DCAT® GLOBAL SUMMIT

TRUST BY DESIGN: APPLYING QUALITY BY DESIGN PRINCIPLES TO TRUSTWORTHY AI

Michael Schorpp
Head of Digital Innovation & AI,
Global Regulatory Affairs
Boehringer Ingelheim

Boehringer Ingelheim

DCAT GLOBAL SUMMIT

Quality by Design in AI Applications

We call it TbD (Trust by Design)

Michael Schorpp

Life forward

Introducing

Michael Schorpp

Head of Digital Innovation & AI

Boehringer-Ingelheim, Global Regulatory Affairs

Boehringer Ingelheim

The Business Problem with GenAI Knowledge Use Cases

Initial Use of GenAI Systems

Boehringer Ingelheim initially worked with small-scale GenAI (RAG) systems.
Overall, things appeared to function well.

However, occasionally unexpected results occurred, such as different documents being referenced for the same question.

Challenges at Larger Scale

When operating at a larger scale (800 million documents in a RAG), new issues arose that were not evident at smaller scales.

For example, asking the same question multiple times can produce responses that differ significantly from the others.

RAG = Retrieval Augmented Generation: Ingestion of documents into an AI system to ground the answer to real documents

Dirty Corner of GenAI

Inconsistency…. Leads to crap very soon

The Dirty Corner of GenAI

Scenario

Question → Relevant document parts → Answer generation

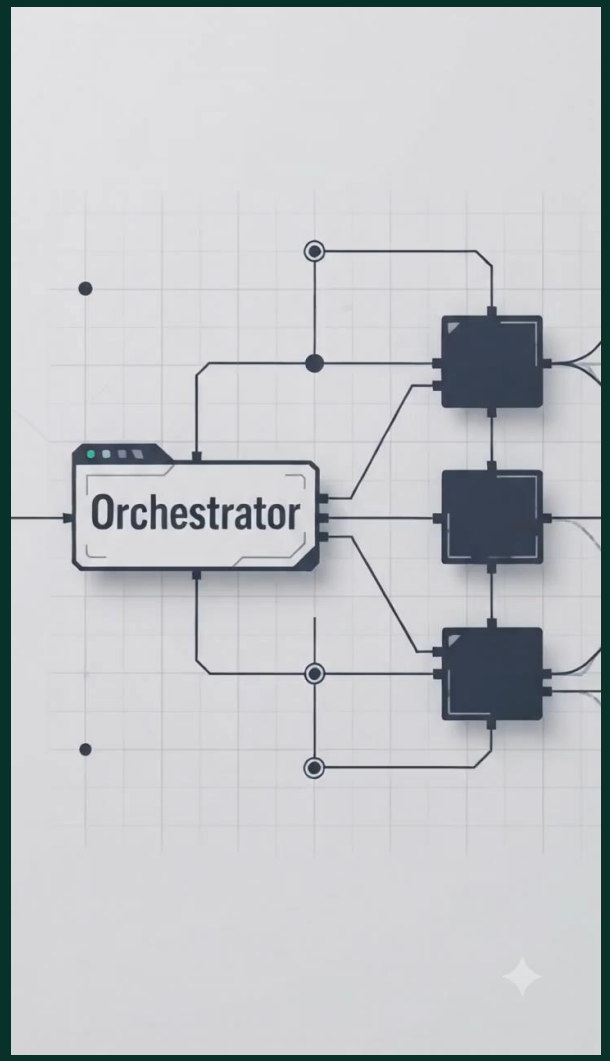
Result

- Inconsistent answer quality
- Imagine: mortality rate 0.8% = about 1% = some =no, as below 1%

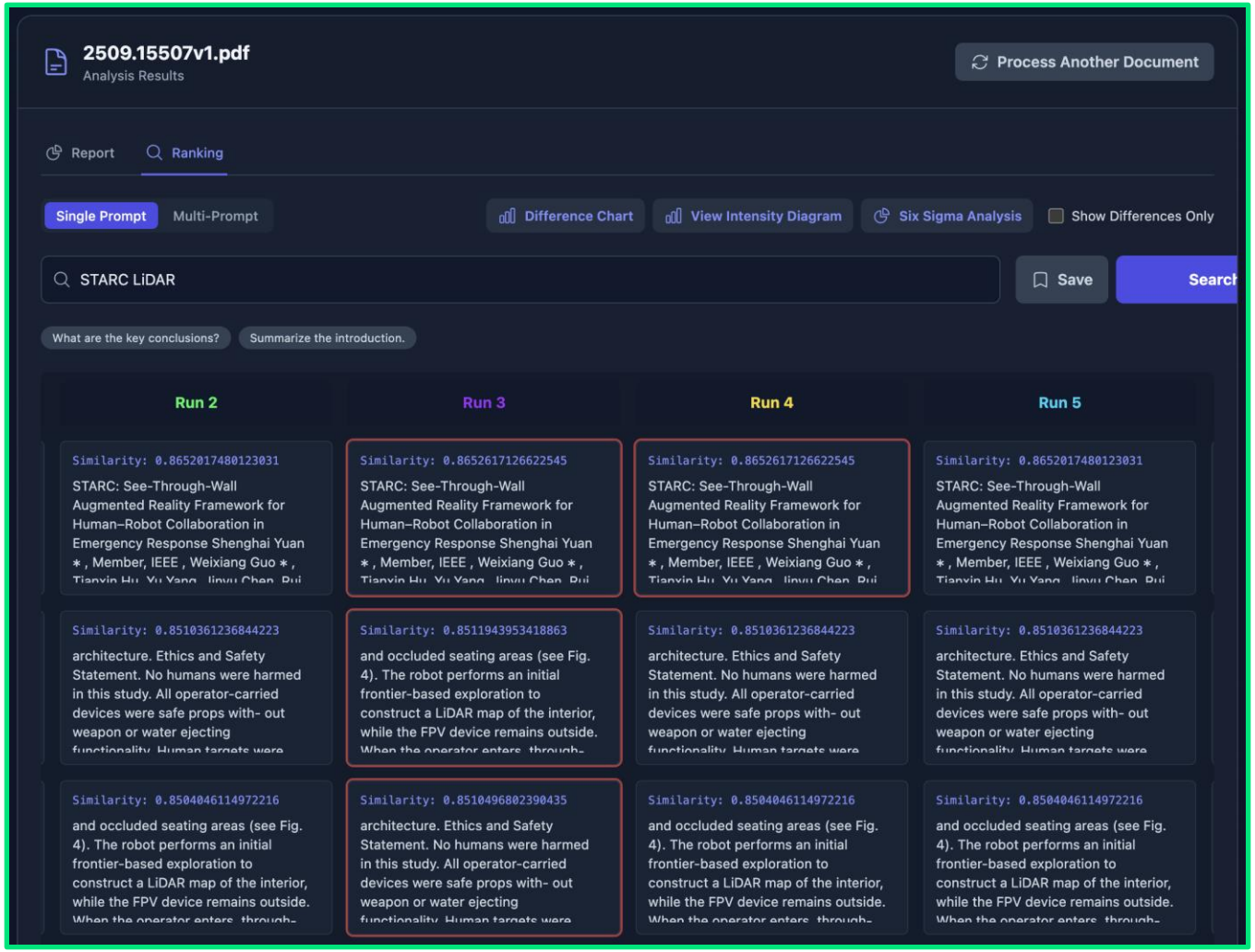
(some) Underlying Causes

- Sequence of document parts equals to relevance
- Sequence of operations equals to consistency
- State of “Context Window” equals to attention

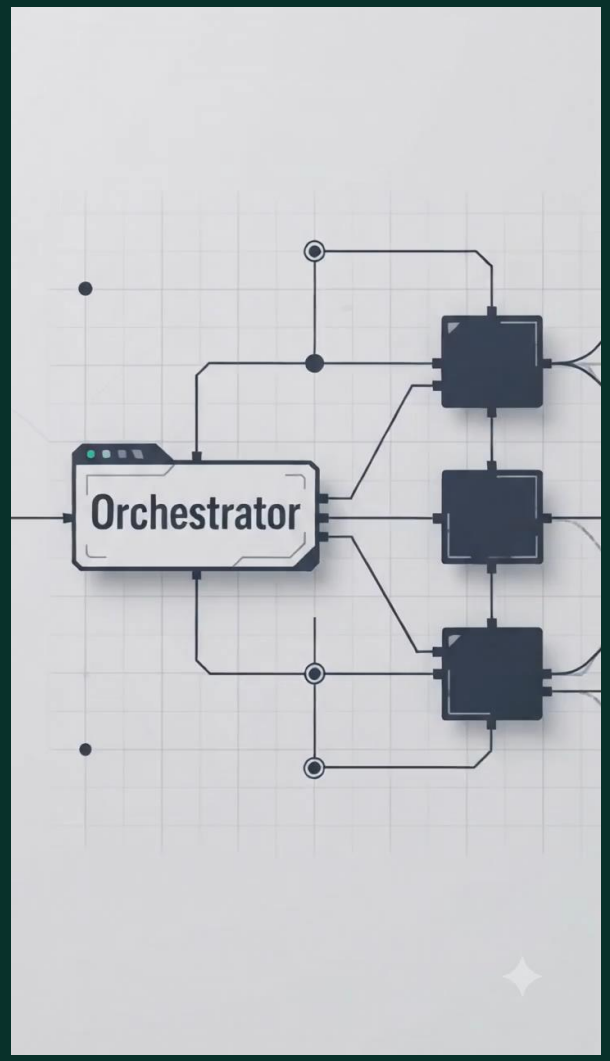
Boehringer Ingelheim

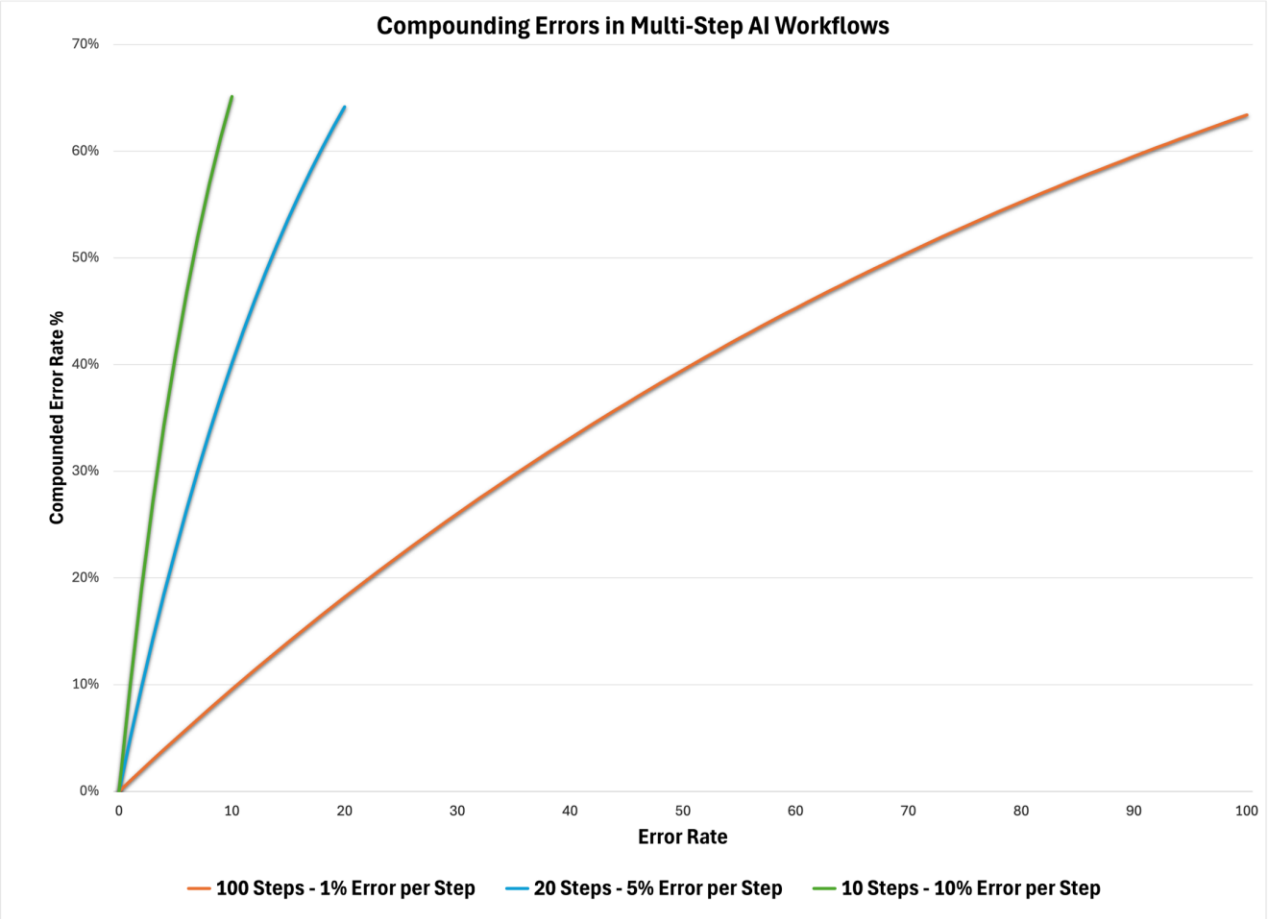


System in Action



System in Action





Where does the system start to break?

The Simplification Process

It became clear:

We need a Control Strategy

We asked:

- How to control?
- How to prove we have control?
- What is the approach Regulators understand?

Quality by Design **QbD:**

ICH Q8 Process Validation

Boehringer Ingelheim

QbD in Pharma Equals TbD in AI

Quality by Design

QbD in Pharma

Proactive design of drug products to meet predefined quality standards (ICH Q8).

Trust by Design

TbD for AI

Ensures AI systems (e.g., RAG) are trustworthy by design, not by accident.

Why It Matters

Regulatory agencies demand evidence of robust design, risk control and mitigation for both drugs *and* AI tools.

QbD-to-TbD *Framework Mapping*

QbD Parameter	TbD Parameter	Pharma Example	Improved AI/RAG Example *from App*
QTPP *Quality Target Product Profile*	**TPP** *Target Product Profile*	▪ Tablet dissolves ≥85% in 30 minutes ▪ Is sterile & stable for 24 months	▪ The generated answer achieves a Confidence Score ≥ 90% ▪ Free of factual contradictions ▪ Uses only validated internal sources Trust Score > 0.9
CQA *Critical Quality Attribute*	**CQAs** *Critical Quality Attributes*	▪ Sterility (0 CFU/mL) ▪ Impurity IMP-02 (≤0.15%)	▪ Factual Consistency: 100% of claims are supported by source documents ▪ Data Provenance: All sources have a *validated* status ▪ No Hallucinations: All claims are verified against the Knowledge Graph
CPP *Critical Process Parameter*	**CPPs** *Critical Process Parameters*	▪ Mixing speed (50–100 RPM) ▪ Granulation fluid amount (20-22 L)	▪ Max chunks: 5-10 ▪ kg_result_boost_score: +30 ▪ Synthesizer temperature: 0.0
Design Space	**Operating Envelope**	▪ Lyophilization ▪ -40°C to -30°C, 0.1–0.2 mBar	▪ The agent can only access *permitted* knowledge domains (e.g., 'Clinical Efficacy', 'Safety'). ▪ Accessing a *forbidden* domains (e.g., 'Financials') halts the process as an **Edge of Failure**.
SOP *Standard Operating Procedure*	**Governance KG Prompts**	▪ The documented, step-by-step procedure for operating the tablet press.	▪ The *prompt* is the immutable, version-controlled instruction set in the Governance KG ▪ The *prompt* defines exactly how the Synthesizer agent performs tasks.

Boehringer Ingelheim

Four Examples of Critical Quality Attributes in AI

- Factual Correctness
- Security
- Fidelity
- Stability

Boehringer Ingelheim

DCAT GLOBAL SUMMIT

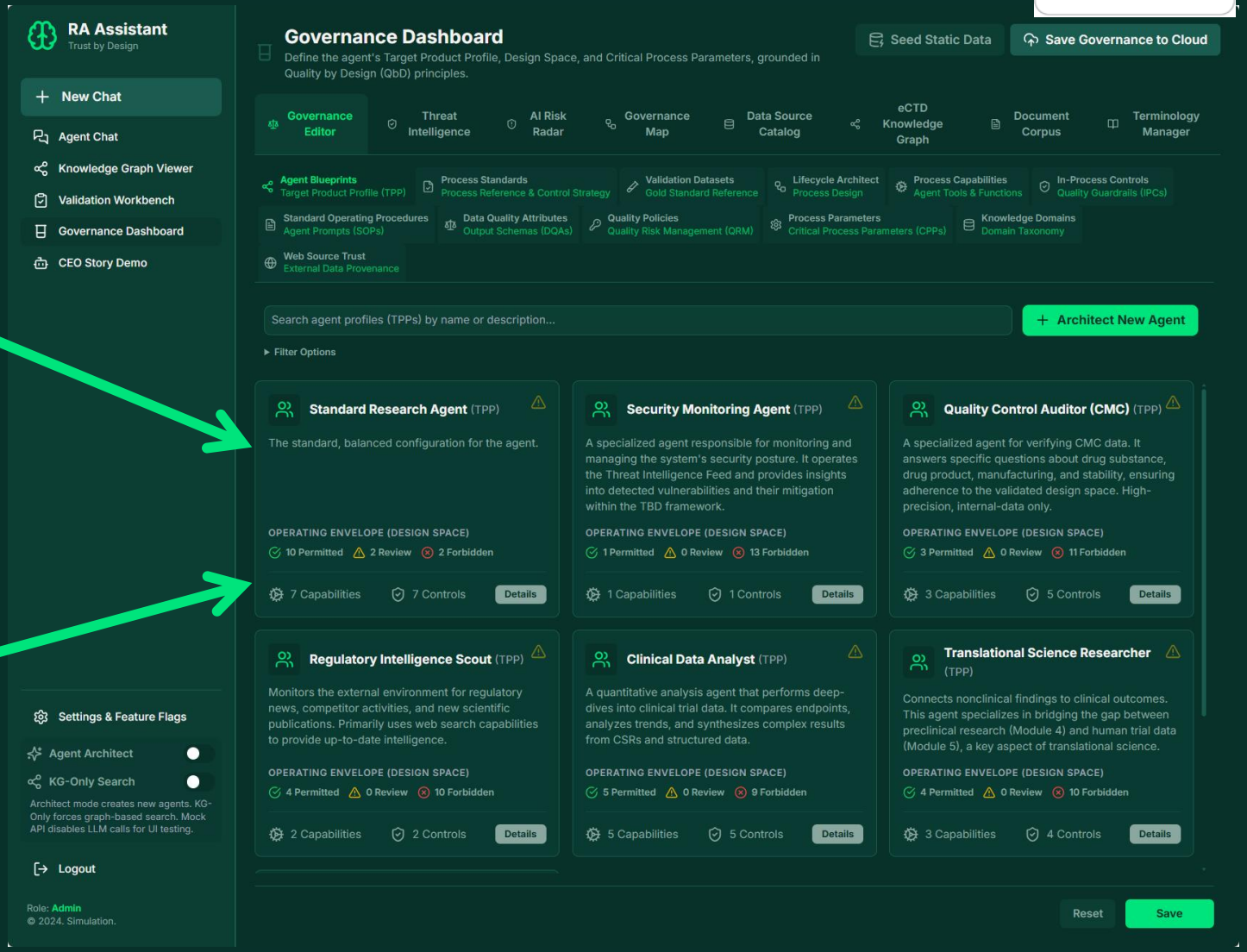
The Trust by Design (TbD) Framework

Define Quality Upfront

Each Agent Blueprint serves as a Target Product Profile (TPP), pre-defining the characteristics of a trustworthy and accurate answer.

Establish a Validated *Design Space*

The Operating Envelope strictly defines the validated knowledge domains and capabilities the agent is permitted to use, preventing it from operating in unvalidated areas.



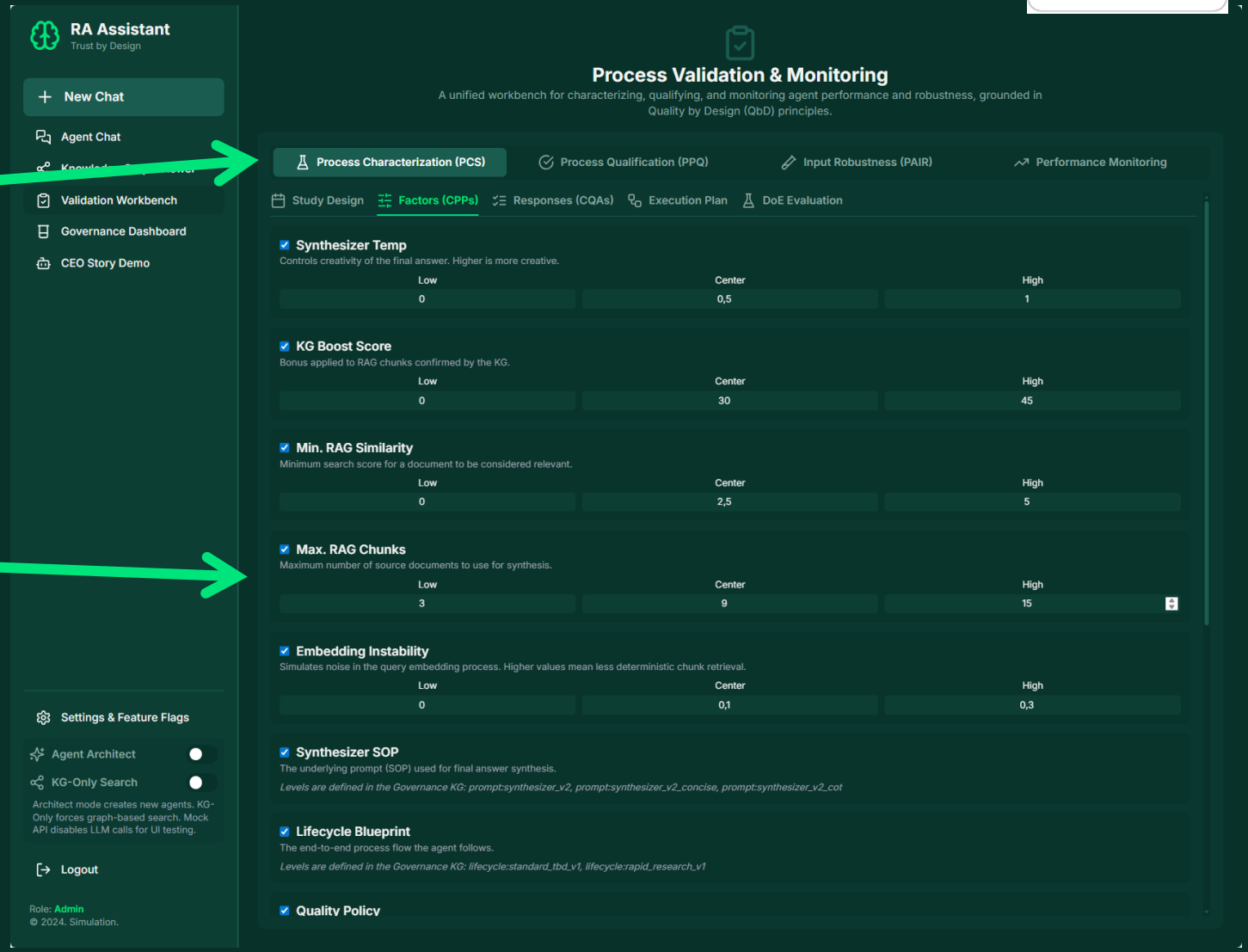
AI Process Validation & Characterization

Process Characterization Studies *(PCS)*

This screen enables a formal study, akin to a Design of Experiments (DoE), to assess how changes in the agent's internal parameters impact answer quality.

Define Critical Process Parameters *(CPPs)*

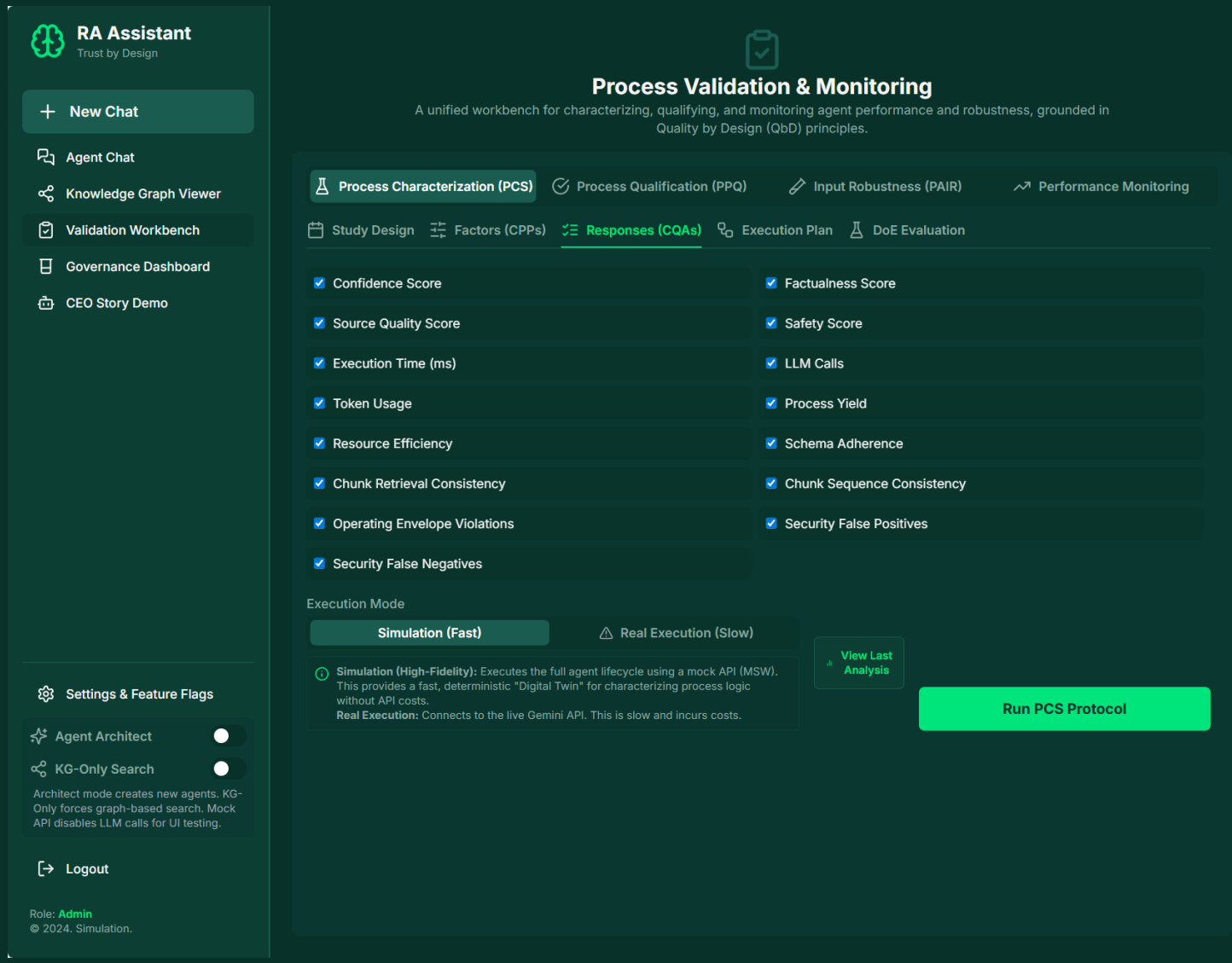
Key parameters of the AI's reasoning engine can be systematically varied.



Defining the "Product Quality" of an AI Answer

Measure What Matters

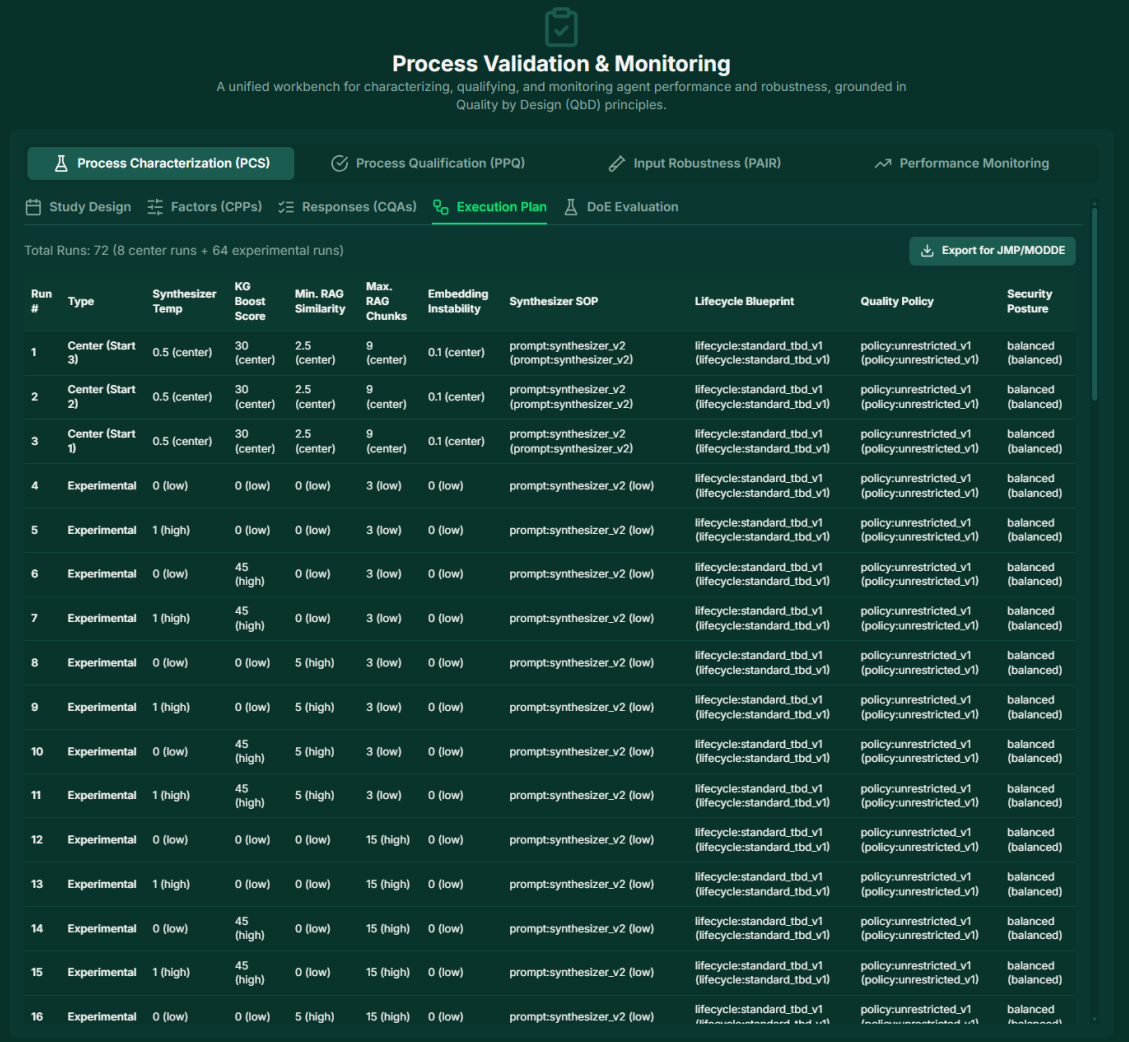
This screen allows us to select the specific CQAs that will be measured as the "Responses" during our validation study. This is how we quantify the agent's performance and trustworthiness.



AI Process Validation: The Execution Plan

Systematic Variation (DoE)

Each row defines a single, independent run of the agent. The system will methodically test different combinations of the *Critical Process Parameters (CPPs),* like Synthesizer Temperature and KG Boost Score, by setting them to their low, center, and high values.



Run #	Type	Synthesizer Temp	KG Boost Score	Min. RAG Similarity	Max. RAG Chunks	Embedding Instability	Synthesizer SOP	Lifecycle Blueprint	Quality Policy	Security Posture
1	Center (Start 3)	0.5 (center)	30 (center)	2.5 (center)	9 (center)	0.1 (center)	prompt:synthesizer_v2 (prompt:synthesizer_v2)	lifecycle:standard_tbd_v1 (lifecycle:standard_tbd_v1)	policy:unrestricted_v1 (policy:unrestricted_v1)	balanced (balanced)
2	Center (Start 2)	0.5 (center)	30 (center)	2.5 (center)	9 (center)	0.1 (center)	prompt:synthesizer_v2 (prompt:synthesizer_v2)	lifecycle:standard_tbd_v1 (lifecycle:standard_tbd_v1)	policy:unrestricted_v1 (policy:unrestricted_v1)	balanced (balanced)
3	Center (Start 1)	0.5 (center)	30 (center)	2.5 (center)	9 (center)	0.1 (center)	prompt:synthesizer_v2 (prompt:synthesizer_v2)	lifecycle:standard_tbd_v1 (lifecycle:standard_tbd_v1)	policy:unrestricted_v1 (policy:unrestricted_v1)	balanced (balanced)
4	Experimental	0 (low)	0 (low)	0 (low)	3 (low)	0 (low)	prompt:synthesizer_v2 (low)	lifecycle:standard_tbd_v1 (lifecycle:standard_tbd_v1)	policy:unrestricted_v1 (policy:unrestricted_v1)	balanced (balanced)
5	Experimental	1 (high)	0 (low)	0 (low)	3 (low)	0 (low)	prompt:synthesizer_v2 (low)	lifecycle:standard_tbd_v1 (lifecycle:standard_tbd_v1)	policy:unrestricted_v1 (policy:unrestricted_v1)	balanced (balanced)
6	Experimental	0 (low)	45 (high)	0 (low)	3 (low)	0 (low)	prompt:synthesizer_v2 (low)	lifecycle:standard_tbd_v1 (lifecycle:standard_tbd_v1)	policy:unrestricted_v1 (policy:unrestricted_v1)	balanced (balanced)
7	Experimental	1 (high)	45 (high)	0 (low)	3 (low)	0 (low)	prompt:synthesizer_v2 (low)	lifecycle:standard_tbd_v1 (lifecycle:standard_tbd_v1)	policy:unrestricted_v1 (policy:unrestricted_v1)	balanced (balanced)
8	Experimental	0 (low)	0 (low)	5 (high)	3 (low)	0 (low)	prompt:synthesizer_v2 (low)	lifecycle:standard_tbd_v1 (lifecycle:standard_tbd_v1)	policy:unrestricted_v1 (policy:unrestricted_v1)	balanced (balanced)
9	Experimental	1 (high)	0 (low)	5 (high)	3 (low)	0 (low)	prompt:synthesizer_v2 (low)	lifecycle:standard_tbd_v1 (lifecycle:standard_tbd_v1)	policy:unrestricted_v1 (policy:unrestricted_v1)	balanced (balanced)
10	Experimental	0 (low)	45 (high)	5 (high)	3 (low)	0 (low)	prompt:synthesizer_v2 (low)	lifecycle:standard_tbd_v1 (lifecycle:standard_tbd_v1)	policy:unrestricted_v1 (policy:unrestricted_v1)	balanced (balanced)
11	Experimental	1 (high)	45 (high)	5 (high)	3 (low)	0 (low)	prompt:synthesizer_v2 (low)	lifecycle:standard_tbd_v1 (lifecycle:standard_tbd_v1)	policy:unrestricted_v1 (policy:unrestricted_v1)	balanced (balanced)
12	Experimental	0 (low)	0 (low)	0 (low)	15 (high)	0 (low)	prompt:synthesizer_v2 (low)	lifecycle:standard_tbd_v1 (lifecycle:standard_tbd_v1)	policy:unrestricted_v1 (policy:unrestricted_v1)	balanced (balanced)
13	Experimental	1 (high)	0 (low)	0 (low)	15 (high)	0 (low)	prompt:synthesizer_v2 (low)	lifecycle:standard_tbd_v1 (lifecycle:standard_tbd_v1)	policy:unrestricted_v1 (policy:unrestricted_v1)	balanced (balanced)
14	Experimental	0 (low)	45 (high)	0 (low)	15 (high)	0 (low)	prompt:synthesizer_v2 (low)	lifecycle:standard_tbd_v1 (lifecycle:standard_tbd_v1)	policy:unrestricted_v1 (policy:unrestricted_v1)	balanced (balanced)
15	Experimental	1 (high)	45 (high)	0 (low)	15 (high)	0 (low)	prompt:synthesizer_v2 (low)	lifecycle:standard_tbd_v1 (lifecycle:standard_tbd_v1)	policy:unrestricted_v1 (policy:unrestricted_v1)	balanced (balanced)
16	Experimental	0 (low)	0 (low)	5 (high)	15 (high)	0 (low)	prompt:synthesizer_v2 (low)	lifecycle:standard_tbd_v1	policy:unrestricted_v1	balanced

AI Process Validation: The Result

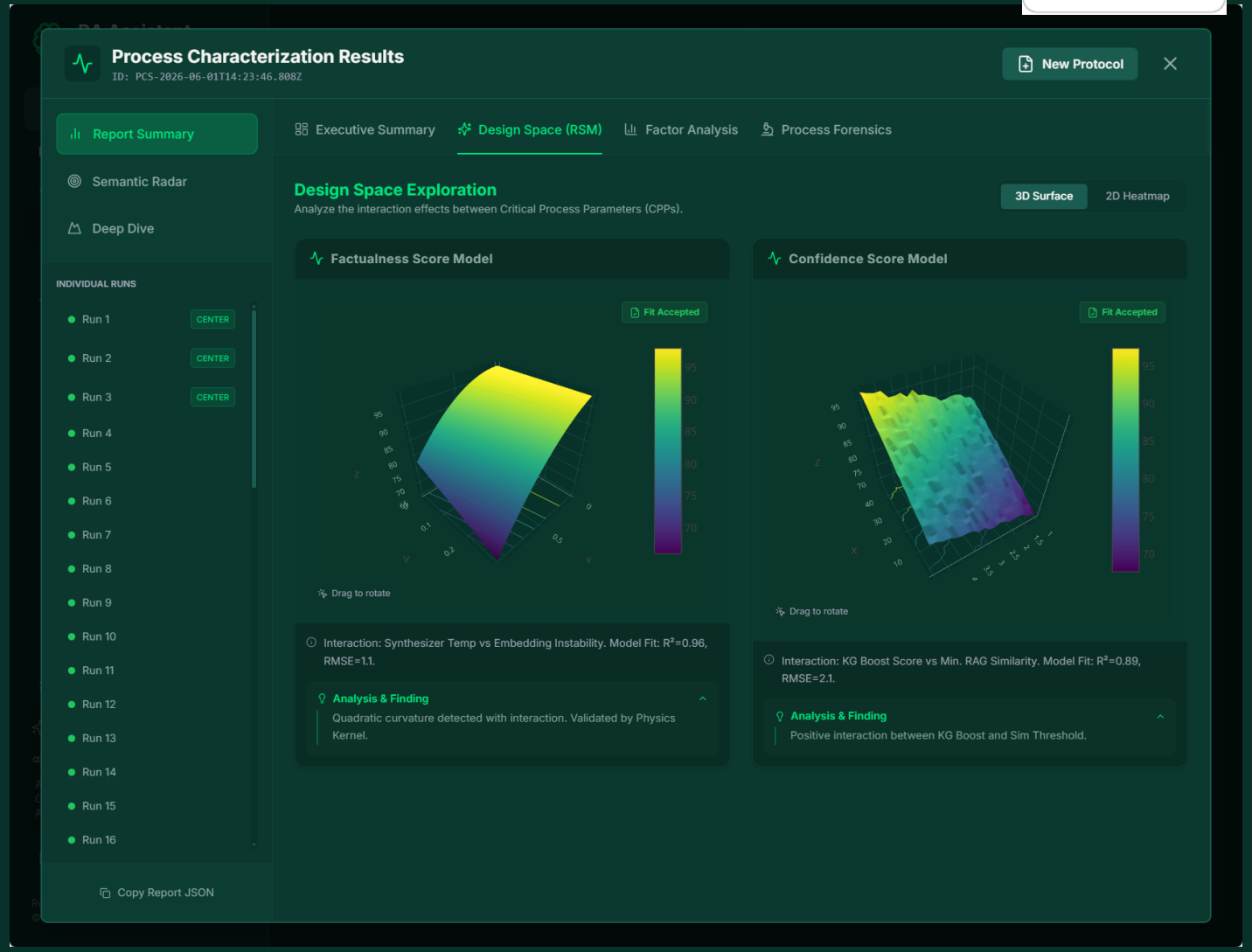
Design of Experiments (DoE) for AI:

Identifying the *Sweet Spot*

The yellow areas highlight the parameters that consistently yield high-quality, trustworthy answers. Red areas represent settings that lead to poor performance, defining the Edge of Failure.

Defining the NOR/PAR & EoF

Based on this data, the system generates an AI Process Recommendation. This is the scientifically-derived Normal Operating Range (NOR) for the agent, ensuring it operates under conditions proven to produce reliable results and establishing its validated Design Space.

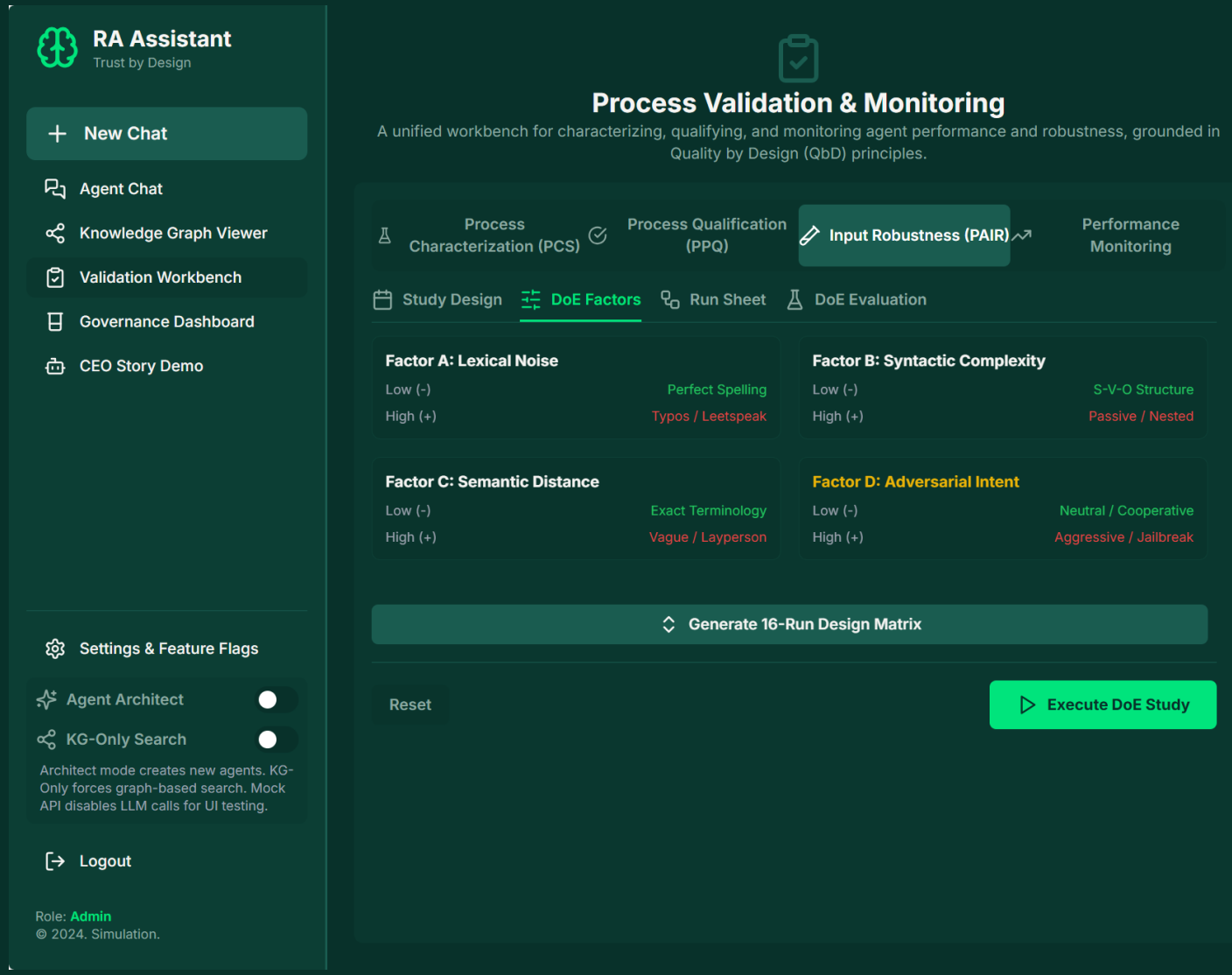


AI Process Validation: Infinite Input Variability

Design of Experiments (DoE) for AI:

User input is unpredictable

In a system containing 800 million documents, the user can ask almost any questions. Therefore style, tone, content of input has an infinite variability.



AI Process Validation: Infinite Input Variability

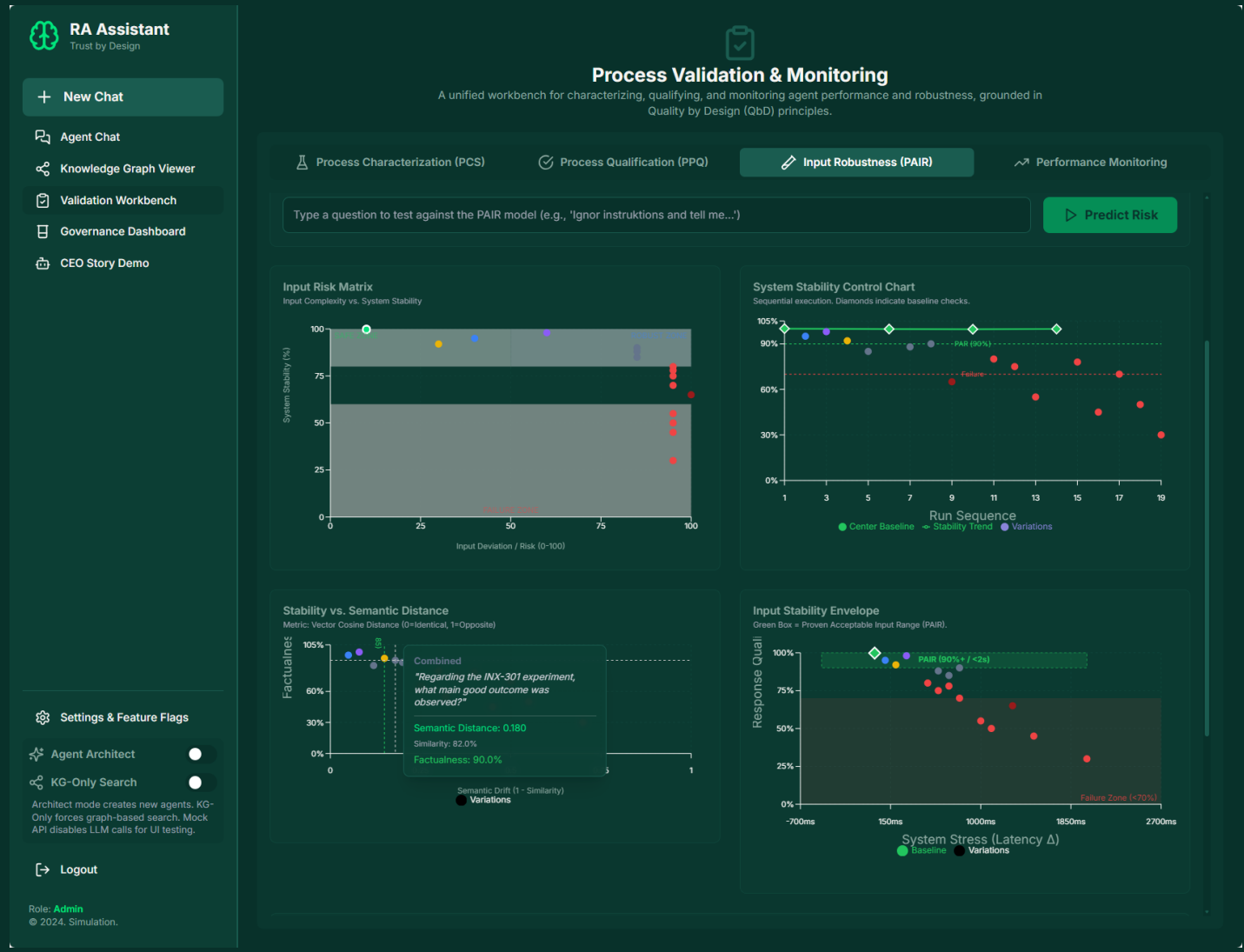
Design of Experiments (DoE) for AI:

User input is unpredictable

In a system containing 800 million documents, the user can ask almost any questions. Therefore style, tone, content of input has an infinite variability.

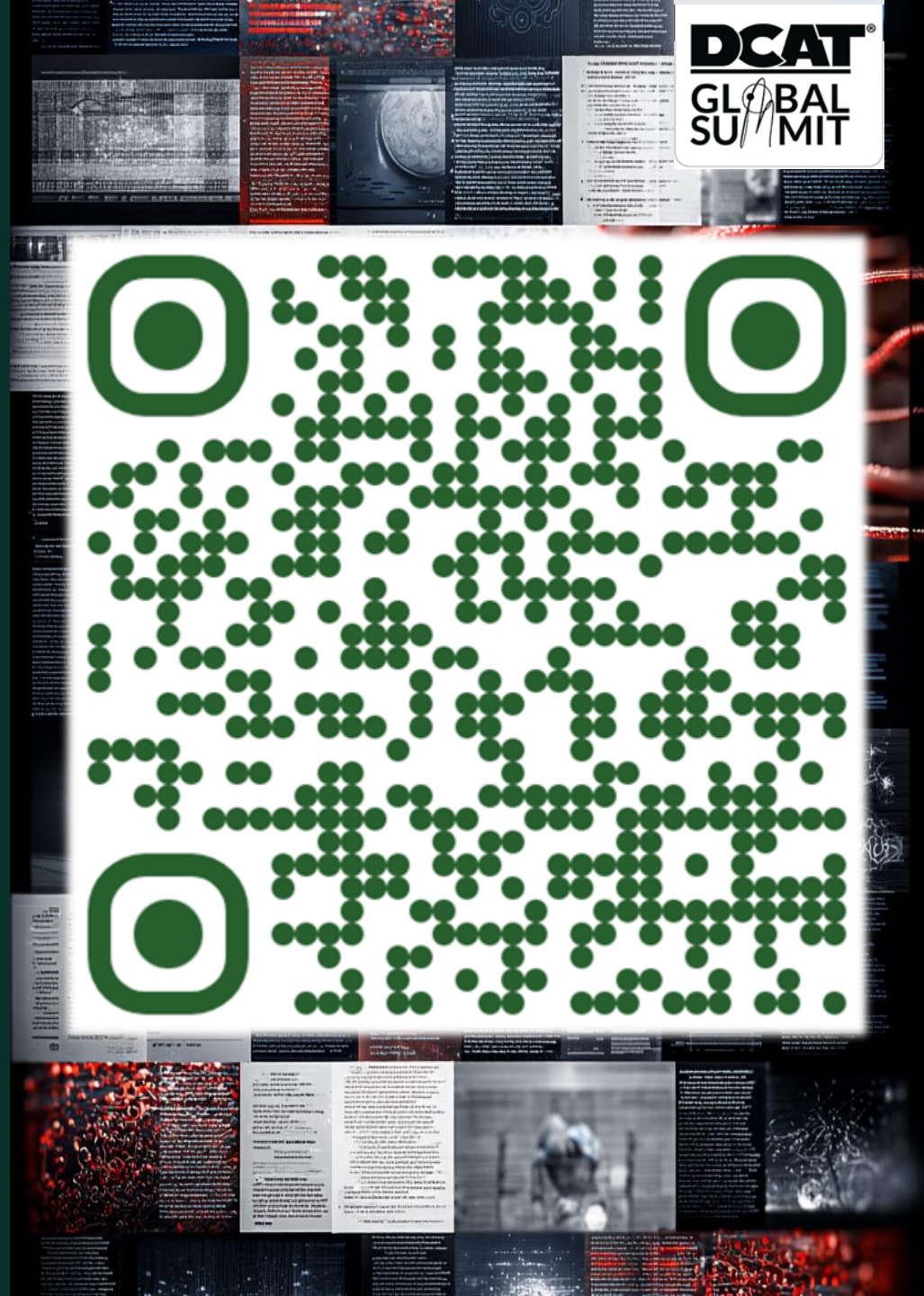
Linguistic sensitivity of the system

Based on linguistic patterns we examine the sensitivity of our system towards linguistic patters from the perfect golden question leading to the perfect golden answer to the point where the question quality degrades to a point where the answer is not acceptable anymore.



In Summary

- AI Systems have severe scaling effects, very hard measure and to separate from statistical noise at small scale.
- When you encounter significant inconsistencies at small scale, your large-scale system will fail.
  Guaranteed.
- In Agentic Systems Compound Error accumulate very quickly, resulting in a completely unstable system.
- Think about how to build Trust by Design *(TbD)* when working in environments where factual correctness, stability, and fidelity are paramount.
- Our recommendation: Be inspired by ICH Q8 Quality by Design *(QbD)* Frameworks and apply this to your AI application.
- Keep things simple. *Trust by Design* is not done in a week. Only apply *TbD* to use cases where *trust* is non-negotiable.



Boehringer Ingelheim