Humigent

Based on Humigent's PMSA 2026 presentation

From Pilot Purgatory to Production Intelligence:

Why Pharma Needs a Native Language Model

A Framework for Building Trusted Enterprise Intelligence in Life Sciences

Authors

Sid Reddy
AI Innovation Leader

Barun Maskara
Partner

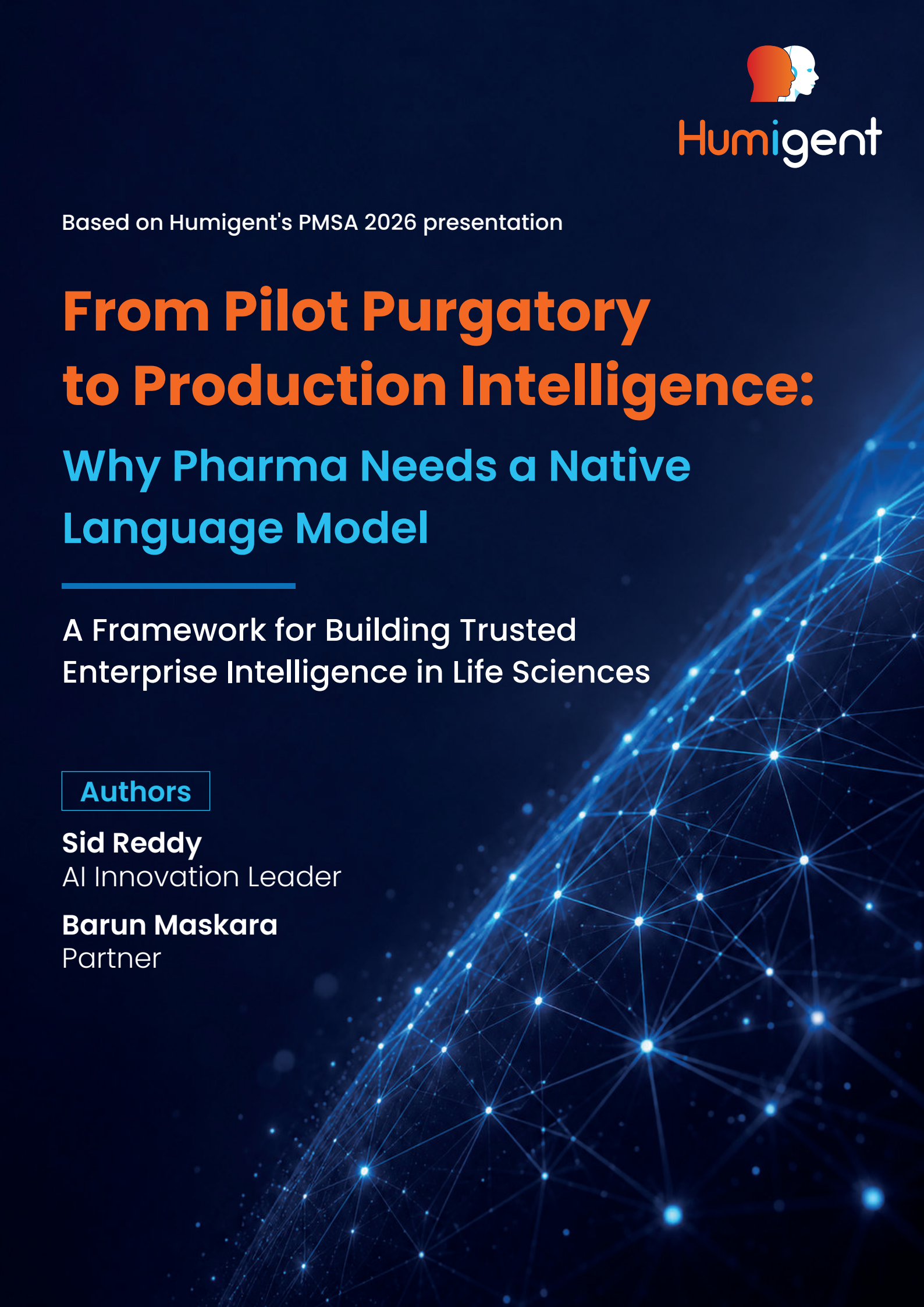


Table of Contents

About the Authors

Sid Reddy

Sid Reddy is a founding contributor to Google's Gemini, Embeddings, and NotebookLM programs - among the first enterprise-grade LLM applications deployed at scale. His career spans leadership roles at Google, Amazon, Meta, and Microsoft, with research grounded in the applied mechanics of AI, from deep learning to reinforcement learning. He holds over 120 publications and patents, a computer science degree from IIT Kharagpur, and a PhD from Arizona State University.

Barun Maskara

Barun Maskara is a life sciences strategist and technologist with more than 20 years of experience across R&D, commercial analytics, strategy consulting, and field operations. He has contributed to more than 70 life sciences patents. He holds a bachelor's degree in Electrical Engineering from IIT Kharagpur, a master's degree in Biomedical Engineering from Johns Hopkins, and an MBA in Finance from the Wharton School.

At Humigent, they are building the infrastructure layer for pharma AI that actually makes it to production.

Executive Summary

Pharma does not have an AI awareness problem. It has an AI production problem.

Across the industry, organizations are piloting LLMs, copilots, analytics assistants, and AI-powered knowledge tools. The appetite is real. Yet too many of these initiatives stall in the same place: past the demo, short of deployment. They impress in sandbox, then break on real data. This is what we call Pilot Purgatory: proof-of-concepts that work in controlled environments but cannot survive the structural complexity of enterprise pharma.

The standard explanations - not enough data, not the right model, not enough prompt engineering - miss the actual problem. Pharma AI fails because it is built on general - purpose architectures that were never designed to reason over fragmented pharma data structures, enforce domain-specific business logic, or produce auditable outputs for regulated environments.

This whitepaper examines why many Pharma AI initiatives fail to move beyond pilot programs despite significant investment, executive sponsorship, and organizational enthusiasm.

The challenge is not a lack of data, model sophistication, or prompt engineering expertise. Rather, it is the mismatch between general-purpose AI architectures and the unique requirements of pharmaceutical enterprises - where fragmented data ecosystems, complex business logic, regulatory oversight, and auditability are essential.

We argue that the next generation of pharma AI will be defined not by larger language models, but by domain-specific intelligence architectures capable of operating reliably within regulated enterprise environments. We explore the architectural principles required to achieve this and illustrate how pharma-native intelligence systems - with a domain model at the core can help organizations move from experimentation to production-scale impact.

01

The AI Inflection Point in Pharma

- Artificial intelligence has moved from experimentation to strategic priority across the pharmaceutical industry.
- Organizations are investing heavily in generative AI, copilots, analytics assistants, medical information tools, and knowledge management platforms. Commercial teams seek faster insights. Medical affairs teams need rapid access to evidence. Market access and field organizations are under pressure to improve decision-making while operating with increasing complexity.
- Yet a common pattern has emerged. While pilot programs often demonstrate promise, many organizations struggle to operationalize AI at enterprise scale.
- The challenge facing pharma leaders is no longer whether AI can create value. The challenge is determining how AI can be trusted, governed, and scaled across mission-critical workflows.
- This distinction separates experimentation from transformation.

02

The Pilot Paradox

- Industry studies and enterprise adoption surveys consistently show that an estimated 80% of pharma AI initiatives begin as a proof of concept. Of those, fewer than 30% reach production - grade deployment.
- That gap exists not because organizations lack funding, enthusiasm, or executive sponsorship - all of those are typically present. It persists because the underlying diagnosis is wrong.
- Three explanations dominate post-mortem discussions of failed pilots:

"We needed more data."

Pharma is not data-poor. It is data-fragmented. Claims, CRM, ERP, formulary, market research, patient services, call planning, and prescriptions all exist - they just live in separate systems with different schemas, identifiers, and access rules. More data in isolation does not solve a structural join problem.

"We needed a better model."

Swapping one general-purpose LLM for another does not resolve the architectural limitations that caused the failure. Every frontier model is probabilistic. The hallucination risk is non-zero by design. A different model will produce different errors - not fewer errors.

"We needed better prompts."

Prompt engineering is a patch, not an architecture. Training an entire field force to write precise technical prompts is not a sustainable production strategy. The system should understand the language of the business, not the other way around.

- The root cause of pilot failure in pharma is architectural, not linguistic. General-purpose LLMs were not built to reason reliably over complex, siloed pharma data. They cannot enforce approved metric definitions, resolve identifier conflicts across source systems, maintain temporal precision in multi-quarter analytical queries, or produce the derivation trace a compliance team can review.

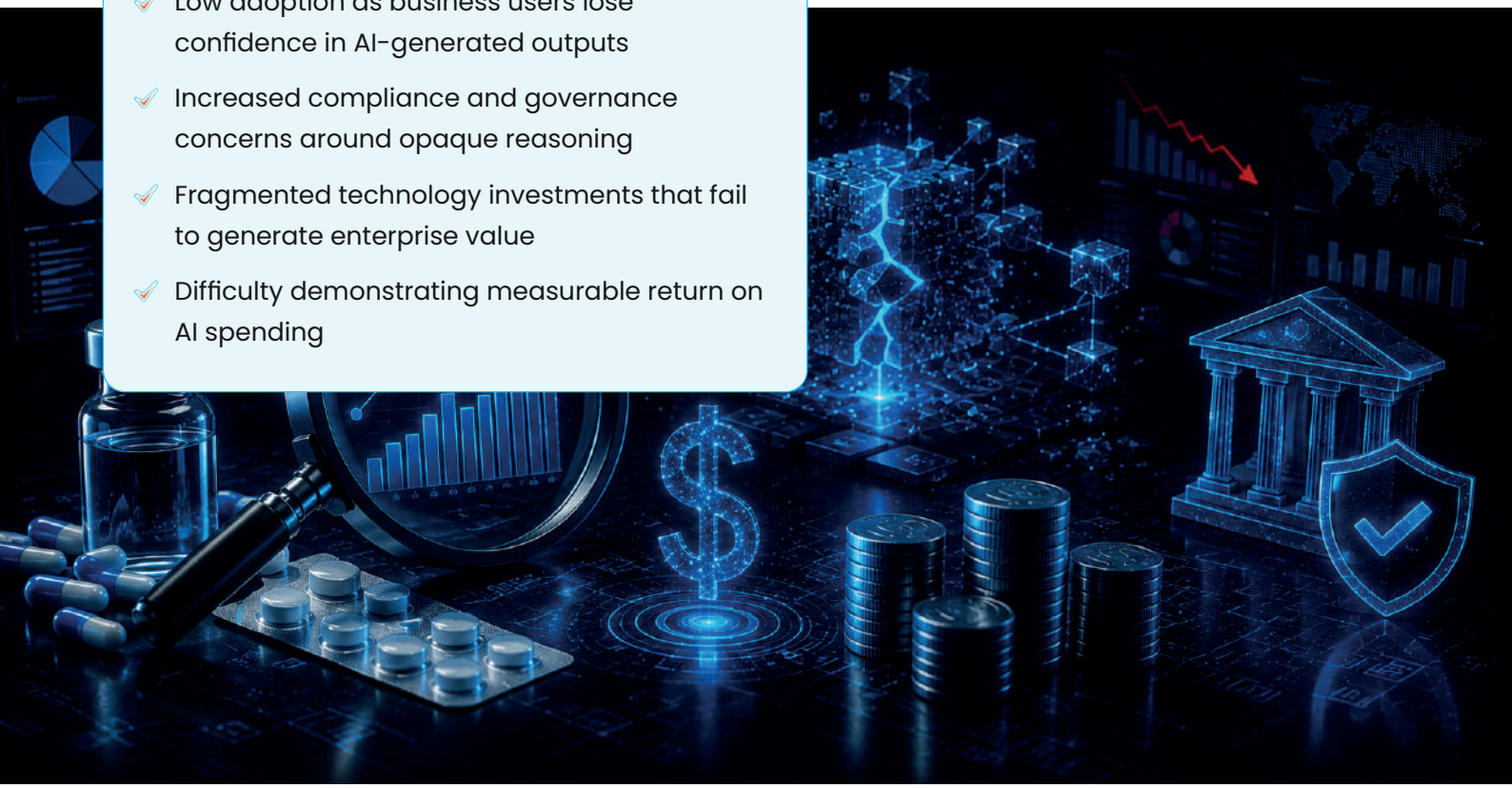
And critically: when they get it wrong, they often do not say so. They return a well-formatted answer with no indication that the logic failed. This is the silent failure problem - and it is the defining production risk for pharma AI.

03

The Hidden Cost of Pilot Purgatory and The Business Impact

The consequences of stalled AI initiatives extend beyond technology budgets. When AI programs remain trapped in pilot mode, organizations face:

- Delayed decision-making due to continued dependence on manual analysis.
- Low adoption as business users lose confidence in AI-generated outputs
- Increased compliance and governance concerns around opaque reasoning
- Fragmented technology investments that fail to generate enterprise value
- Difficulty demonstrating measurable return on AI spending



- The true cost of pilot purgatory is not the failed pilot itself. It is the opportunity cost of insights, efficiency, and competitive advantage that never reach production.
- For pharmaceutical organizations operating in highly competitive and highly regulated markets, that cost compounds quickly.

04

What General-Purpose LLMs Cannot Reliably Do in Pharma

To understand why a pharma-native model matters, it helps to see exactly where generic LLMs fail.

They are not native to pharma data structures.
Pharma data is not just text to be retrieved. It is a dense ecosystem of structured tables with overlapping identifiers, system-specific schemas, and business definitions that live in data dictionaries, not in the data itself. A model can generate plausible SQL without knowing whether the correct join key in a given context is NPI_ID, HCP_ID, or Provider_ID - and these are not

They cannot apply approved business logic.
Terms like "new patient start," "call attainment," "formulary restriction," or "line of therapy" are not just vocabulary - they are derived constructs with specific attribution windows, eligibility rules, and time period requirements that vary by brand, market, and organizational convention.
A generic model will substitute its own interpretation when the correct one is not in the prompt.

They fail silently on schema drift.
When a data vendor updates a column name in a quarterly release, a general-purpose model has no way to detect the discrepancy. It will continue to query the old schema and return confidently incorrect results.

They do not produce auditability by default.
Logging prompts and responses is not an audit trail. A production pharma system must capture source lineage, metric definitions, validation steps, join logic, access permissions, and reasoning path - not as an afterthought, but as part of how every answer is generated.

They create structural dependency.
Closed frontier models are subject to pricing changes, API deprecations, and terms-of-service modifications. For regulated enterprise workflows involving proprietary commercial data, that dependency is a governance risk, not just an infrastructure preference.

05

A new blueprint for pharma intelligence

- One approach to addressing these challenges is the development of domain-specific model designed for pharmaceutical environments. Humigent's implementation of this approach is HumigentLM, a 30-billion-parameter language model built specifically for life sciences.
- It is not a general-purpose model wrapped in a pharma user interface. Domain knowledge is baked directly into the model weights — pharma vocabulary, commercial logic, medical nuance, and analytical conventions are native to how the model understands and generates language, not added on at inference time.

Its design rests on four principles:

Pharma-native understanding.

HumigentLM is trained and tuned on domain-specific corpora: clinical trial literature, commercial data dictionaries, market access frameworks, formulary logic, field operations data, and medical information sources. When it encounters terms like TRx, NRx, NBRx, MSL inquiry, step edit, or IC attainment, it reasons with them in the way a pharma professional would - not as generic vocabulary.

Structured data awareness.

Many high-value pharma questions cannot be answered with text retrieval alone. They require reasoning over tables, schemas, metric definitions, and cross-system joins. HumigentLM is designed to support this kind of structured analytical reasoning as part of a broader system architecture - not as a standalone SQL generator, but as the domain intelligence layer within a governed pipeline.

Humigent

Production-safe deployment.

HumigentLM is designed to run inside client-controlled environments. Data does not leave the enterprise perimeter. For pharma organizations managing proprietary commercial strategy, sensitive clinical data, and regulated workflows, this is not a feature preference - it is a compliance requirement.

Cost-efficient scale.

A pharma-native model optimized for a narrower domain can be materially faster and more economical than frontier models at production scale. When a field force of thousands is generating queries across hundreds of brands and geographies, latency and cost per query are not afterthoughts - they determine whether adoption actually happens.

06

The Architecture Around the Model

- A model alone does not solve the production problem. Successful enterprise deployments combine domain intelligence, governed data access, semantic understanding, reasoning controls, workflow orchestration, and auditability into a unified operating framework.
- HumigentLM is designed to sit inside a layered system - the Humigent Pharma Cognitive Engine - that translates domain understanding into governed, auditable outputs.

The architecture has five integrated layers:

Data Layer:
Connects to fragmented source systems - CRM, claims, prescriptions, ERP, formulary, biomarkers, market research, patient services, call planning, and field activity data.

Semantic Layer:
Encodes pharma ontology, approved metric definitions, column aliases, data relationships, and compliance constraints. This layer ensures the system interprets data using the business logic of the organization, not its own inference.

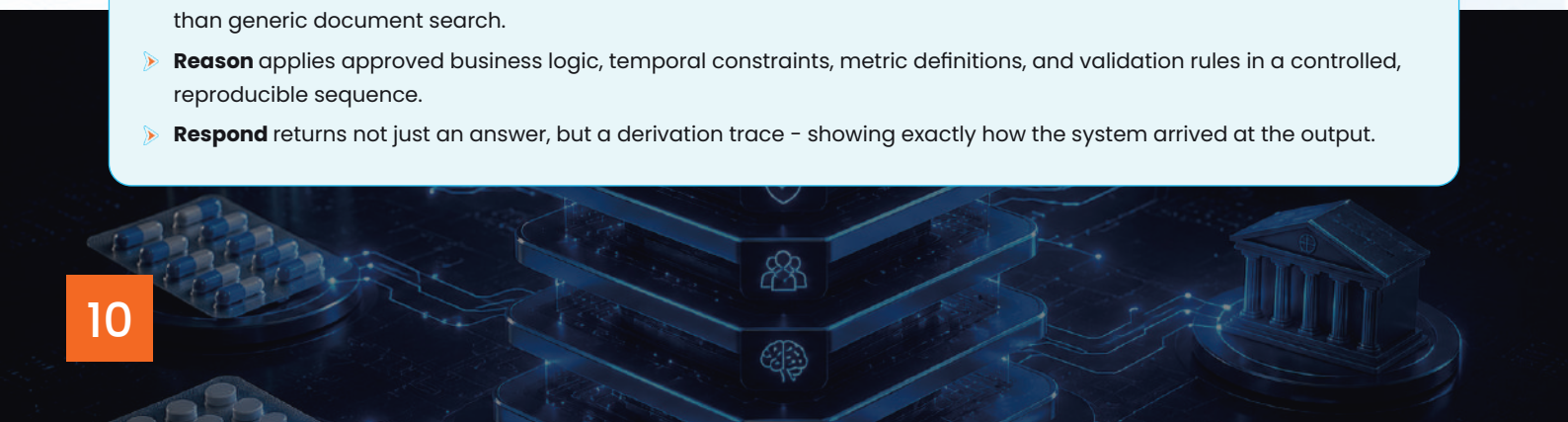
Foundation Layer:
HumigentLM provides pharma-native reasoning at the base of the AI layer. This is what distinguishes the Pharma Cognitive Engine from a generic LLM with a custom interface.

Agent and Orchestration Layer:
Specialized agents handle defined functions - schema resolution, data validation, authentication and authorization, summarization, audit, and workflow-specific reasoning. An orchestrator routes, prioritizes, and sequences these agents so the system behaves as

Output and Audit Layer:
Every response includes a confidence score, source lineage, reasoning path, validation checks, and a complete audit log. This transforms an AI response into an enterprise artifact that compliance teams can review and analytics teams can defend.

The R³ framework - Retrieve, Reason, Respond - governs how HumigentLM operates within this system:

- **Retrieve** grounds the task in the right enterprise data, informed by pharma vocabulary and data relationships rather than generic document search.
- **Reason** applies approved business logic, temporal constraints, metric definitions, and validation rules in a controlled, reproducible sequence.
- **Respond** returns not just an answer, but a derivation trace - showing exactly how the system arrived at the output.



07

Where Generic AI Breaks: A Production Example

- The clearest way to understand the value of this architecture is to examine a query that appears straightforward but breaks most general-purpose implementations.
- "Which of our top 50 HCPs by TRx volume saw a decline in new patient starts last quarter, despite an increase in rep call frequency?"
- This is a routine question for a brand team or field analytics leader. But answering it correctly requires:
  - A three-table join across CRM, claims, and field force planning data
  - Correct HCP identity resolution across systems that may use NPI_ID, HCP_ID, and Provider_ID as non-equivalent identifiers
  - Calculation of "new patient starts" as a derived metric with a specific attribution window - not a raw row count
  - Temporal logic that defines "top 50 HCPs" based on the prior quarter's TRx volume, not the current quarter
  - Quarter-over-quarter delta calculations across separate fact tables
  - Validation that directional claims in the output are numerically consistent

A generic LLM given this query will typically:

- Select an incorrect join key, producing HCP mismatches without flagging the error
- Calculate new patient starts as a raw claim count, ignoring the 90-day attribution rule
- Rank HCPs by current-quarter volume rather than prior-quarter volume
- Return a polished, formatted table that looks authoritative and is factually wrong

- This is the silent failure problem in its most dangerous form: incorrect analysis delivered with high confidence and no error signal.
- Here is how the Pharma Cognitive Engine handles the same query:
  - The schema agent queries the data catalog, identifies NPI_ID as the correct join key across all three source systems, and resolves the identifier alias conflict before the query runs. The semantic layer supplies the approved definition of "new patient start" including the correct attribution window. The reasoning agent applies that definition to the claims data. The retrieval agent identifies the prior-quarter TRx cohort. The validation agent checks directional consistency in the output. The audit agent logs the full derivation trail.
- The result is not just a correct answer. It is an answer with proof.

08

Evaluation: What Actually Matters for Pharma AI

Generic AI benchmarks were not designed for pharma. A model that performs well on standard reasoning tasks may still fail on commercial analytics questions that require domain-specific join logic, derived metrics, and multi-system data integration.

Humigent evaluates AI performance across dimensions that matter in production pharma settings:

Faithfulness

Does the answer stay grounded in the actual source data?

Answer Relevancy

Does it address the question the user actually asked?

Structured Reasoning

Can it handle multi-step analytical queries with correct logic sequencing?

Schema Awareness

Can it navigate complex, multi-system data without identifier errors?

Latency

Is it fast enough for real-time workflow adoption?

Cost Efficiency

Can it scale economically across thousands of daily queries?

Privacy and Deployment Control

Can it run inside client-controlled environments?

Auditability

Does the answer stay grounded in the actual source data?

Compliance Readiness

Can legal, medical, regulatory, and risk stakeholders review the output path?

Early internal evaluations on pharma-specific market research and commercial analytics tasks show meaningful improvements over general-purpose models across these dimensions. Additional benchmarking is currently underway across broader datasets, use cases and client environments. Formal benchmark results will be published as validation efforts expand.

09

The Strategic Case for a Pharma-Native Model

The argument for HumigentLM is not purely technical. It is strategic.

Reduced dependency on closed platforms.
Frontier models are powerful, but they are controlled by third parties. Pricing, deployment options, fine-tuning access, and model behavior can all change. A pharma-native model gives life sciences organizations more control over a critical operational capability.

Client-specific adaptation.
Different therapeutic areas, brand portfolios, commercial models, and governance environments require different configurations. A domain model and architecture can be adapted to client-specific data, metrics, and workflows without routing every use case through a generic external system.

Economics at scale.
Pilot economics are not production economics. Once AI becomes embedded in daily decision workflows, cost per query and system responsiveness become business-critical. A purpose-built model optimized for a narrower domain can offer substantially better economics for high-volume operational use cases.

Compounding domain advantage.
A general-purpose model is broad by design. A domain model can become progressively deeper. As more pharma workflows are validated, more reasoning patterns are encoded, and more client environments contribute to model refinement, the gap between a domain model and a generalist grows - in favor of the domain

Organizational Trust
Enterprise AI adoption is ultimately a trust challenge. Users must trust the outputs. Compliance teams must trust the controls. Leadership teams must trust the governance model. Architectures that prioritize transparency, auditability, and explainability are more likely to achieve sustained adoption than those focused solely on model capability.

10

Questions Pharma Leaders Should Ask Before Choosing an AI Architecture

When evaluating AI solutions for production deployment, the right question is not "Which model are you using?" The right questions are:

01 Is the model pharma-native, or is it a general-purpose model with a pharma interface?

02 Does it understand structured pharma data and the business logic behind approved metrics?

03 Can it run in a client-controlled environment, with data never leaving the enterprise perimeter?

04 How does it handle schema ambiguity and changing data definitions across vendor updates?

05 How does it distinguish raw data fields from derived metrics with attribution rules?

06 How does it validate multi-step analytical logic before returning an answer?

07 Does every answer include source lineage, reasoning path, and a complete audit log?

08 How are access control, privacy, and compliance constraints enforced at the system level?

09 What happens when the system is uncertain — does it flag uncertainty or generate a confident wrong answer?

10 Can the architecture extend across multiple functions, brands, and workflows without rebuilding from scratch?

The answers to these questions determine whether an AI initiative will remain a pilot or become a production intelligence system that analysts, compliance teams, and commercial leaders can actually trust.

11

The Bottom Line

- The pharma industry has largely moved beyond asking whether AI can generate value. Business leaders, analytics teams, medical affairs, and commercial functions have all seen what AI can do in a sandbox. The more important question is whether AI can be trusted within the realities of enterprise operations.
- Production AI requires more than a chatbot, more than prompt engineering, and more than access to the most powerful frontier model available. It requires a pharma-native foundation, a governed semantic layer, controlled reasoning, and transparent outputs.
- Organizations that solve these challenges will move beyond isolated pilots and unlock AI as a scalable enterprise capability. Those that do not risk remaining trapped in a cycle of experimentation without impact.
- The future of pharmaceutical AI will belong not to the organizations with the largest models, but to those with the most trusted intelligence architectures.

Next Steps

- Organizations evaluating AI for commercial, medical affairs, market access, or enterprise analytics should assess whether their current AI strategy is designed for experimentation or for production.
- Humigent works with life sciences organizations to design and operationalize production-ready AI systems built around pharma-native intelligence.
- To learn more, schedule an executive briefing or *request a demo* or contact us at info@humigent.ai

Humigent

About Humigent

Humigent enables Life Sciences organizations to unlock the full potential of AI through intelligent data, advanced analytics, and purpose-built solutions. We help commercial teams make faster, smarter, and more confident decisions while preserving the human judgment that drives impact.

US Office

33 S Wood Ave Suite 660, Iselin, NJ 08830

India Office

948, 9th floor, Tower B1, Spaze I-Tech Park, Sector 49, Sohna Road, Gurgaon, Haryana, 122 018