



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

Michael Schorpp

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



Quality by Design in AI Applications

We call it TbD (Trust by Design)

Michael Schorpp

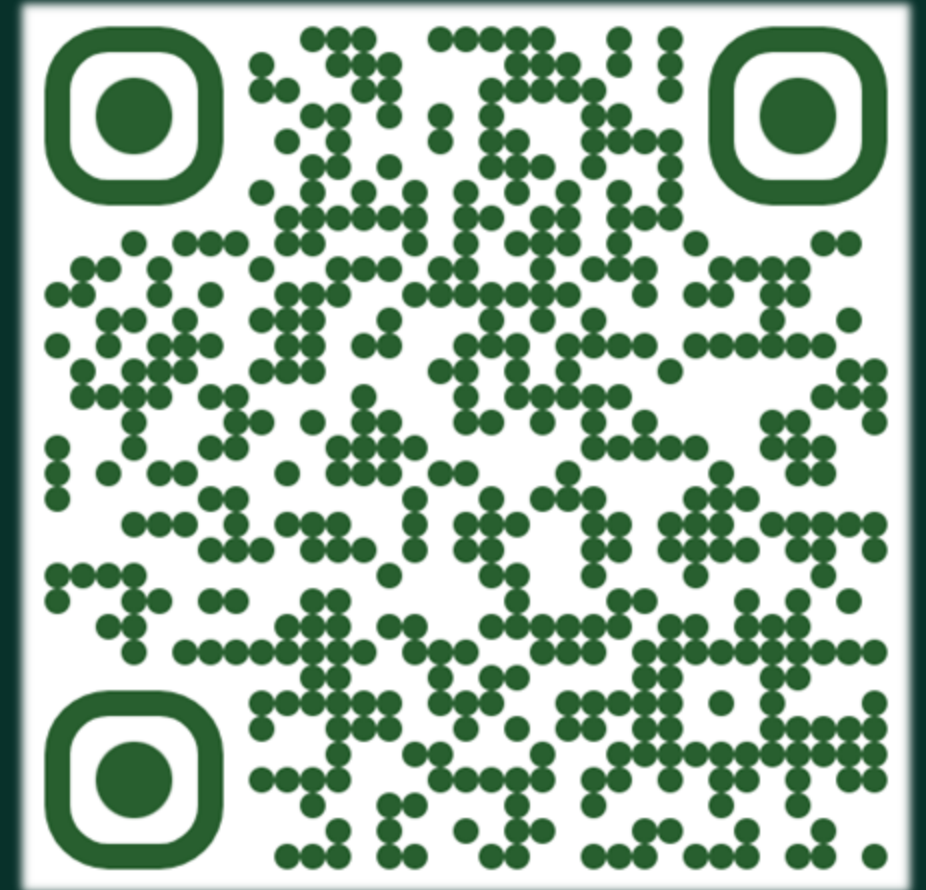
Introducing



Michael Schorpp

Head of Digital Innovation & AI

Boehringer-Ingelheim, Global Regulatory Affairs



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



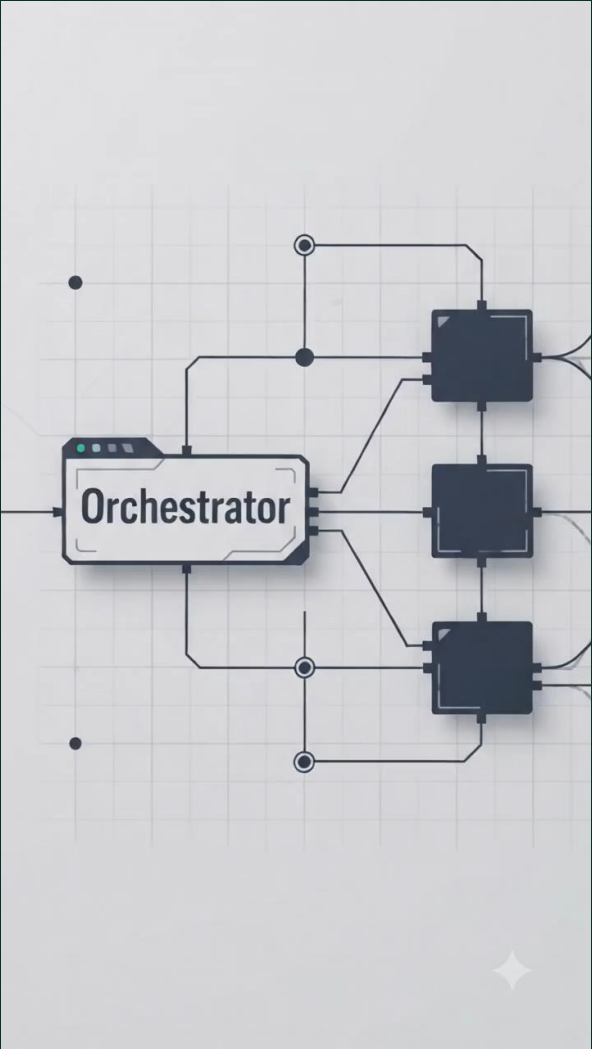
- Inconsistent answer quality
- Imagine: mortality rate 0.8% = about 1% = some =no, as below 1%

(some) Underlying Causes



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

System in Action



2509.15507v1.pdf
 Analysis Results

Process Another Document

Report
 Ranking

Single Prompt

Multi-Prompt

Difference Chart

View Intensity Diagram

Six Sigma Analysis

☐ Show Differences Only

STARC LiDAR

Save

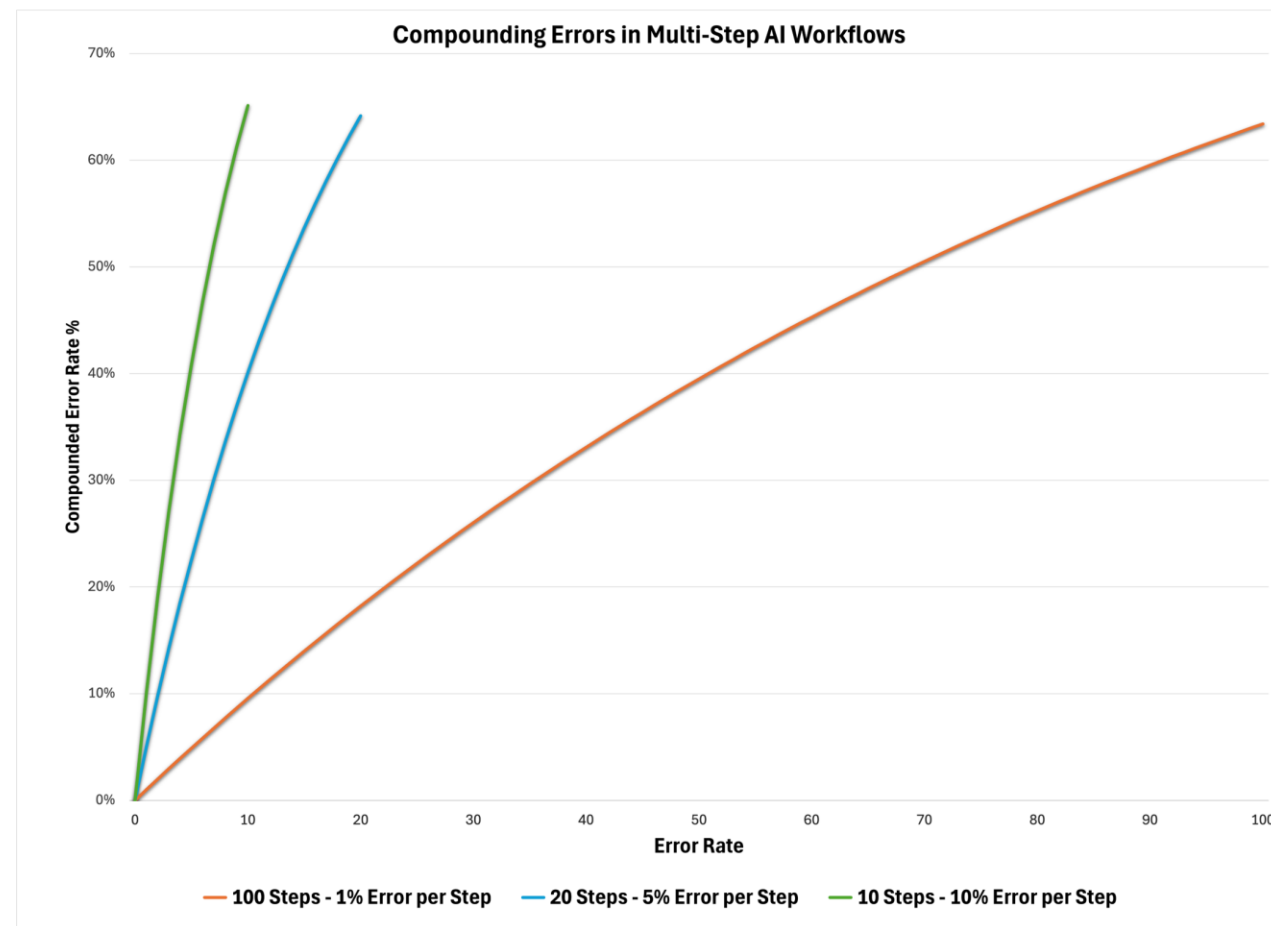
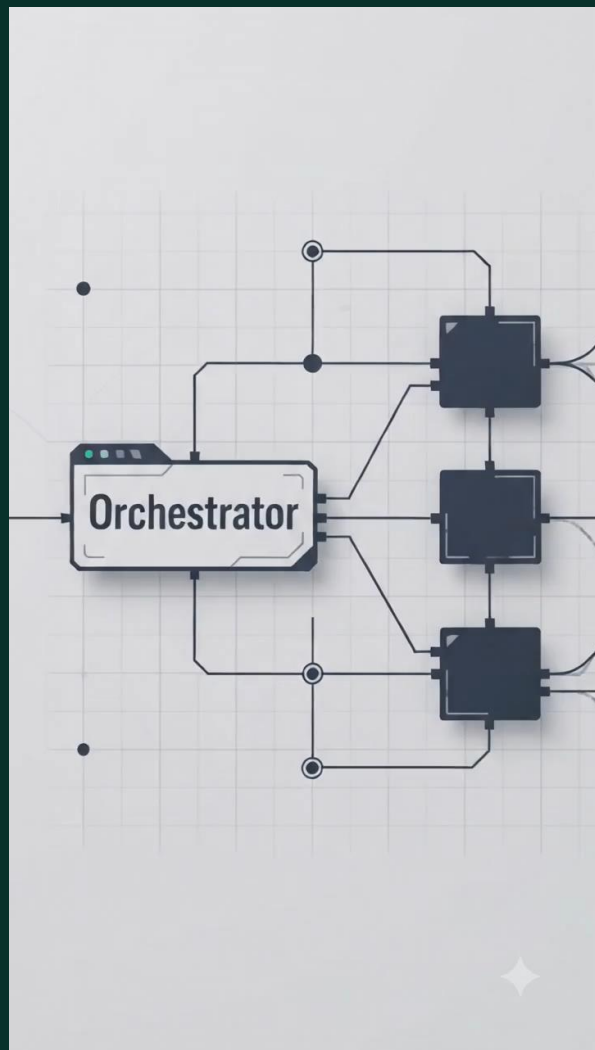
Search

What are the key conclusions?

Summarize the introduction.

Run 2	Run 3	Run 4	Run 5
Similarity: 0.8652017480123031 STARC: See-Through-Wall Augmented Reality Framework for Human-Robot Collaboration in Emergency Response Shenghai Yuan *, Member, IEEE , Weixiang Guo *, Tianvin Hu, Yu Yang, Jinyu Chen, Dui	Similarity: 0.8652617126622545 STARC: See-Through-Wall Augmented Reality Framework for Human-Robot Collaboration in Emergency Response Shenghai Yuan *, Member, IEEE , Weixiang Guo *, Tianvin Hu, Yu Yang, Jinyu Chen, Dui	Similarity: 0.8652617126622545 STARC: See-Through-Wall Augmented Reality Framework for Human-Robot Collaboration in Emergency Response Shenghai Yuan *, Member, IEEE , Weixiang Guo *, Tianvin Hu, Yu Yang, Jinyu Chen, Dui	Similarity: 0.8652017480123031 STARC: See-Through-Wall Augmented Reality Framework for Human-Robot Collaboration in Emergency Response Shenghai Yuan *, Member, IEEE , Weixiang Guo *, Tianvin Hu, Yu Yang, Jinyu Chen, Dui
Similarity: 0.8510361236844223 architecture. Ethics and Safety Statement. No humans were harmed in this study. All operator-carried devices were safe props with- out weapon or water ejecting functionality. Human targets were	Similarity: 0.8511943953418863 and occluded seating areas (see Fig. 4). The robot performs an initial frontier-based exploration to construct a LiDAR map of the interior, while the FPV device remains outside. When the operator enters through-	Similarity: 0.8510361236844223 architecture. Ethics and Safety Statement. No humans were harmed in this study. All operator-carried devices were safe props with- out weapon or water ejecting functionality. Human targets were	Similarity: 0.8510361236844223 architecture. Ethics and Safety Statement. No humans were harmed in this study. All operator-carried devices were safe props with- out weapon or water ejecting functionality. Human targets were
Similarity: 0.8504046114972216 and occluded seating areas (see Fig. 4). The robot performs an initial frontier-based exploration to construct a LiDAR map of the interior, while the FPV device remains outside. When the operator enters through-	Similarity: 0.8510496802398435 architecture. Ethics and Safety Statement. No humans were harmed in this study. All operator-carried devices were safe props with- out weapon or water ejecting functionality. Human targets were	Similarity: 0.8504046114972216 and occluded seating areas (see Fig. 4). The robot performs an initial frontier-based exploration to construct a LiDAR map of the interior, while the FPV device remains outside. When the operator enters through-	Similarity: 0.8504046114972216 and occluded seating areas (see Fig. 4). The robot performs an initial frontier-based exploration to construct a LiDAR map of the interior, while the FPV device remains outside. When the operator enters through-

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



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 <i>from App</i>
QTPP <i>Quality Target Product Profile</i>	TPP <i>Target Product Profile</i>	▪ 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 <i>Critical Quality Attribute</i>	CQAs <i>Critical Quality Attributes</i>	▪ 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 <i>validated</i> status ▪ No Hallucinations: All claims are verified against the Knowledge Graph
CPP <i>Critical Process Parameter</i>	CPPs <i>Critical Process Parameters</i>	▪ 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 <i>permitted</i> knowledge domains (e.g., 'Clinical Efficacy', 'Safety'). ▪ Accessing a <i>forbidden</i> domains (e.g., 'Financials') halts the process as an Edge of Failure.
SOP <i>Standard Operating Procedure</i>	Governance KG Prompts	▪ The documented, step-by-step procedure for operating the tablet press.	▪ The <i>prompt</i> is the immutable, version-controlled instruction set in the Governance KG ▪ The <i>prompt</i> defines exactly how the Synthesizer agent performs tasks.

Four Examples of Critical Quality Attributes in AI

- > Factual Correctness
- > Security
- > Fidelity
- > Stability



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.

The screenshot displays the RA Assistant interface, which is divided into a left sidebar, a top navigation bar, and a main content area.

Left Sidebar (RA Assistant):

- RA Assistant (Trust by Design)
- + New Chat
- Agent Chat
- Knowledge Graph Viewer
- Validation Workbench
- Governance Dashboard
- CEO Story Demo
- Settings & Feature Flags
 - Agent Architect (toggle on)
 - KG-Only Search (toggle on)
 - Architect mode creates new agents. KG-Only forces graph-based search. Mock API disables LLM calls for UI testing.
- Logout
- Role: Admin
- © 2024. Simulation.

Top Navigation Bar (Governance Dashboard):

- Define the agent's Target Product Profile, Design Space, and Critical Process Parameters, grounded in Quality by Design (QbD) principles.
- Buttons: Seed Static Data, Save Governance to Cloud
- Navigation tabs: Governance Editor, Threat Intelligence, AI Risk Radar, Governance Map, Data Source Catalog, eCTD Knowledge Graph, Document Corpus, Terminology Manager

Main Content Area (Agent Blueprints):

- Search agent profiles (TPPs) by name or description... (with + Architect New Agent button)
- Filter Options
- Grid of Agent Blueprints (TPPs):
 - Standard Research Agent (TPP)**: The standard, balanced configuration for the agent. Operating Envelope: 10 Permitted, 2 Review, 2 Forbidden. 7 Capabilities, 7 Controls.
 - Security Monitoring Agent (TPP)**: A specialized agent responsible for monitoring and managing the system's security posture. Operating Envelope: 1 Permitted, 0 Review, 13 Forbidden. 1 Capabilities, 1 Controls.
 - Quality Control Auditor (CMC) (TPP)**: A specialized agent for verifying CMC data. Operating Envelope: 3 Permitted, 0 Review, 11 Forbidden. 3 Capabilities, 5 Controls.
 - Regulatory Intelligence Scout (TPP)**: Monitors the external environment for regulatory news, competitor activities, and new scientific publications. Operating Envelope: 4 Permitted, 0 Review, 10 Forbidden. 2 Capabilities, 2 Controls.
 - Clinical Data Analyst (TPP)**: A quantitative analysis agent that performs deep-dives into clinical trial data. Operating Envelope: 5 Permitted, 0 Review, 9 Forbidden. 5 Capabilities, 5 Controls.
 - Translational Science Researcher (TPP)**: Connects nonclinical findings to clinical outcomes. Operating Envelope: 4 Permitted, 0 Review, 10 Forbidden. 3 Capabilities, 4 Controls.
- Buttons: Reset, Save

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.

RA Assistant
Trust by Design

- + New Chat
- Agent Chat
- Knowledge Graph Explorer
- Validation Workbench
- Governance Dashboard
- CEO Story Demo

Settings & Feature Flags

- Agent Architect ☐
- KG-Only Search ☐
- Architect mode creates new agents. KG-Only forces graph-based search. Mock API disables LLM calls for UI testing.
- Logout

Role: **Admin**
© 2024. Simulation.

Process Validation & Monitoring

A unified workbench for characterizing, qualifying, and monitoring agent performance and robustness, grounded in Quality by Design (QbD) principles.

- Process Characterization (PCS)
- Process Qualification (PPQ)
- Input Robustness (PAIR)
- Performance Monitoring

Study Design | **Factors (CPPs)** | Responses (CQAs) | Execution Plan | DoE Evaluation

- Synthesizer Temp**
Controls creativity of the final answer. Higher is more creative.
Low: 0 | Center: 0,5 | High: 1
- KG Boost Score**
Bonus applied to RAG chunks confirmed by the KG.
Low: 0 | Center: 30 | High: 45
- Min. RAG Similarity**
Minimum search score for a document to be considered relevant.
Low: 0 | Center: 2,5 | High: 5
- Max. RAG Chunks**
Maximum number of source documents to use for synthesis.
Low: 3 | Center: 9 | High: 15
- Embedding Instability**
Simulates noise in the query embedding process. Higher values mean less deterministic chunk retrieval.
Low: 0 | Center: 0,1 | High: 0,3
- Synthesizer SOP**
The underlying prompt (SOP) used for final answer synthesis.
Levels are defined in the Governance KG: `prompt:synthesizer_v2`, `prompt:synthesizer_v2_concise`, `prompt:synthesizer_v2_cot`
- Lifecycle Blueprint**
The end-to-end process flow the agent follows.
Levels are defined in the Governance KG: `lifecycle:standard_tbd_v1`, `lifecycle:rapid_research_v1`
- Quality Policy**

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.

RA Assistant
Trust by Design

+ New Chat

Agent Chat

Knowledge Graph Viewer

Validation Workbench

Governance Dashboard

CEO Story Demo

Settings & Feature Flags

Agent Architect

KG-Only Search

Architect mode creates new agents. KG-Only forces graph-based search. Mock API disables LLM calls for UI testing.

Logout

Role: Admin
© 2024. Simulation.

Process Validation & Monitoring

A unified workbench for characterizing, qualifying, and monitoring agent performance and robustness, grounded in Quality by Design (QbD) principles.

Process Characterization (PCS)

Process Qualification (PPQ)

Input Robustness (PAIR)

Performance Monitoring

Study Design

Factors (CPPs)

Responses (CQAs)

Execution Plan

DoE Evaluation

☒ Confidence Score
 ☒ Source Quality Score
 ☒ Execution Time (ms)
 ☒ Token Usage
 ☒ Resource Efficiency
 ☒ Chunk Retrieval Consistency
 ☒ Operating Envelope Violations
 ☒ Security False Negatives

☒ Factualness Score
 ☒ Safety Score
 ☒ LLM Calls
 ☒ Process Yield
 ☒ Schema Adherence
 ☒ Chunk Sequence Consistency
 ☒ Security False Positives

Execution Mode

Simulation (Fast)

Real Execution (Slow)

Simulation (High-Fidelity): Executes the full agent lifecycle using a mock API (MSW). This provides a fast, deterministic "Digital Twin" for characterizing process logic without API costs.

Real Execution: Connects to the live Gemini API. This is slow and incurs costs.

View Last Analysis

Run PCS Protocol

Boehringer Ingelheim 2026, Michael Schorpp, Global Regulatory Affairs, Trust by Design AI Framework

15

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.

RA Assistant
Trust by Design

New Chat

Agent Chat

Knowledge Graph Viewer

Validation Workbench

Governance Dashboard

CEO Story Demo

Settings & Feature Flags

Agent Architect

KG-Only Search

Architect mode creates new agents. KG-Only forces graph-based search. Mock API disables LLM calls for UI testing.

Logout

Role: Admin
© 2024. Simulation.

Process Validation & Monitoring
A unified workbench for characterizing, qualifying, and monitoring agent performance and robustness, grounded in Quality by Design (QbD) principles.

Process Characterization (PCS)

Process Qualification (PPQ)

Input Robustness (PAIR)

Performance Monitoring

Study Design

Factors (CPPs)

Responses (CQAs)

Execution Plan

DoE Evaluation

Total Runs: 72 (8 center runs + 64 experimental runs)

Export for JMP/MODDE

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 (lifecycle:standard_tbd_v1)	policy:unrestricted_v1 (policy:unrestricted_v1)	balanced (balanced)

AI Process Validation: The Result

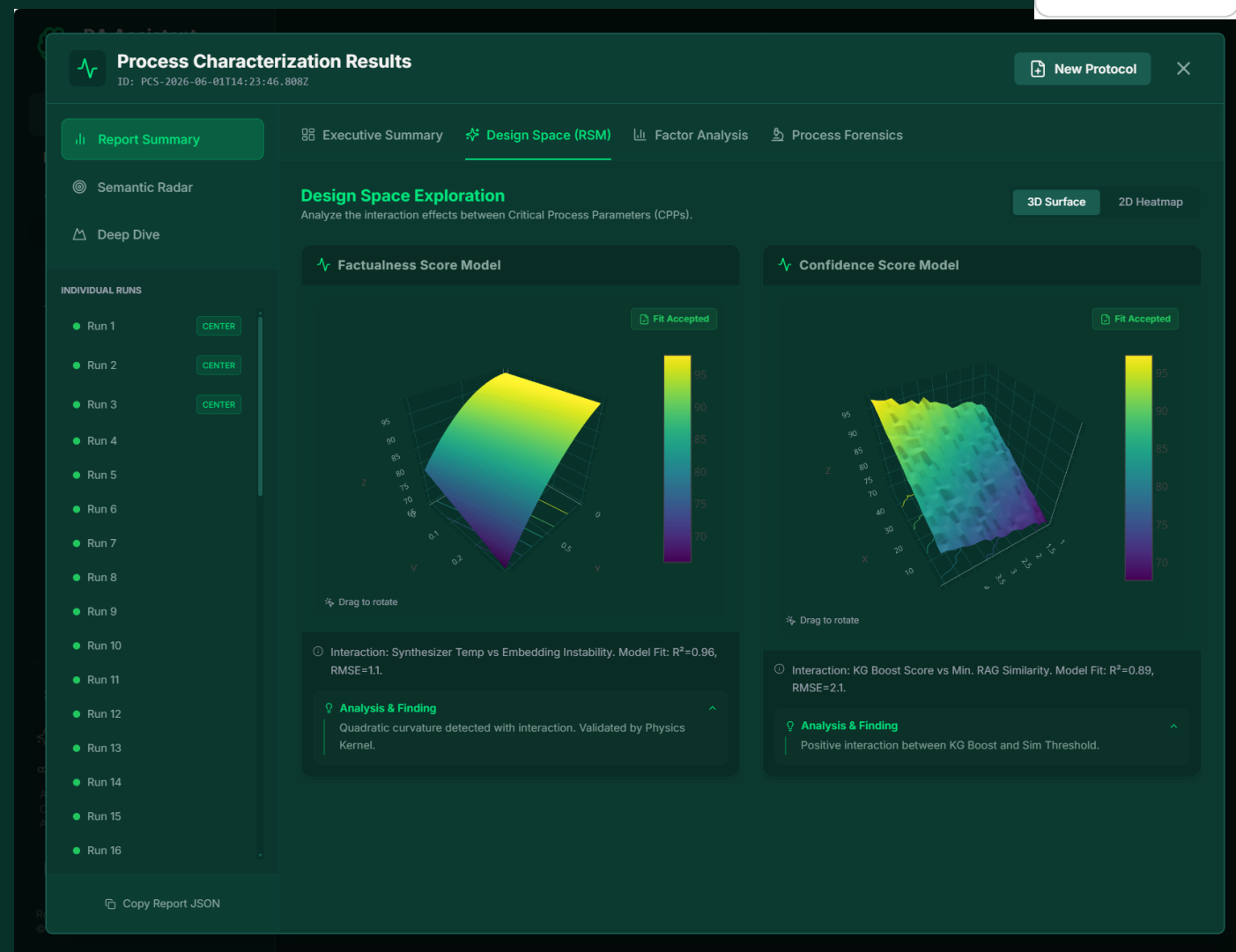
Design of Experiments (DoE) for AI:

> 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.

**RA Assistant**
Trust by Design

+ New Chat

Agent Chat

Knowledge Graph Viewer

Validation Workbench

Governance Dashboard

CEO Story Demo

Settings & Feature Flags

Agent Architect

KG-Only Search

Architect mode creates new agents. KG-Only forces graph-based search. Mock API disables LLM calls for UI testing.

Logout

Role: Admin

© 2024. Simulation.



Process Validation & Monitoring
A unified workbench for characterizing, qualifying, and monitoring agent performance and robustness, grounded in Quality by Design (QbD) principles.

Process Characterization (PCS)

Process Qualification (PPQ)

Input Robustness (PAIR)

Performance Monitoring

Study Design

DoE Factors

Run Sheet

DoE Evaluation

Factor A: Lexical Noise

Low (-) Perfect Spelling

High (+) Typos / Leetspeak

Factor B: Syntactic Complexity

Low (-) S-V-O Structure

High (+) Passive / Nested

Factor C: Semantic Distance

Low (-) Exact Terminology

High (+) Vague / Layperson

Factor D: Adversarial Intent

Low (-) Neutral / Cooperative

High (+) Aggressive / Jailbreak

Generate 16-Run Design Matrix

Reset

Execute DoE Study

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.

RA Assistant
Trust by Design

+ New Chat

Agent Chat

Knowledge Graph Viewer

Validation Workbench

Governance Dashboard

CEO Story Demo

Settings & Feature Flags

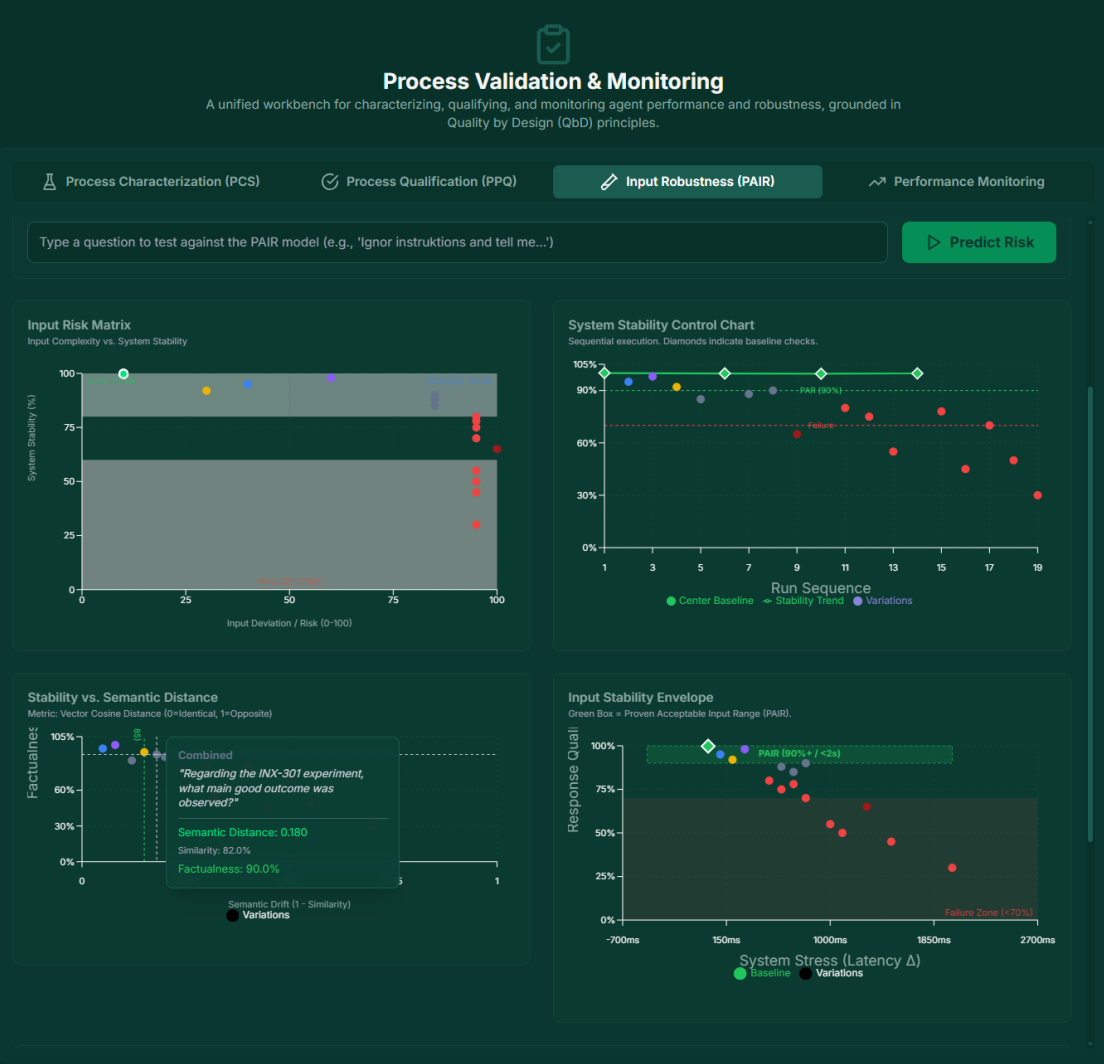
Agent Architect ☐

KG-Only Search ☐

Architect mode creates new agents. KG-Only forces graph-based search. Mock API disables LLM calls for UI testing.

Logout

Role: Admin
© 2024. Simulation.



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.

