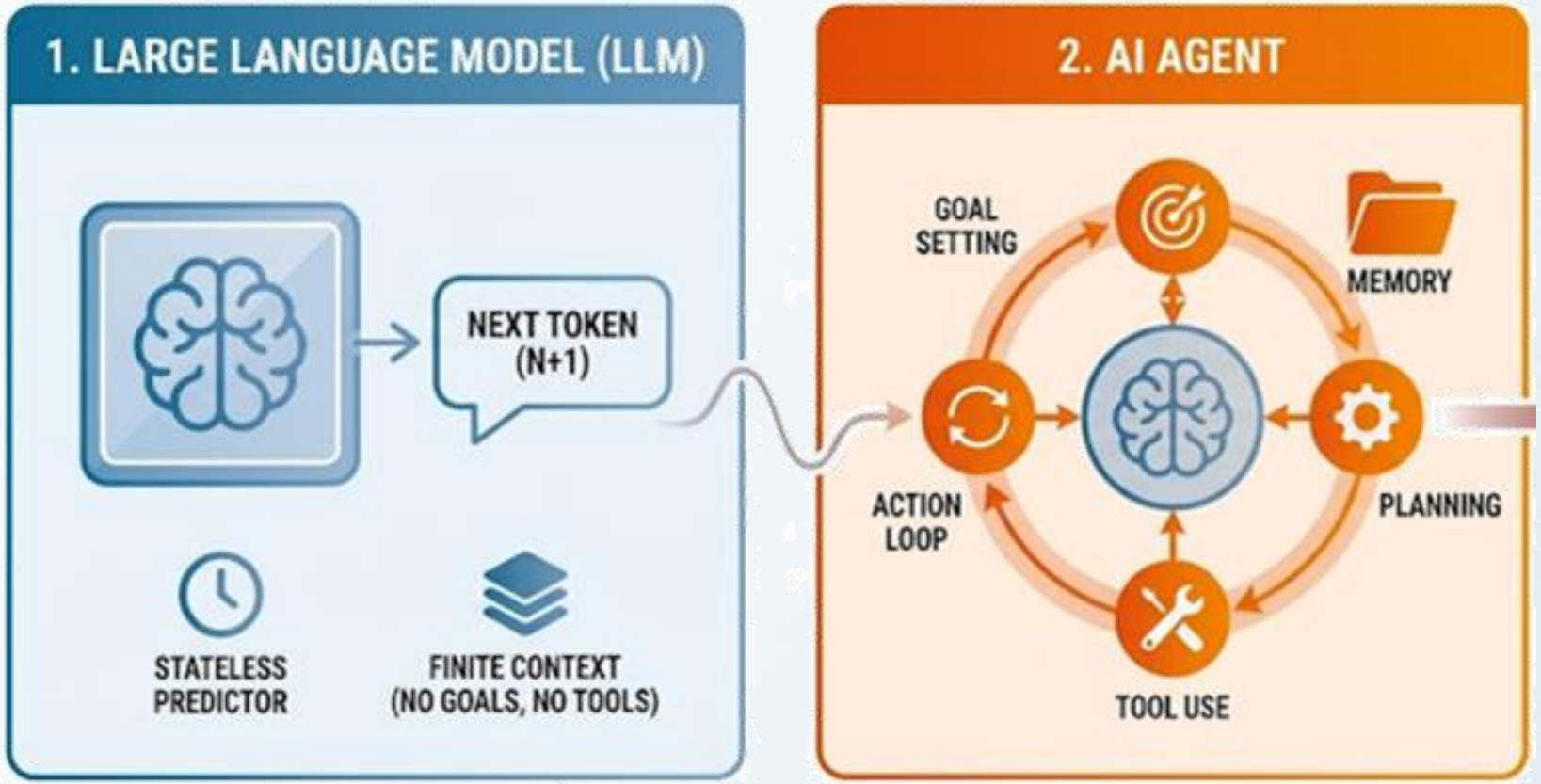
Agentic Measurement Deep Dive

Q&A

What are we exploring?
From Predictors to Agentic Systems



What are we exploring?
From Predictors to Agentic Systems



What are we exploring?
From Predictors to Agentic Systems

NIST ITL's AI Measurement Role

#1 ADVANCING TESTING, EVALUATION, VERIFICATION, AND VALIDATION (TEVV) FOR TRUSTWORTHY AI

Transforming the measurement of AI – especially system trustworthiness – via TEVV to ensure that AI is deployed and used responsibly, reliably, and efficiently

#2 EMPOWERING INFORMED DECISION-MAKING

Providing resources for managing AI benefits and risks, empowering industry, government to make informed decisions about AI trustworthiness and use

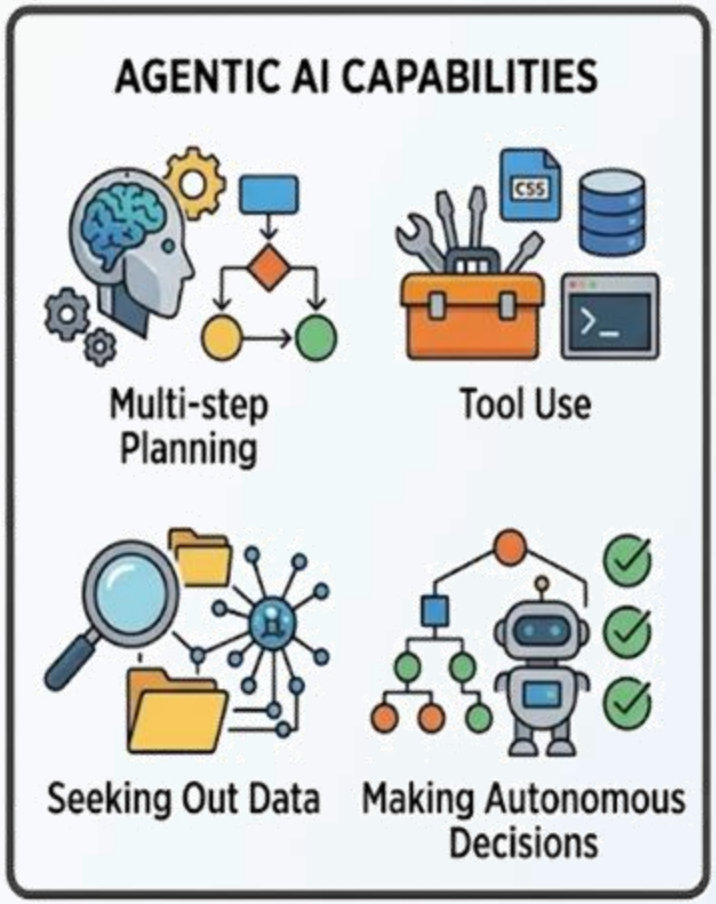
#3 APPLYING AI TO HIGH PRIORITY DOMAINS

Enabling the U.S. to be at the forefront in applying AI to high-priority domains by developing innovative approaches to address scientific, technical, and measurement challenges.

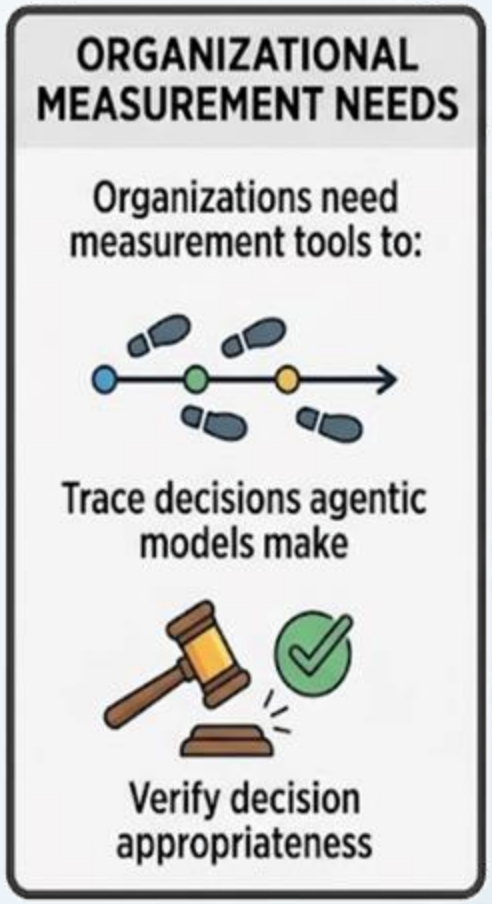
Shared Challenge, Collaborative Solution

- NIST is part of a community working to measure agentic AI
  - Open-source measurement development
- We're actively seeking input about your practical measurement challenges and domain-specific use cases
  - Collaboratively build a robust, shared foundation for measuring agentic-AI
- **What are Your Agentic-AI measurement challenges?**

Agentic AI - Measurement Probes

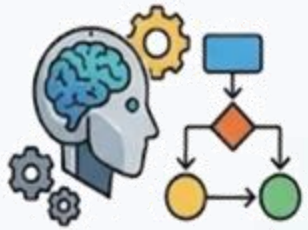


+



Agentic AI - Measurement Probes

AGENTIC AI CAPABILITIES



Multi-step Planning

Tool Use

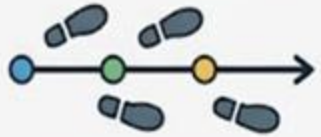
Seeking Out Data

Making Autonomous Decisions

+

ORGANIZATIONAL MEASUREMENT NEEDS

Organizations need measurement tools to:



Trace decisions agentic models make

Verify decision appropriateness

TRACEABILITY DEPENDS UPON RECONSTRUCTING...

Chain of Reasoning

Usage of Tools

Evidence Behind Decisions

What are Measurement Probes

What they are:

- Automated tools acting as “adversarial verifiers” that stress-test AI outputs against trusted information

How they work:

- Isolates every claim the AI makes and verifies against the cited source material
- **Powered by LM-Judge:** Uses strict, rubric-based schemas to evaluate specific criteria
- Generates structured feedback the AI can use to autonomously self-correct before finalizing the output

How we insert them:

- Injected into the workflow and run automatically at inference, when the AI is generating output

Impact - Why we want probes

Unlocking High Stakes Applications

- Turns prohibitive risks (like hallucinations) into manageable, measurable properties
- Empowers enterprise to safely scale AI workloads through constraint adherence

Measurement Science Contribution

- Providing the foundational measurement infrastructure needed to build justified trust and accelerate U.S. AI innovation

Delivering the Auditable “Why”

- Transforms "the AI said so" into "here is what the AI found, where it found it, and how it justified the conclusions"

Agenda

The Agentic AI Stack

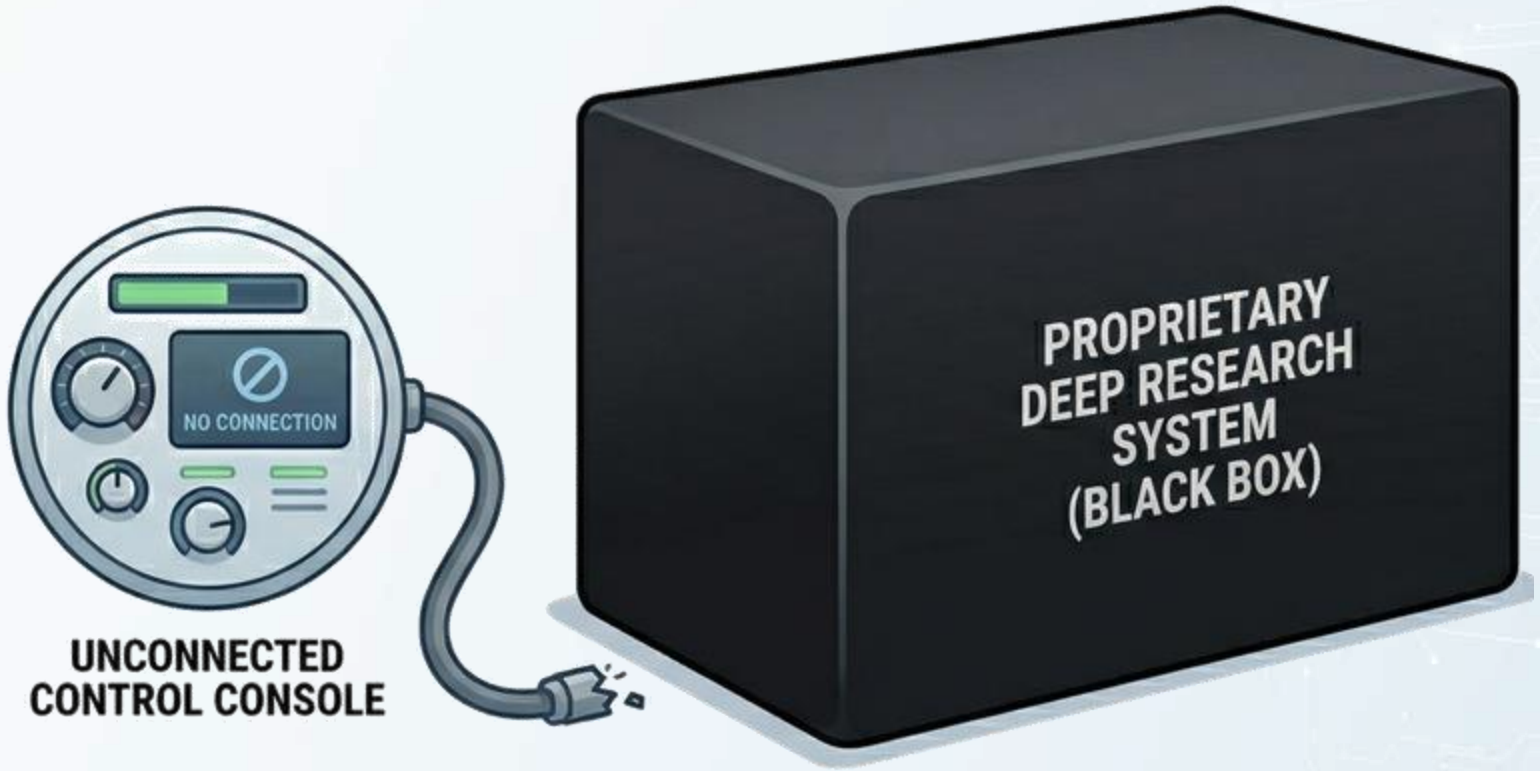
Deep Research as a Case Study

Measurement Probes

Agentic Measurement Deep Dive

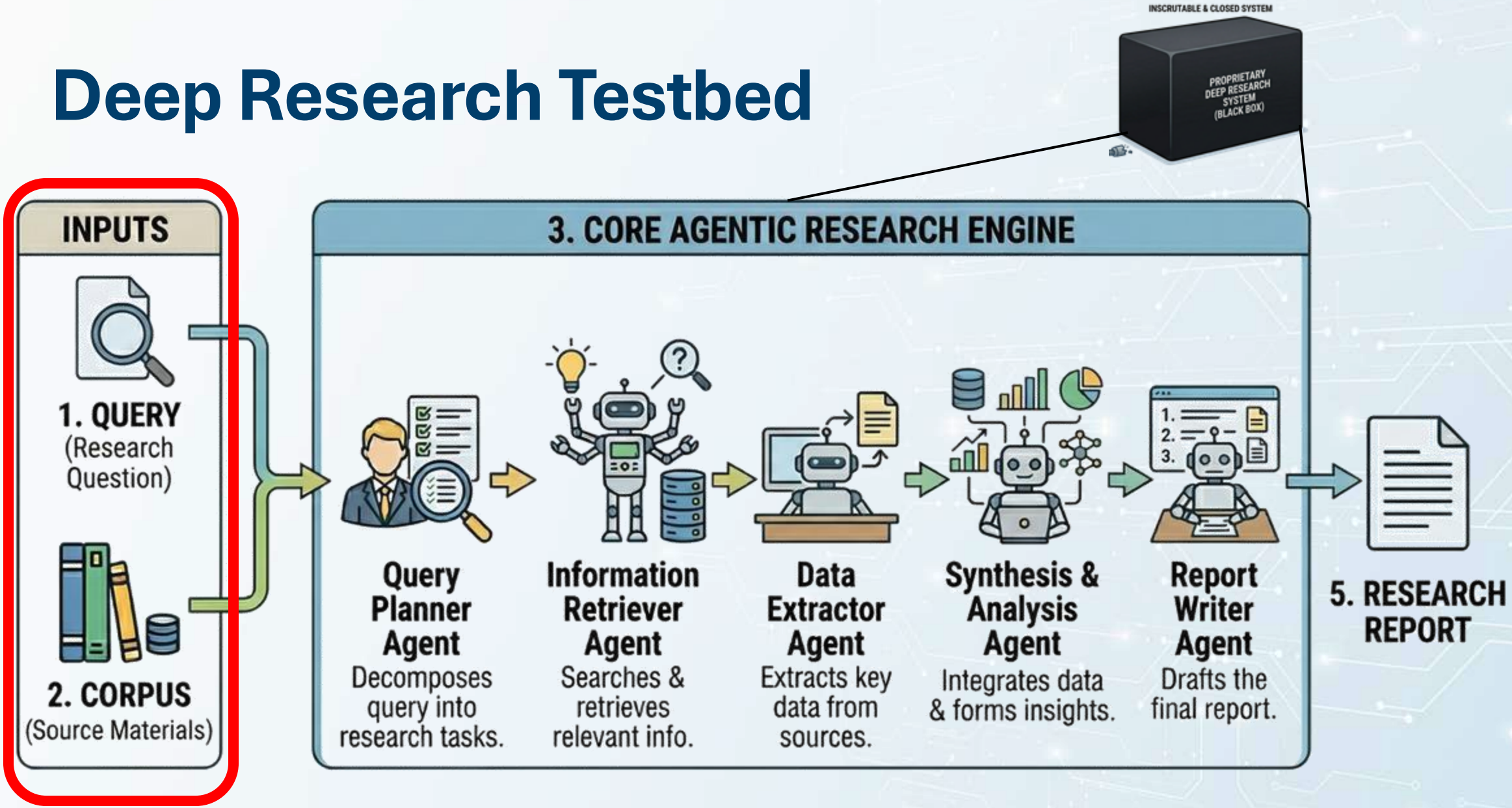
Q&A

Agentic AI - Testbed/Exemplar

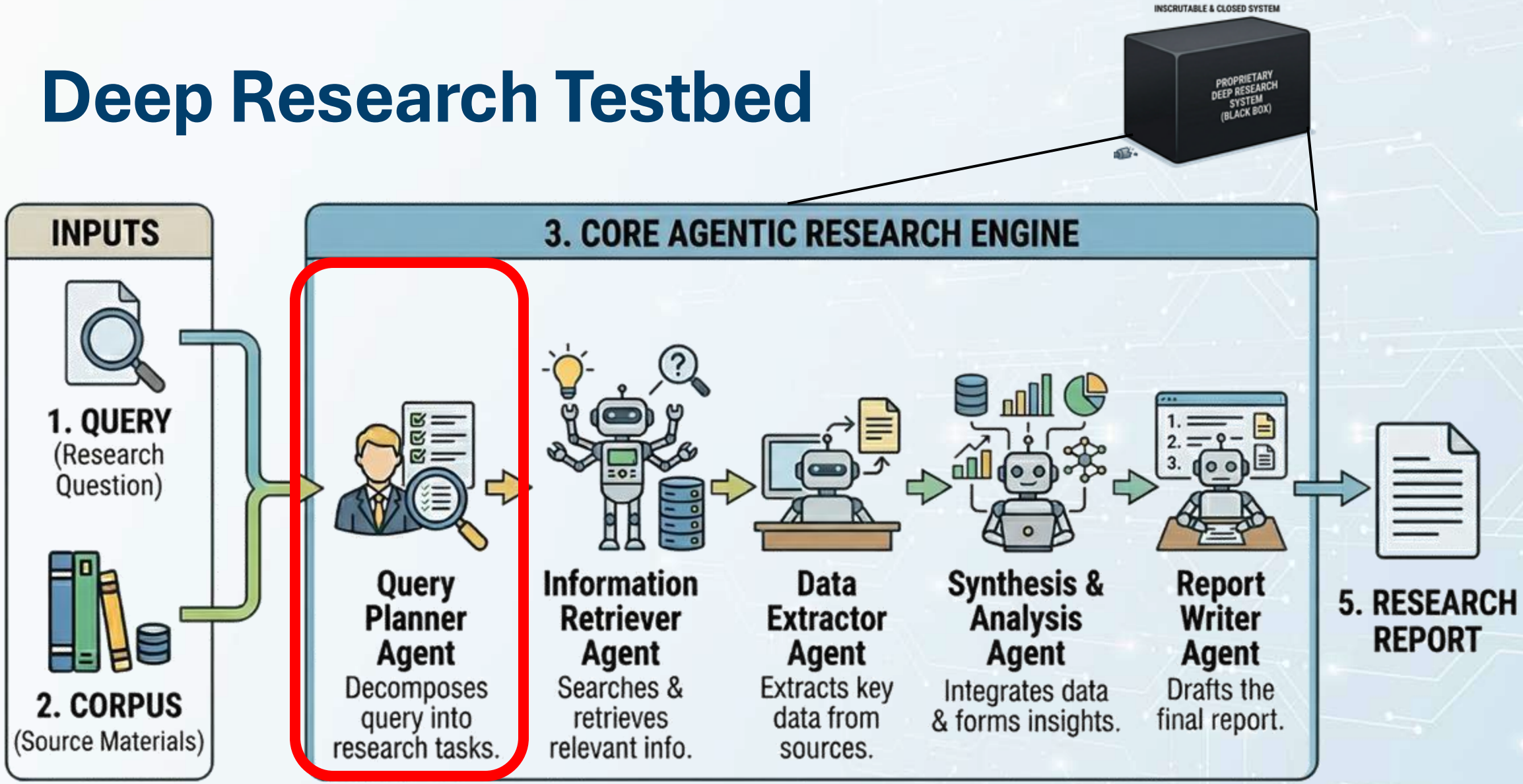


Measurement Toolkit

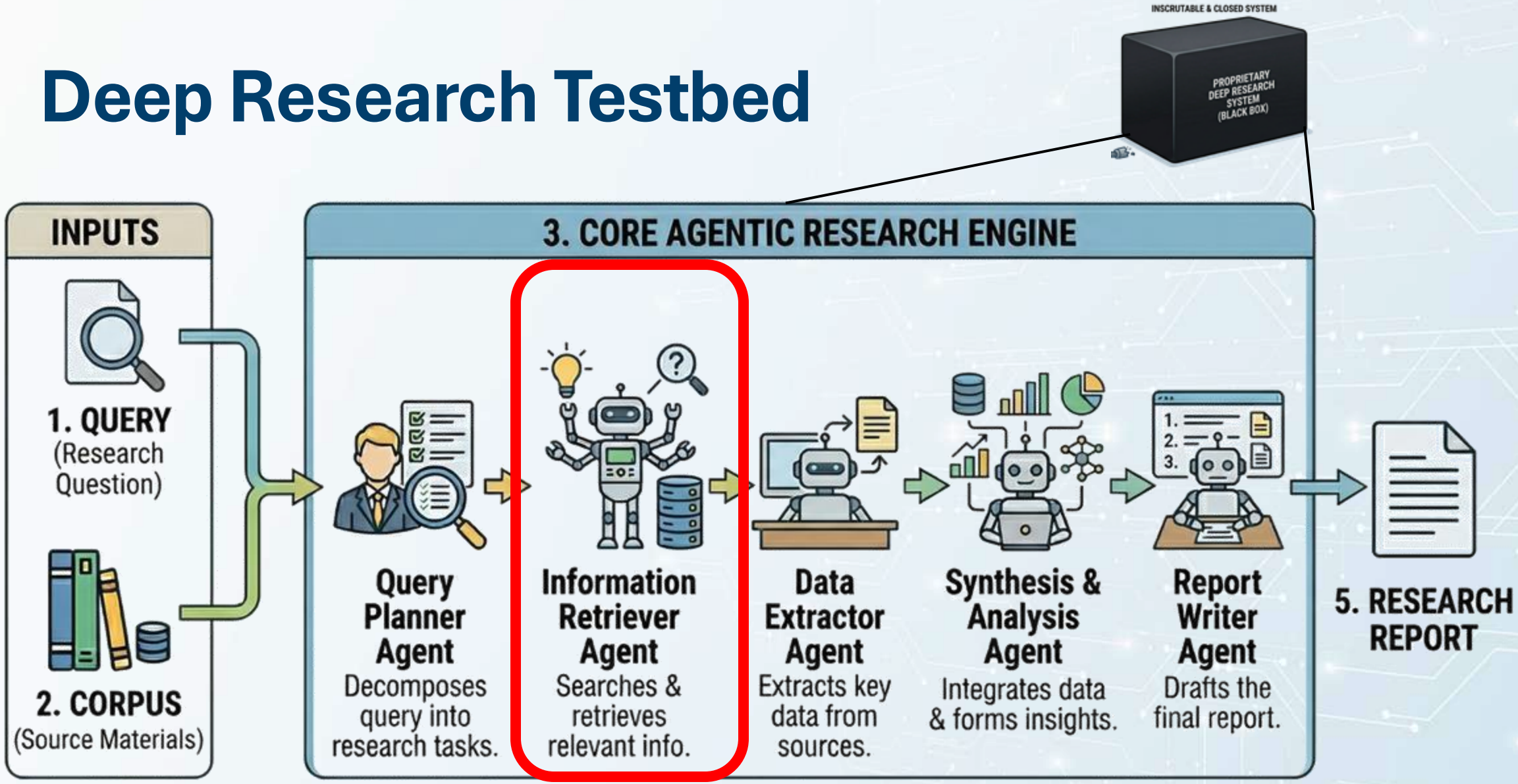
Deep Research Testbed



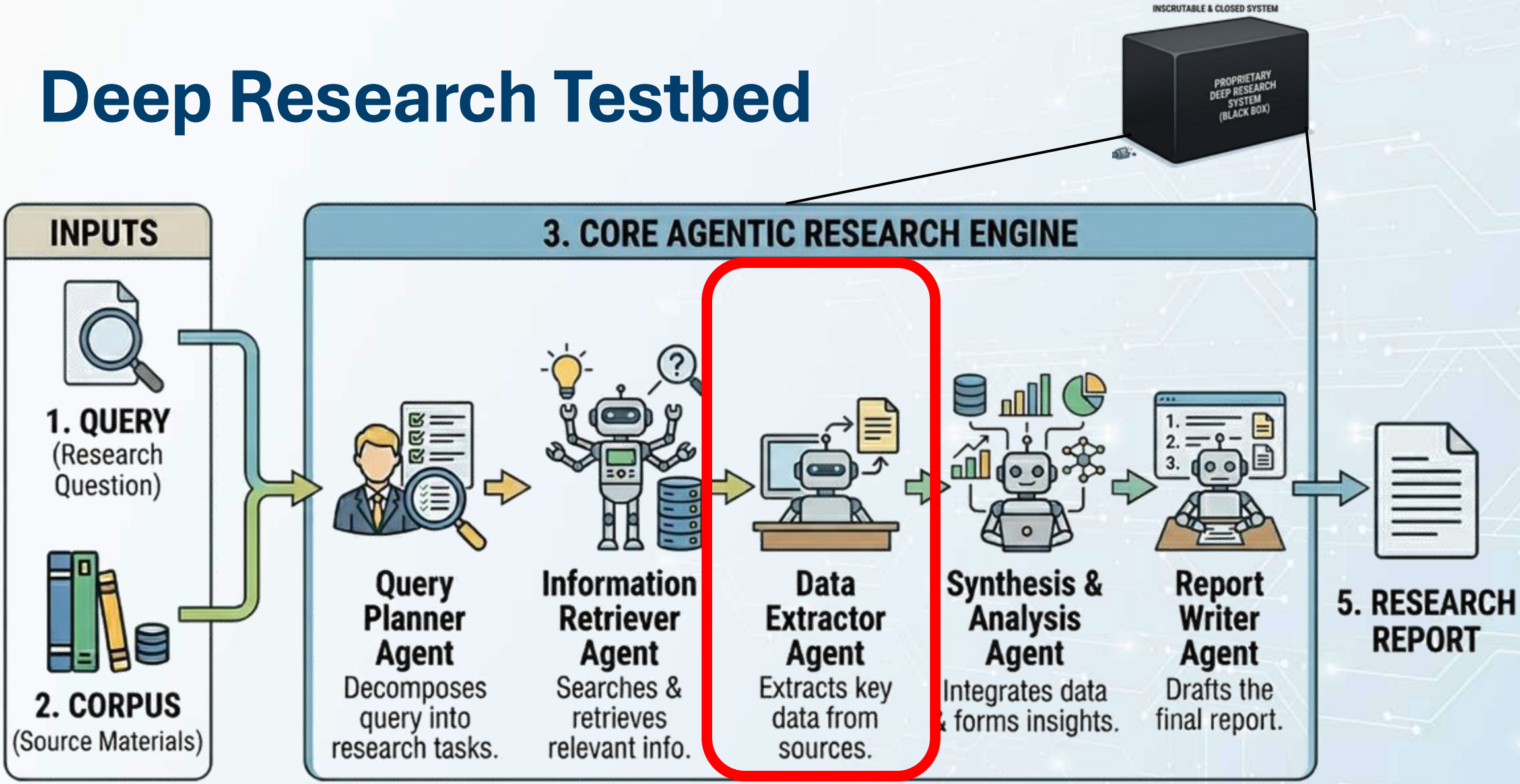
Deep Research Testbed



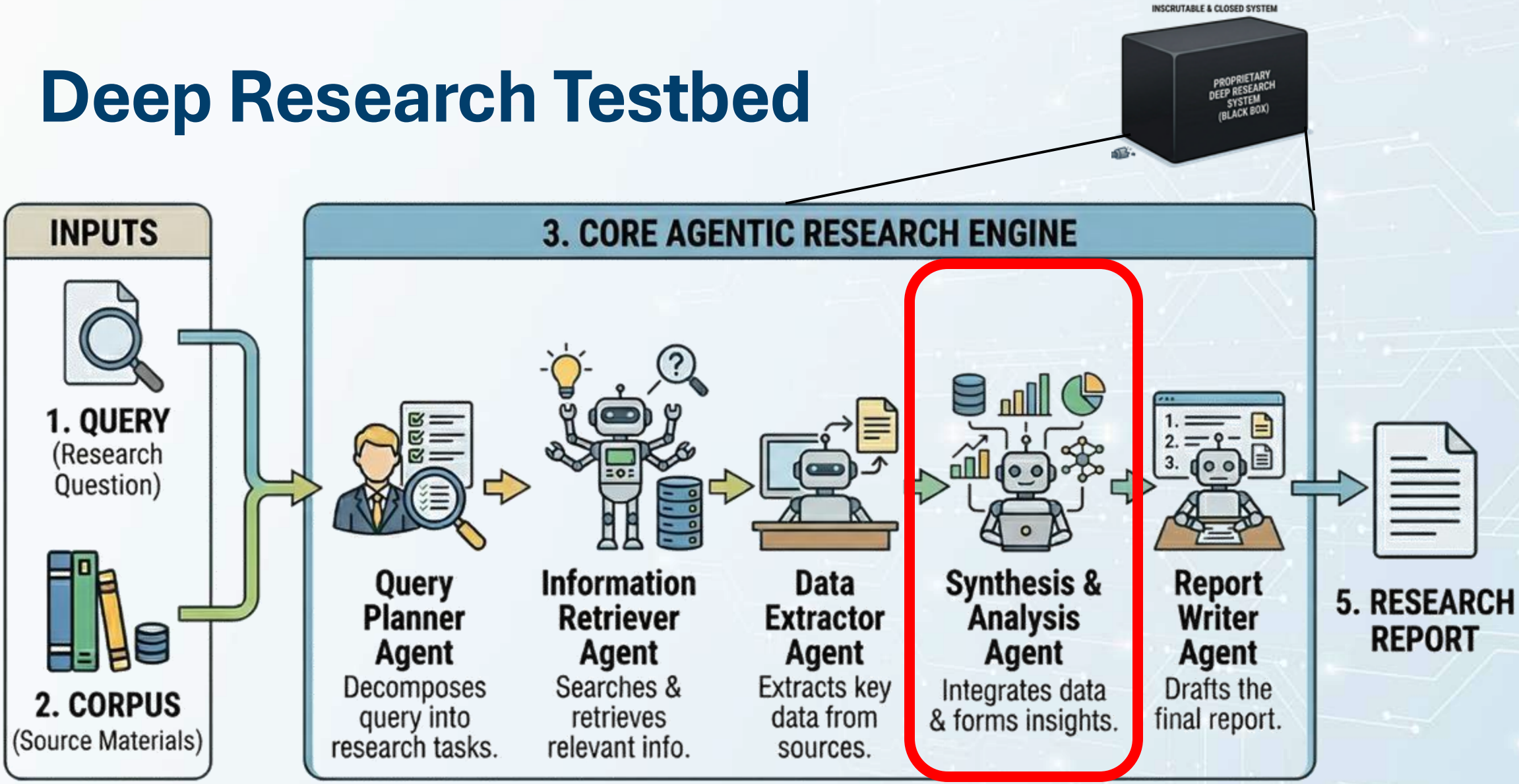
Deep Research Testbed



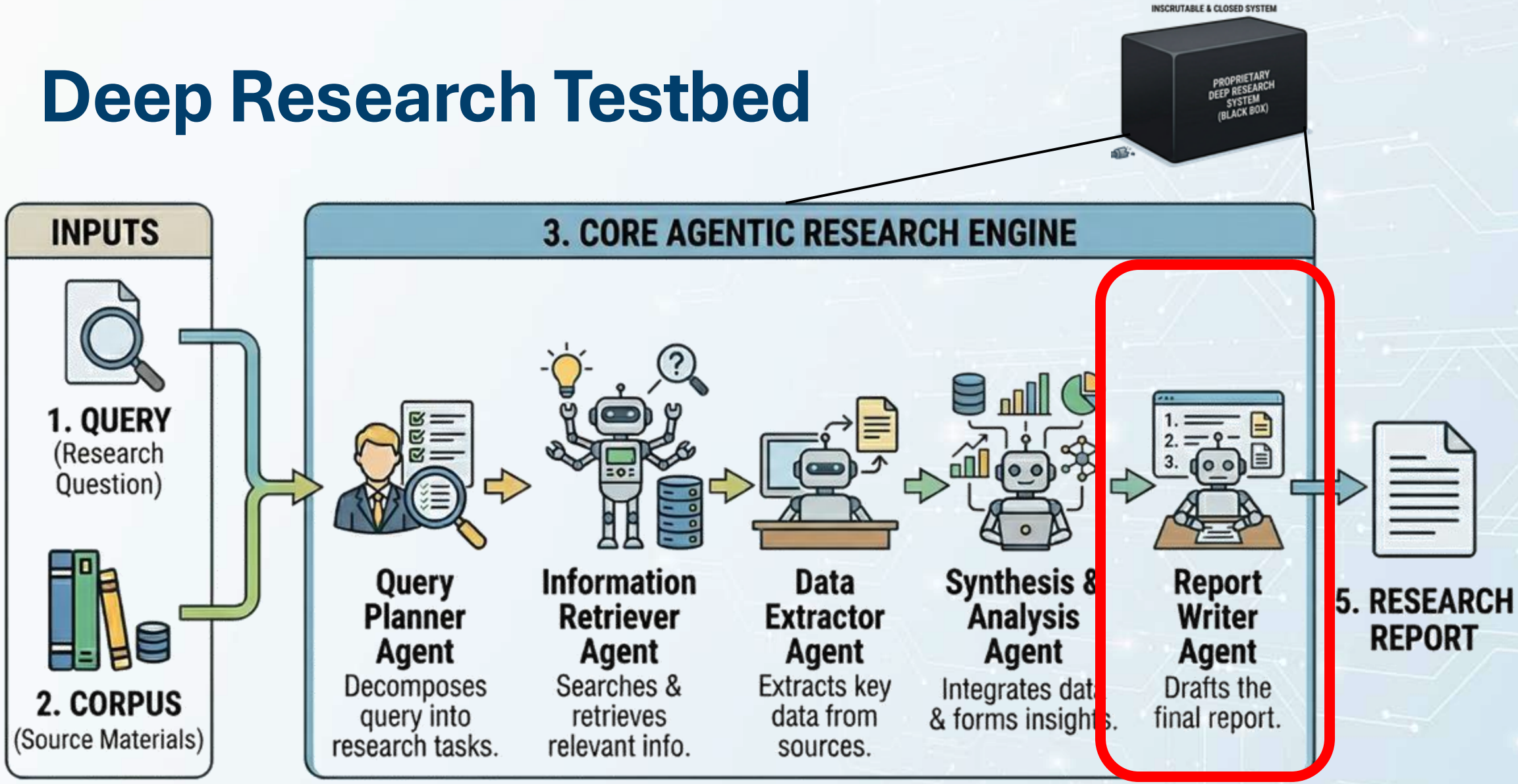
Deep Research Testbed



Deep Research Testbed



Deep Research Testbed

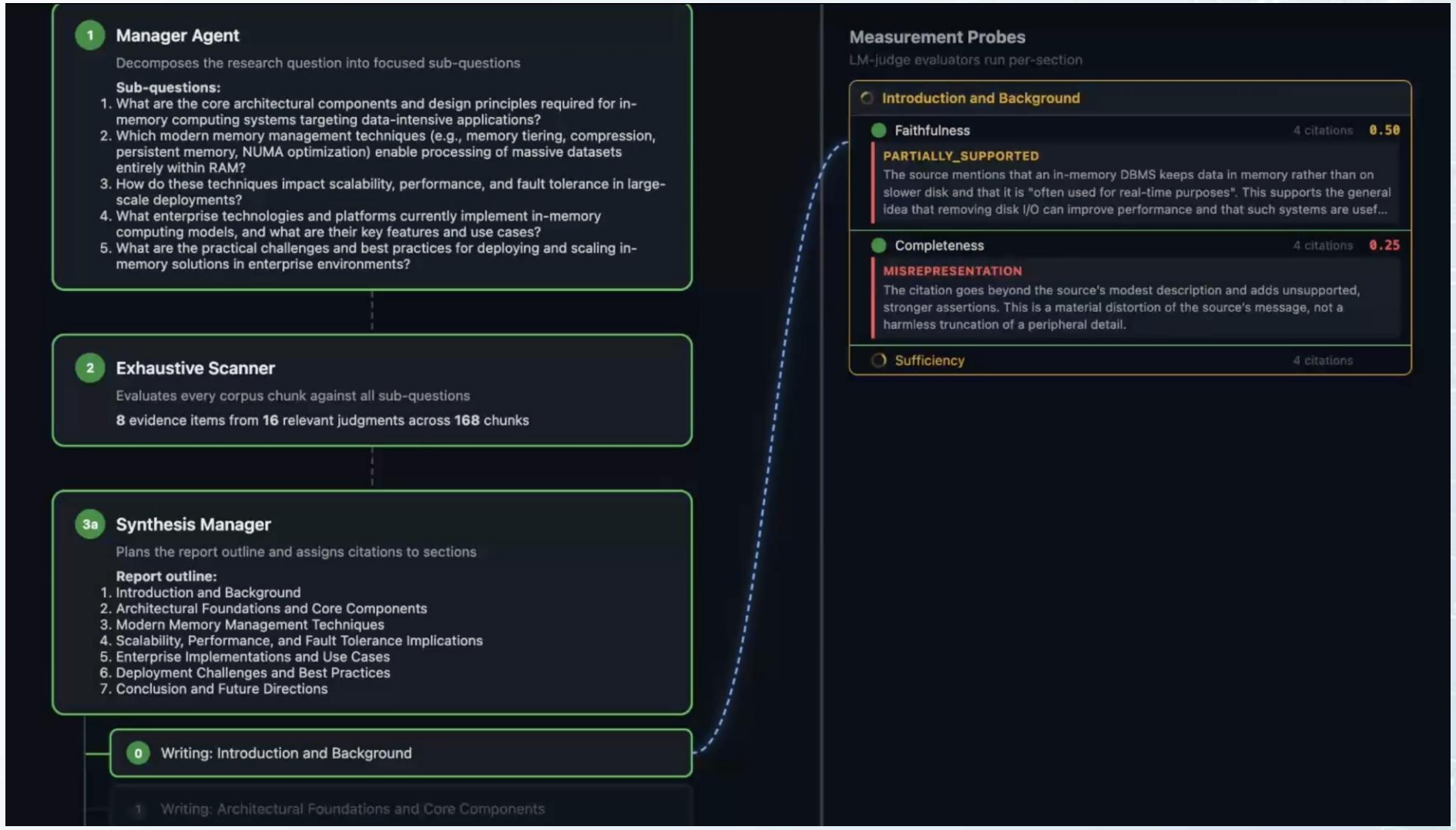


Why Deep Research Is a Perfect Test Case

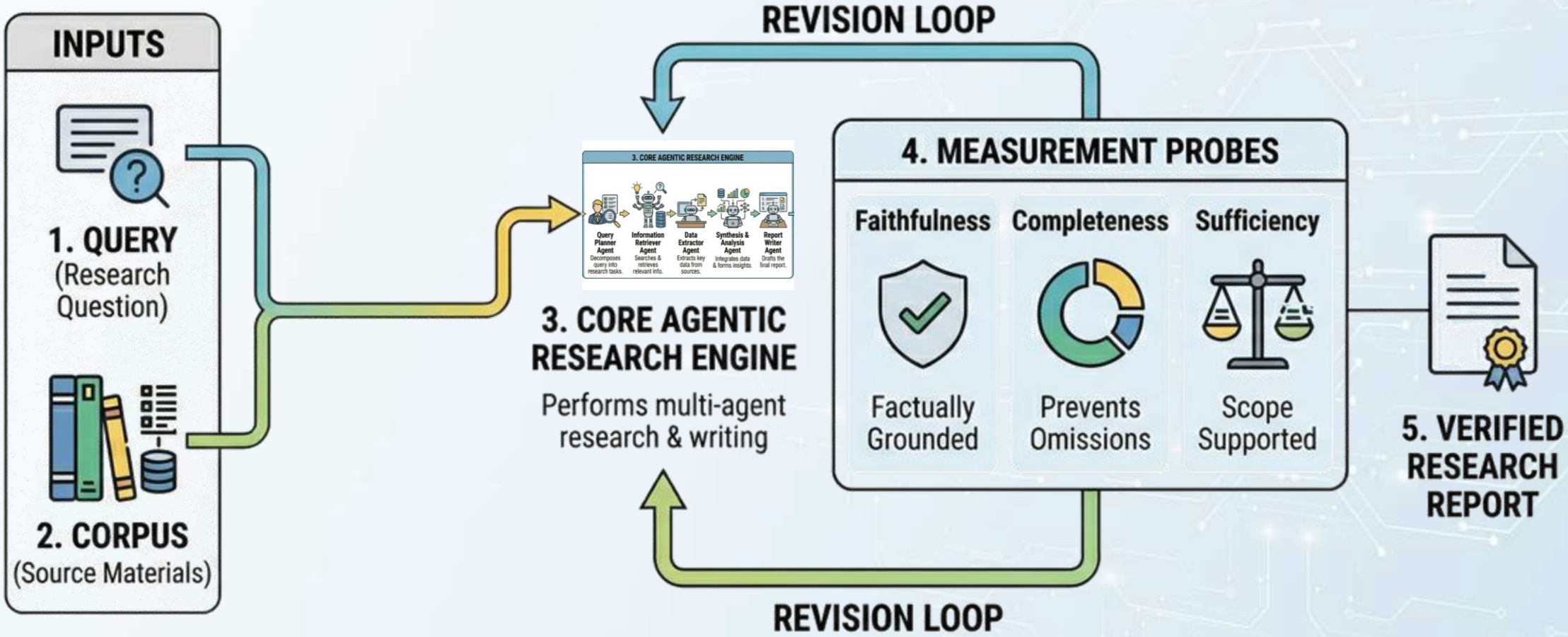
1. **Verifiable ground truth** (source documents define universe of correct answers)
2. **Citation traceability**
3. **Realistic complexity** (requires multi-step reasoning, synthesis across sources, judgment about relevance)
4. **Domain generality** (same pipeline works for cybersecurity standards, legal docs, scientific literature, etc.)
5. **Clear failure modes**
6. **Controlled scope** (fixed local corpus eliminates information retrieval)

Deep Research Testbed/Exemplar

Measurement Probe Results



Deep Research Testbed + Measurement Probes



Agenda

The Agentic AI Stack

Deep Research as a Case Study

Measurement Probes

Agentic Measurement Deep Dive

Q&A

Example Measurement Probes

FAITHFULNESS
(ANTI-HALLUCINATION)

- **Core Question:** Does the source passage actually support the claim?
- **Main Failure Mode Detected:** Errors of Commission (unsupported claims)

COMPLETENESS
(ANTI-CHERRY-PICKING)

- **Core Question:** Did the text capture the source's full message and nuance?
- **Main Failure Mode Detected:** Missing Hedges, Scope Erasure, Misrepresentation

SUFFICIENCY
(ANTI-OVERREACHING)

- **Core Question:** Does the source carry the burden of proof the claim requires?
- **Main Failure Mode Detected:** Scope/Sample Extrapolation, Rhetorical Inflation

●●●

Open Deep Research Pipeline

Real-time visualization of the agentic research workflow

Question

Analyze the architectural requirements and scalability of in memory computing for data-intensive applications. Specifically, evaluate how modern memory management techniques enable the processing of massive datasets entirely within RAM, and identify existing enterprise technologies that implement this model.

Corpus dir

./example-corpus

Model

gpt-oss-120b

Run Pipeline

1 **Manager Agent**

Decomposes the research question into focused sub-questions

2 **Exhaustive Scanner**

Evaluates every corpus chunk against all sub-questions

3a **Synthesis Manager**

Plans the report outline and assigns citations to sections

3c **Final Report Assembly**

Concatenates sections and appends references

Measurement Probes

LM-judge evaluators run per-section

Agenda

The Agentic AI Stack

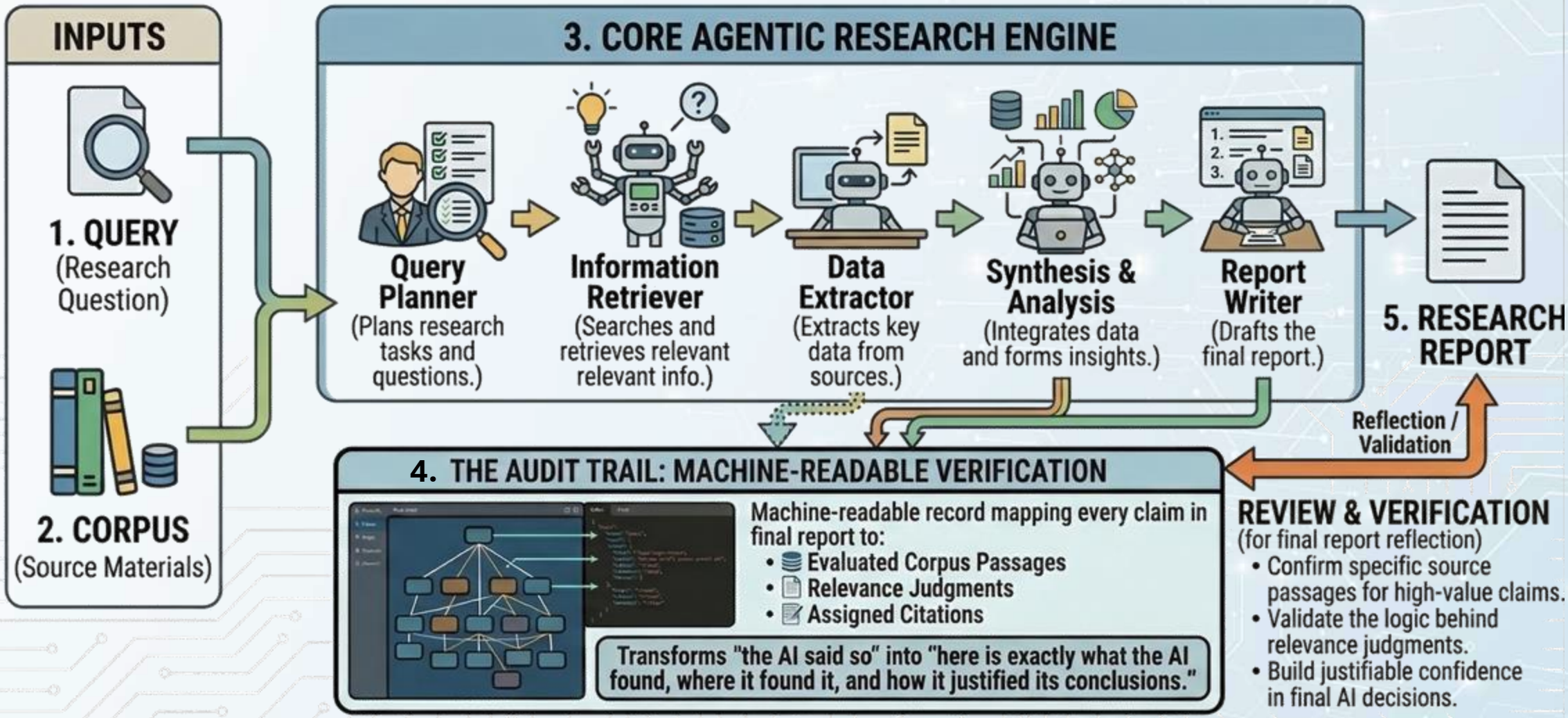
Deep Research as a Case Study

Measurement Probes

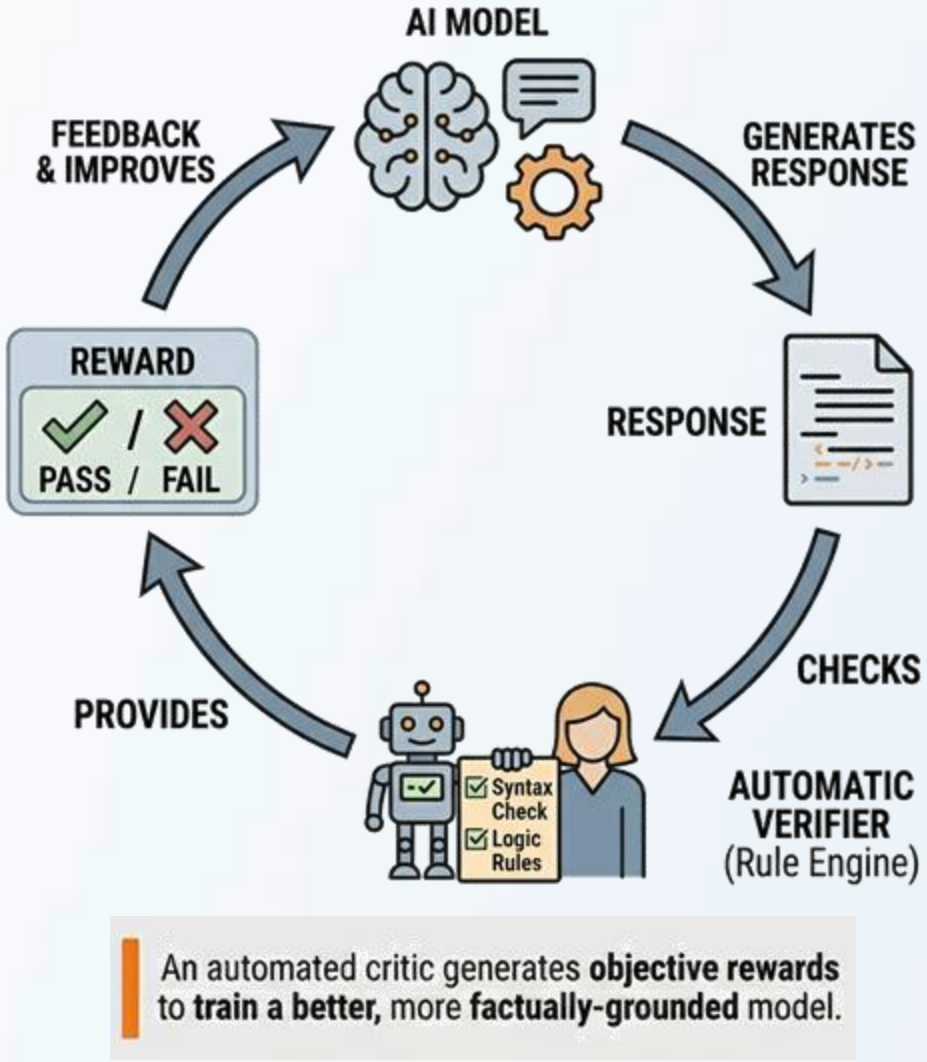
Agentic Measurement Deep Dive

Q&A

Building the Audit Trail



Reinforcement Learning from Verifiable Rewards (RLVR)



An automated critic generates **objective rewards** to **train a better,** more **factually-grounded** model.

Base LM next-token predictors

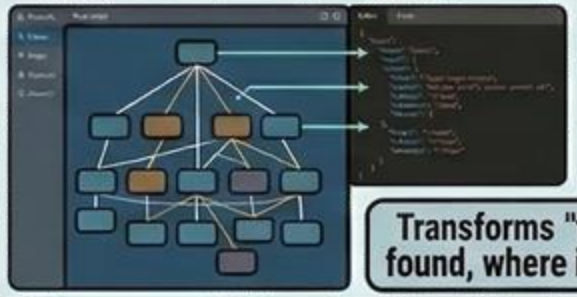
+ RLHF chatbots
  - (human feedback : subjective, expensive)
+ RLVR agents
  - (automated verifiers : objective, scalable)

Repurpose verifiers to measure any model's factual grounding

Repurposing RLVR for Agentic Verifiers: Using the Audit Trail

Probes are Built on the Audit Trail

THE AUDIT TRAIL: MACHINE-READABLE VERIFICATION



Machine-readable record mapping every claim in final report to:

- Evaluated Corpus Passages
- Relevance Judgments
- Assigned Citations

Transforms "the AI said so" into "here is exactly what the AI found, where it found it, and how it justified its conclusions."

Adapting RLVR to Inference Time Measurement Probes

ANY AGENTIC AI MODEL

RESPONSE + CITATIONS

ADVERSARIAL VERIFIER AGENTS
(Measurement Probes)
(Constrained to specific checks)

TAKE SPECIFIC CLAIM

TAKE CITED PASSAGE

DETERMINE SUPPORT
(Entailment Check)

Exampies of probes:

- FAITHFULNESS PROBE (anti-hallucination)
- COMPLETENESS PROBE (anti-cherry-picking)
- SUFFICIENCY PROBE (anti-overreaching)

SELF-CORRECTION LOOP

REVISION directed toward factual alignment

QUANTIFIABLE SAFETY SCORES
(e.g. 1.0 rationale
0.5 with rationale
0.0 with rationale)

MACHINE-READABLE AUDIT TRAIL

Verifiers are repurposed as ***probes*** at ***inference time*** to ***measure grounding*** and ***enable self-correction***, creating an automated measurement system.

generated by
gpt-oss-120b
grounded in
Jean-Claude Franchitti. Introduction to Computer Science. Open Stax, 2024.

Example Markdown Report

```
# Introduction and Context

In-memory computing is a paradigm that relocates the entire data set of a workload from traditional, slower storage tiers into volatile main memory so that the CPU can read and write data at RAM-speed rather than disk-speed[^2]. For data-intensive applications—such as real-time analytics, high-frequency trading, and large-scale machine learning—this shift can reduce query latency from seconds or minutes to milliseconds and increase throughput by orders of magnitude[^2].

The technical foundation of any in-memory system rests on the four pillars of technology architecture: computer, memory, storage, and network[^1]. Modern implementations must orchestrate these components so that massive data structures fit within the RAM of a cluster, are placed optimally across nodes, and are accessed via high-bandwidth, low-latency interconnects. An operating system mediates this interaction, providing isolation and controlled access to CPU, memory, persistent storage, and network interfaces[^4].

...
```

Probe: Citation Faithfulness

For every [^N] citation

- judge it against the source passage

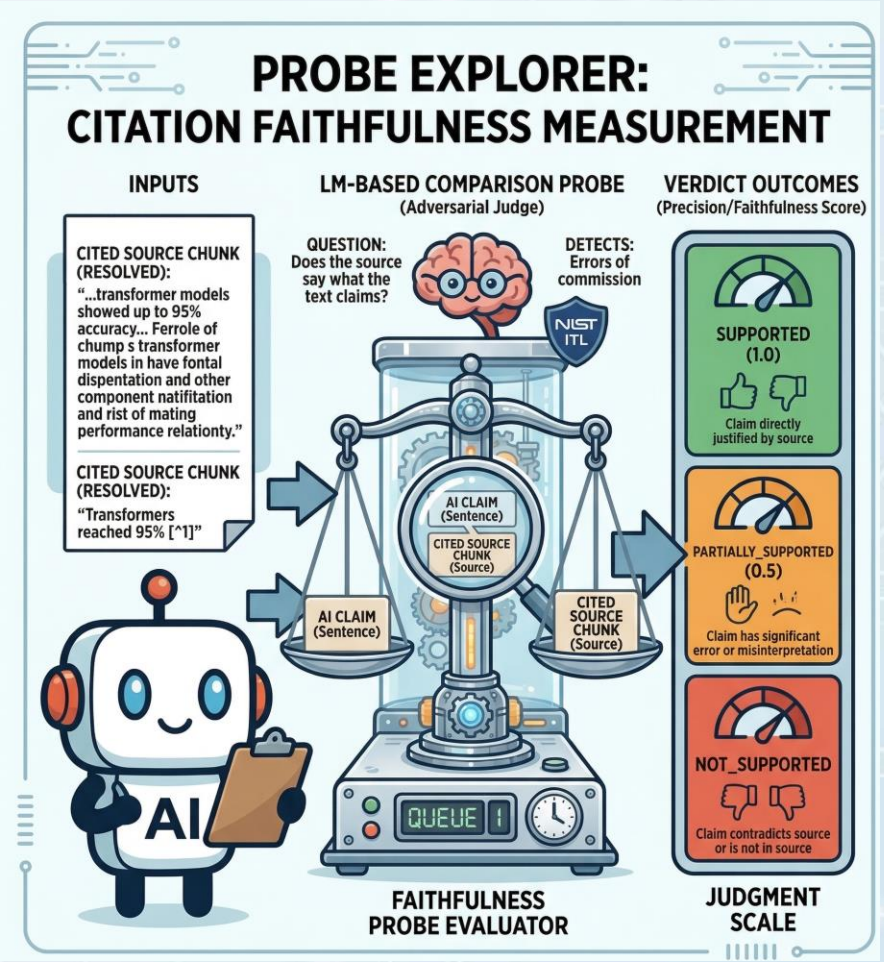
The anti-hallucination verifier

Evaluates errors of commission:

- Direct Support: The source explicitly states or clearly implies the claim
- No Contradictions: The sentence does not conflict with the cited text

LM-Judge Prompt

- https://github.com/usnistgov/agentic-research-measurement-probes/blob/main/src/probes/_prompts.py#L3

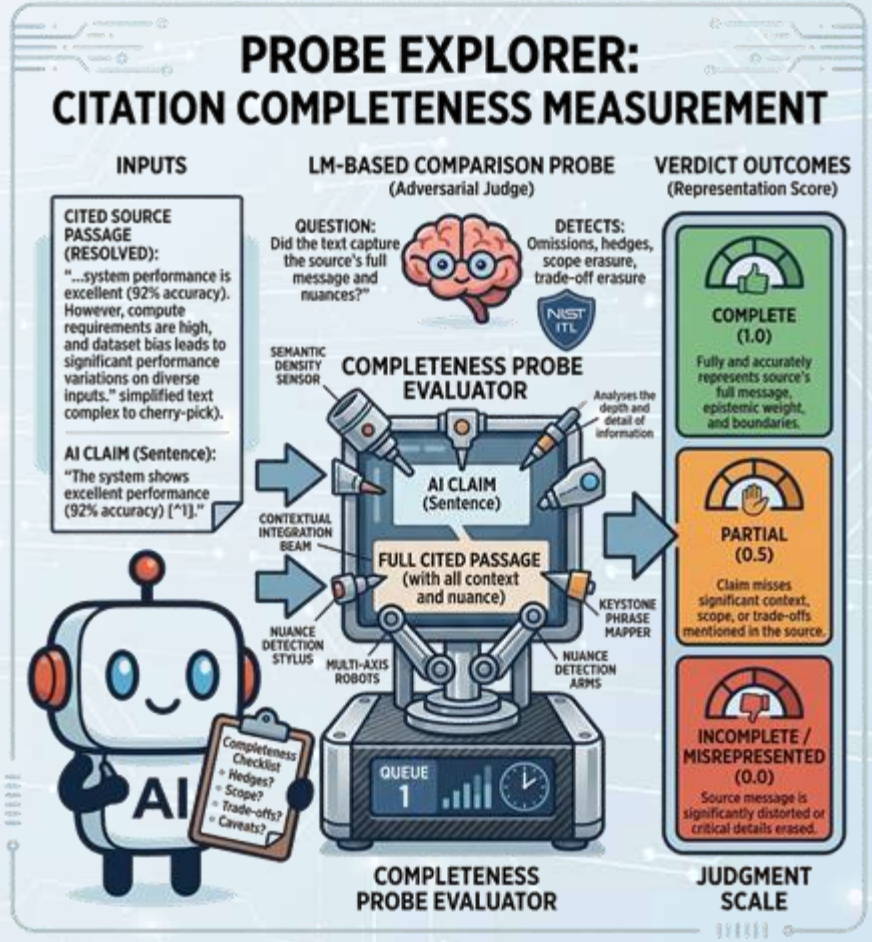


Probe: Citation Completeness

The anti-cherry-picking verifier

Evaluates Representational Delta

- Representational Fidelity: Captures intent
- Epistemic Integrity: Matches source certainty
  - **e.g., "suggests" must not become "proves"**
- Balanced Evidence: Avoids cherry-picking
  - **preserves key trade-offs**



LM-Judge Prompt

- https://github.com/usnistgov/agentic-research-measurement-probes/blob/main/src/probes/_prompts.py#L41

Probe: Citation Sufficiency

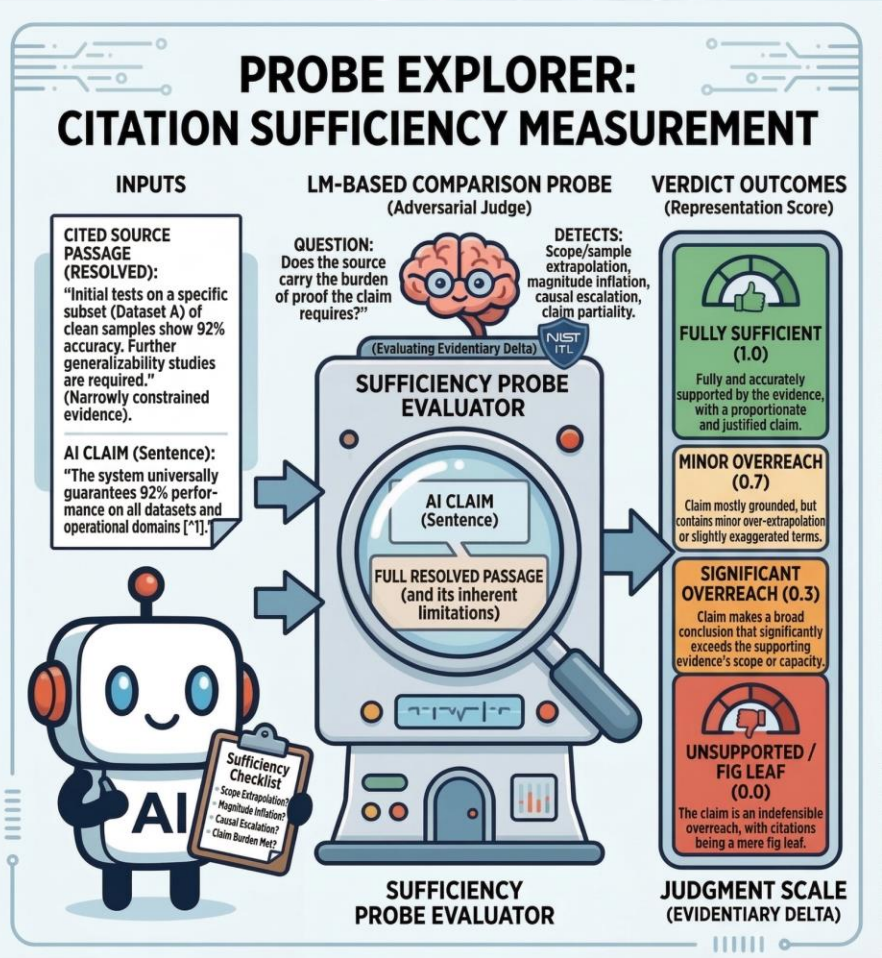
The anti-overreach verifier

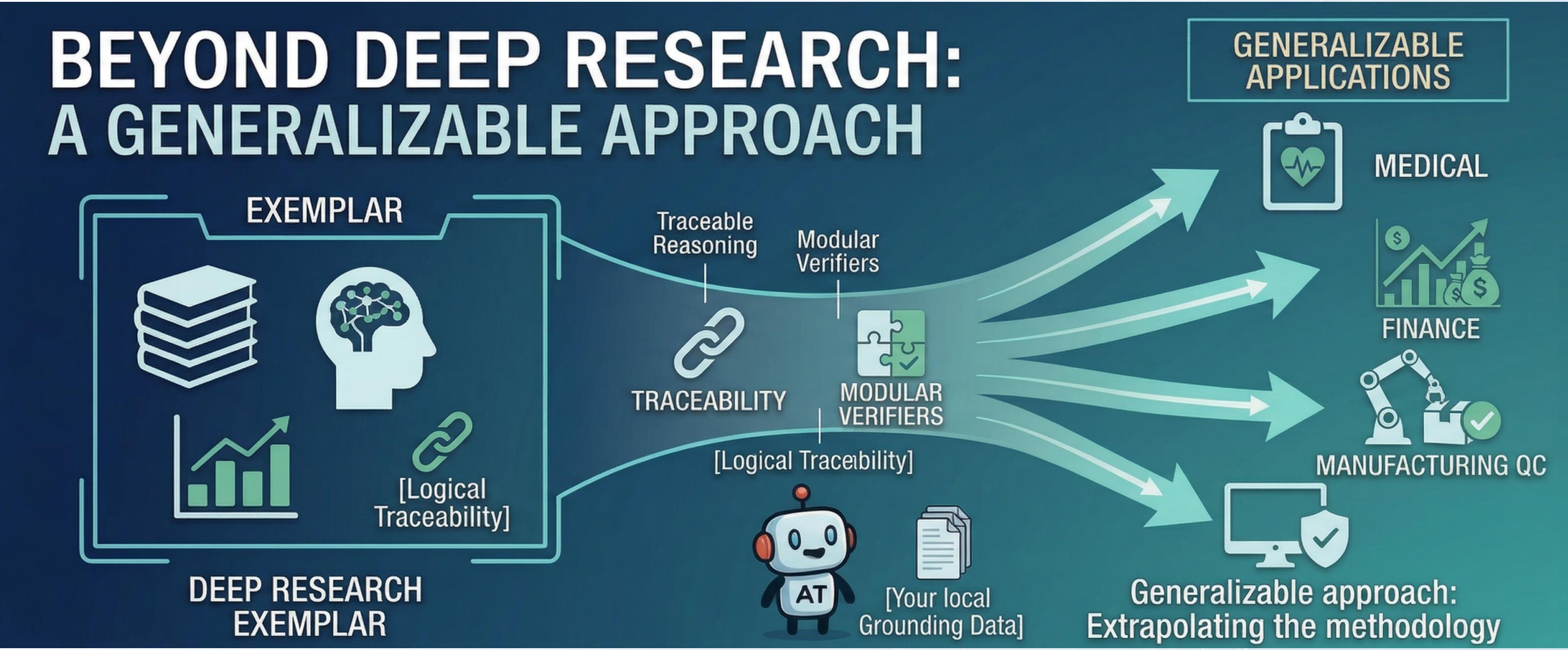
Evaluates Evidentiary Delta

- Evidentiary Sufficiency
  - **Does source bear the claim's full "burden of proof"**
- Scope Extrapolation
  - **Are localized findings improperly inflated into broad claims**

LM-Judge Prompt

- https://github.com/usnistgov/agentic-research-measurement-probes/blob/main/src/probes/_prompts.py#L93




BEYOND DEEP RESEARCH:
A GENERALIZABLE APPROACH
EXEMPLAR
[Logical Traceability]
DEEP RESEARCH EXEMPLAR
Traceable Reasoning
Modular Verifiers
TRACEABILITY
MODULAR VERIFIERS
[Logical Traceability]
AT
[Your local Grounding Data]
GENERALIZABLE APPLICATIONS
MEDICAL
FINANCE
MANUFACTURING QC
Generalizable approach:
Extrapolating the methodology

The Path Forward

Expand Set of Measurement Probes

- Section level probes (i.e. coverage/cohesiveness/justification)
- Evidence Justification (Adversarial):
  - **Poke holes in claims based on evidence at hand**

Turn Static Measurement Probes into Dynamic Verifiers

- Allow the Agentic AI Deep Research to correct issues surfaced by probes

Evaluate Improvement in accuracy & grounding of Deep Research

- Can smaller models support large research agents?

Research into Measurement of argument evidentiary support

Thank You

Current State

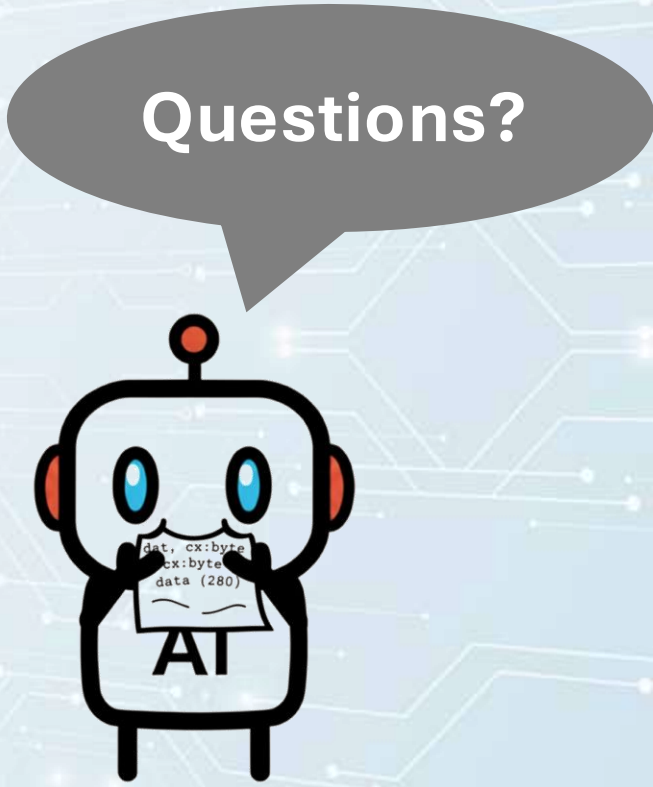
- Demonstrator of verifiable, citation-grounded Agentic AI Open Deep Research
- https://github.com/usnistgov/agentic-research-measurement-probes

Vision

- Documented, reproducible methodology for auditing agent behavior against defined document corpus

Call to Action

- What are Your Agentic-AI measurement challenges?
- Submit your ideas via email at itl-ai-program@nist.gov



NIST INFORMATION TECHNOLOGY LABORATORY

Contact Us:

Scan the code to subscribe for AI-related updates from NIST's Information Technology Laboratory (ITL)

Or email us: itl-ai-program@nist.gov

Next webinar: Details Coming Soon!

Just released: Concept Note on Artificial Intelligence Risk Management Framework Profile on Trustworthy AI in Critical Infrastructure