KOMODO

WHITE PAPER

Benchmarking Marmot Against Foundational Models

How Komodo Health's Specialized Healthcare AI Agent Improves Analytic Performance

March 2026

Table of Contents

Disclosure Statement

Executive Summary

Generative AI offers significant promise for healthcare analytics, yet its impact is often restricted by a critical context gap. While foundational large language models (LLMs) demonstrate high fluency — the ability to generate natural, human-like text by predicting the next token in a sequence — this does not equate to logical reasoning or true domain understanding. Consequently, LLMs are frequently described as "fluent, but not intelligent," because they excel at identifying language patterns, but they fall short of the domain expertise, data access, and methodological precision necessary for high-stakes analytics that directly impact patient care.

Relying on general-purpose LLMs for tasks that impact patient care is a strategic risk. A standard LLM can generate a speculative memo on how a market event might impact brand performance, but it cannot reliably illustrate real-world impacts or account for the technical shortcomings of claims data without generating methodological hallucinations.

This study benchmarks Marmot™, Komodo Health's specialized AI agent, against a leading foundational LLM. Across core commercial workflows, Marmot consistently demonstrated a **2.3x improvement in accuracy** compared to this baseline foundational model. These findings prove the hypothesis that to deliver reliable results, AI agents require more than just data access; they require a specialized semantic layer that encodes the specific business rules of healthcare to increase their usefulness as a strategic tool in the Life Sciences space.

Introduction: Today's Challenges in Healthcare AI

The current generation of LLMs has solved the problem of fluency. However, most questions in Life Sciences analytics require multi-step logic, such as cohort construction, temporal filters, and Line of Therapy (LoT) derivation, where small errors compound quietly. This presents a fundamental challenge: using probabilistic models to behave like precise, auditable computational engines in environments where silent errors carry significant commercial and regulatory risk.

Generic AI models often struggle to coordinate multi-step tool execution under token and context limits while failing to account for complex business rules. Marmot addresses this challenge by using LLMs within a tightly controlled analytical framework, where governed assets, such as canonical codesets, and deterministic execution harnesses explicitly bound how the model generates insights and interacts with data.

The Fundamental Challenge: Beyond Basic Fluency

Healthcare analytics presents unique challenges that pattern recognition alone cannot solve:

- **Data Complexity:** Healthcare data is fragmented, incomplete, non-standardized, and longitudinal.
- **Market Dynamics:** Even widely accepted industry use case analytics require unique business rules and inclusion criteria to accurately represent the specific market basket of interest.
- **The Hallucination Risk:** In healthcare analytics, a plausible but inaccurate finding is more dangerous than no answer at all, potentially leading to multimillion-dollar misallocations.

Bridging the Gap: Architecting an Analytic Specialist

To achieve meaningful performance gains, Marmot was engineered with several healthcare-specific customizations that move beyond the limitations of raw LLM reasoning:

- **System Prompt Customization:** Marmot's system prompt contains significant metadata regarding Komodo's Healthcare Map®, including schema overviews, coverage details, and data considerations. This ensures the agent understands the nuances of the data that it queries.
- **Specialized Tooling:** Marmot is equipped with a library of custom tooling, including code search, query optimization, and specialized skills, allowing for a systematic approach that is grounded in subject matter expertise.

- **Advanced Agentic Orchestration:** The architecture utilizes parallel execution, interleaved thinking, and context engineering to manage token and context limits. Specialized sub-agents coordinate multi-step tool execution.
- **Validation and User-Guided Features:** Marmot leverages tools like PubMed search for external validation and Research Planner. The Research Planner is key to ensuring user control over the analysis, as it proactively asks questions, shares methodology options, and flags potential issues needing clarification rather than proceeding with flawed assumptions.

Marmot's Comprehensive and Complete Data Foundation

Marmot is anchored in Komodo's Healthcare Map, which aggregates real-world data characterizing patient health status and U.S.-based health service utilization between 2015 and 2026 (current) for a U.S.-insured population of an estimated 330 million lives. This longitudinal fabric is built from 60+ highly curated and continuously refined data sources and features de-duplicated individual claim events with NPI-level granularity for account-level analyses.

The PLAID Schema

Marmot is integrated with PLAID (Patient-Level Analytics and Insights Derivative), Komodo's standard commercial data schema. This standardized structure allows the agent to navigate seven core tables with deterministic precision:

- **Medical and Pharmacy Events:** De-duplicated Mx and Rx events (at the service line level) providing a complete view of patient treatment.
- **Provider and Plans Reference Tables:** Granular healthcare provider (HCP) and organization (HCO) information, including specialty and practice location, alongside payer-level details.
- **Patient Context Tables**: Longitudinal enrollment data, demographics (age/gender), and geographic attributes (ZIP3) for all patients over specific time intervals.

Komodo Drug Projections

For complex use cases like Market Share, the agent also utilizes data from Komodo Drug Projections (KDP). KDP provides volume estimates for over 10,000 drugs, covering both medical and pharmacy benefits. By combining projected and observed data, Marmot can account for the inherent fragmentation in U.S. healthcare data, providing a scaled national view of drug utilization that foundational models cannot replicate through simple claim counts.

Methodology: The Agentic Benchmark

To fairly evaluate performance, we moved beyond simple Q&A tests to what the industry recognizes as "Agentic Benchmarks." Effective evaluation requires testing an agent's ability to navigate environments and execute workflows, not just retrieve facts. We applied this principle to healthcare analytics.

Experimental Groups

We tested three distinct agent configurations to isolate the value of specialized tooling, domain knowledge, and data access for quantitative healthcare analysis:

1. **Tier 1 (The Baseline):** Foundational LLM (Claude Sonnet 4.5) via standard API with no tools or supplemental data
2. **Tier 2 (The Intermediate):** Foundational LLM (Claude Sonnet 4.5) with raw access to Komodo's Healthcare Map but lacking curated tooling or specific healthcare context
3. **Tier 3 (The Specialist):** Marmot, leveraging Komodo's Healthcare Map, custom tooling, and specific system prompts/guardrails

Evaluation Use Cases

We selected three distinct high-complexity use cases designed to stress-test specific healthcare analytical skills, ranging from cohort definition to complex attribution logic:

1. **Market Share (NBRx):** Determining new-patient-start market share, requiring appropriate market basket identification and trend analysis over time.
2. **HCP/HCO Targeting:** Identifying and ranking top prescribers, requiring multi-layered filtering (Specialty, Region, Decile) and complex schema navigation.
3. **Survival Analysis:** Estimating the probability of treatment discontinuation using Kaplan-Meier statistical methods, requiring the ability to correctly identify censored data points within a patient journey.

Grading Criteria (The Golden Dataset)

Outputs were scored against a "Golden Dataset," a version-controlled answer key consisting of ~140 criteria across 14 unique evaluations, validated by Subject Matter Experts. To ensure a holistic assessment, we utilized a three-part evaluation framework for every test case:

1. **Analytic Plan:** Did the agent choose the correct data source and disclose business rules (e.g., lookback periods)?
2. **Expected Output:** Did the final patient counts or provider lists match the validated benchmark?
3. **Output Insights:** Did the agent provide contextual explanations and flag data limitations?

Results: A Case Study in Specialization

Understanding the Metrics

To quantify performance, we utilized an automated LLM as a judge to evaluate every response against a rigorous checklist of verified criteria derived from the Golden Dataset.

- **The Experiment:** For each use case, we executed the same prompt multiple times across all three models.
- **The Scoring:** The judge evaluated every response against specific, binary criteria (e.g., "Did the agent select the correct data table?" "Is the calculated value within 5% of the ground truth?").
- **The Percentage:** The reported Accuracy Score is the average percentage of criteria passed across all experimental runs.

Overall Findings

Marmot (Tier 3) consistently achieved the highest accuracy, averaging ~80% across all tasks. Providing raw data access (Tier 2) improved performance over the baseline but still resulted in a failure rate of ~44%. This demonstrates that access to data is necessary but not sufficient without the semantic tools to interpret it.

Figure 1. Chart of Comparative Accuracy by Use Case for Each Agent Configuration

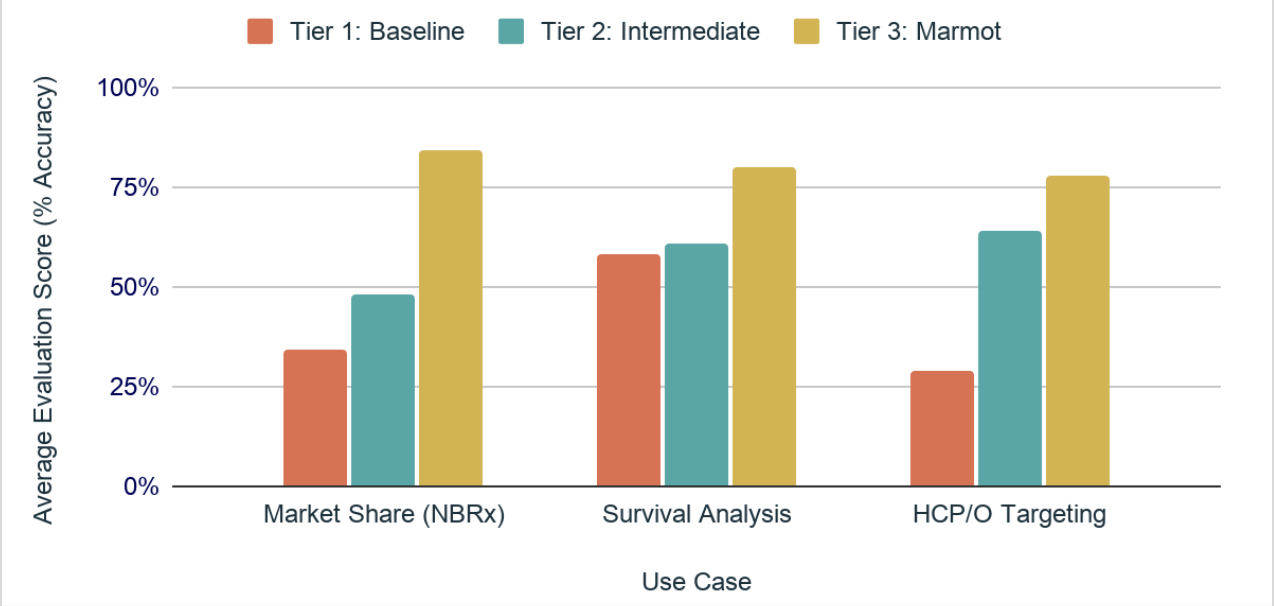

Figure 2. Table of Comparative Accuracy by Use Case for Each Agent Configuration

Use Case	Count of Unique Evaluations	Tier 1: Foundational	Tier 2: Intermediate	Tier 3: Marmot
Market Share (NBRx)	4	34%	48%	84%
HCP/O Targeting	8	29%	64%	78%
Survival Analysis	2	58%	61%	80%

The Prevalence of Methodological Hallucinations

The most significant driver of the performance gap was not a lack of general intelligence, but rather a phenomenon commonly referred to as "Methodological Hallucination." While foundational models can often retrieve data and perform calculations, they consistently fail to apply the specific, unwritten business rules required for valid healthcare analytics. This failure mode manifested in two primary ways across our high-complexity use cases:

1. The Volume Versus Value Trap (HCP Targeting)

HCP/HCO Targeting represented the widest performance chasm (Marmot 78% versus Baseline 29%). This is because "Targeting" is not a universal math problem; it is a multi-step, proprietary workflow that requires layering specific business rules. Foundational models defaulted to generic heuristics, equating "High Volume" with "High Value." Without specific context injection, they failed to apply critical filters:

- **Missing Benefit Design:** They often counted only prescription claims rather than filtering for the specific benefit type (e.g., distinguishing between Prescription Drug versus Medical Benefit for infusion therapies).
- **Ignoring Context:** They frequently recommended high-volume prescribing Primary Care Physicians for specialized therapies, failing to identify the true decision-making specialist. They also didn't account for coverage gaps, which may elevate physicians inappropriately.
- **The Result:** A target list that was mathematically "correct" based on raw counts, but not actionable for a commercial sales team.

2. The Fluency Illusion (Market Share)

In Market Share analysis, foundational models struggled to generate visuals overall. While they did generate plausible-looking charts on occasion, they masked fundamental logic errors:

- **The Fluency Illusion:** When generated, outputs looked authoritative, charts had the right drug names and sensible axes.
- **The Logic Failure:** Under the hood, the agent neglected to consider nuances that were crucial to the fundamental accuracy of the analyses. The agent incorrectly calculated market share based on the total number of claims (raw transaction count) rather than adjusted dispenses (normalized day supply). This distinction is vital, particularly when comparing treatments like oral oncolytics (typically a 90-day supply) with IV therapies (1-day supply). Furthermore, when tasked with an indication-specific market share analysis for multiple sclerosis treatments, the agent recognized the need

for an attribution factor for the multi-indication drug rituximab; however, it lacked the necessary mechanism to implement this filtering methodology correctly.

- The Result: A seemingly impressive visual representation (though not always in the requested format) that, in reality, significantly mischaracterized the true competitive landscape.

Standardization Closes the Gap: Survival Analysis

In standardized statistical tasks like Survival Analysis (Kaplan-Meier), the performance gap between Marmot (80%) and the Foundational Baseline (58%) was narrower than in other areas.

- **Why Foundational Models Performed Decently**: The mathematical formula for a Kaplan-Meier curve is universal. It is well-documented in the public internet data used to train foundational models (e.g., Wikipedia, medical textbooks). Therefore, the LLM "knows" the math.
- **Where Foundational Models Fell Short**: Even knowing the formula, foundational models struggled with data retrieval. They often failed to correctly identify "censored" patients within the specific schema of the Healthcare Map, leading to incorrect inputs for the correct formula.
- **Marmot's Edge**: Marmot's advantage here wasn't just knowing the math, but its ability to correctly query the underlying event tables to feed that math.

Discussion: Why Context Matters

Our findings support the idea that for specialized tasks, the appropriate context is more critical than the raw intelligence of a language model. Even analyses performed by humans can yield plausible yet inaccurate conclusions when the appropriate context is not applied, and the same applies to AI-enabled ones. Marmot's superior performance over Foundational LLMs can be attributed to our "Five Layers of Defense" architecture, which wraps the LLM in necessary context.

- **The Data Layer: Komodo's Healthcare Map**
  - Foundational models rely on training data which typically cuts off at a certain date, or a generic web search.
  - *The Marmot Advantage:* Marmot is anchored by Komodo's Healthcare Map, a near-real-time, longitudinal dataset of 330M+ patient journeys. It doesn't "guess" the number of patients; it counts them. This ground truth is the foundation of accuracy.
- **The Semantic Layer: Komodo's Curated Tools**
  - Tier 2 (The Intermediate) often fails because it tries to write raw SQL against complex, normalized database schemas. It doesn't know how to join table A to table B correctly.
  - *The Marmot Advantage:* Marmot doesn't ask the LLM to write raw code from scratch for every task. We provide vetting tooling/skills for complex tasks like HCP Targeting or Market Share. The LLM only needs to select the right tool; the tool handles the complex math and logic.
- **The Guardrail Layer: Komodo's Prompts**
  - Foundational models will try to answer any question, even if it requires dangerous assumptions (e.g., inferring an off-label indication, hallucinating information).
  - *The Marmot Advantage:* Marmot is injected with strict system prompt components that enforce compliance and business logic. It knows not to infer indications, not to use prohibited commercial terms, and always to disclose limitations. These guardrails prevent "methodological hallucinations" before they happen.
- **The Presentation Layer: Komodo's Context**
  - Foundational models will take liberties in data visualization and interpretation of insights, causing confusion to users.
  - *The Marmot Advantage:* By integrating user-defined business needs with charting tools, Marmot ensures that the resulting data insights and visual representations are precisely tailored to customer requirements.

- **The Stability Layer: Komodo's Benchmarking**
  - Tier 2 (The Intermediate) often struggles when relying directly on foundational frontier models. As organizations move from one model to the next, outputs shift — requiring repeated change management, recalibration of expectations, and time spent adapting to new result patterns. This creates ongoing operational friction and inconsistency in insights.
  - *The Marmot Advantage:* Marmot doesn't simply swap from one frontier model to another. It thoughtfully introduces and governs model updates and performs regression testing, so results remain stable and consistent over time. This reduces the burden of change management on users and ensures that improvements in model capability don't come at the cost of reliability. It's not just about adding context — it's about managing change intelligently, so outcomes don't unexpectedly vary.

Limitations

While these results demonstrate a significant performance advantage for specialized agents, several limitations must be noted:

- **Scope of Therapeutic Areas:** This benchmark focused on a limited subset of therapeutic areas, including rheumatoid arthritis, multiple sclerosis, atrial fibrillation, and breast cancer. Performance may vary in therapeutic areas with highly specialized or rare disease coding requirements.
- **Commercial Use Case Selection:** This study was restricted to three specific commercial workflows: Market Share, HCP Targeting, and Survival Analysis. While representative of core analytics, these do not encompass the full breadth of healthcare commercial use cases.

Conclusion

While foundational models are rapidly advancing, this study demonstrates that for high-stakes healthcare analytics, they cannot yet replace specialized, integrated agents. The optimal path forward is not to choose between "AI" and "Data," but to integrate them. Marmot represents this convergence, where general intelligence is anchored by ground-truth data and proprietary business logic to deliver results you can trust.

About Komodo Health

Komodo Health is a technology platform company creating the new standard for real-world data and machine learning that connects the dots between individual patient journeys and large-scale analytics by pairing the industry's most complete view of patient encounters with enterprise software health outcomes. Across Life Sciences, payers, providers, and developers, Komodo helps its customers unearth patient-centric insights at scale — marrying clinical data with advanced algorithms and AI-powered software solutions to inform decision-making, close gaps in care, address disease burden, and create a more cost-effective, value-driven healthcare system. For more information, visit Komodohealth.com.