

# **Breaking the Celebrity LLM Monopoly:**

**A Comprehensive LLM Performance  
Report**

# Table of Contents

**1**

## **Escape the "Premium" LLM Trap**

Stop defaulting to expensive celebrity LLMs for every task and workflow to eliminate massive budget leakage before it destroys your AI ROI.

**2**

## **Right-Sizing Work for Maximum Impact**

Intelligently segment enterprise tasks by complexity to match each one with the fastest, highest-quality, and lowest-cost model available.

**3**

## **Rigorous Testing That Reveals the Truth**

Explore our standardized evaluation pipeline, 2,800+ real enterprise records, and unbiased LLM judging that levels the playing field across 20+ models.

**4**

## **Proven Results: Premium Performance at a Fraction of the Cost**

See benchmark evidence showing 88–95% pass rates, 4–6× faster latency, near-zero formatting errors, and 94–98% cost savings on routine and agentic workloads.

**5**

## **Dynamically Controlling LLM Costs**

Implement strategic LLM orchestration to break free from costly hype, secure your data, and build scalable AI workflows.

# Executive Summary

In the current enterprise landscape, organizations are falling into a "**Premium Trap**," defaulting to expensive, "celebrity" Large Language Models (LLMs) for every task regardless of complexity. This lack of discernment leads to massive budget leakage, as top-tier proprietary models are frequently over-engineered for routine, high-volume tasks. Using a "supercar to deliver mail across the street" is impressive, but it is not a scalable economic strategy.

## The Rise of the Krista LLM and the 80/20 Rule

Our comprehensive evaluation of over 20 open- and closed-source, open-source, and task-optimized models proves that "Celebrity AI" is often unnecessary for foundational enterprise workflows. We have identified that **roughly 80% of enterprise workloads**—including routine data processing, agentic automation, and standard Q&A—can be handled with near-parity accuracy by the **Krista LLM**.

- **Near-Parity Performance:** Krista LLM achieved an **88.5% pass rate** in generic data tasks and a **94.9% pass rate** in agentic AI, matching closed-source leaders in critical metrics like structural integrity and formatting reliability (0.00% error rate).
- **Latency Advantage:** Krista LLM is up to **4x to 6x faster** than premium proprietary models, delivering responses in as little as **1.15 to 1.49 seconds**.
- **Drastic Cost Reduction:** By right-sizing workloads, organizations can achieve **94–98% of the quality** of "celebrity" models at a fraction of the cost—often **1% of the price**.
- **Ultimate Security:** Krista LLM is completely contained inside of your private instance offering the highest level of privacy and security.

## Strategic Orchestration: The Krista "Conductor"

True operational efficiency requires moving beyond a single-model strategy to **Strategic LLM Routing**. Krista serves as the "**conductor**" for your orchestrated AI ecosystem, automatically selecting the best model for each unique request based on intent, required precision, and cost.

- **Krista LLM as the Efficiency Engine:** Routine, high-volume tasks are automatically routed to the Krista LLM to maximize speed and ROI.
- **Strategic Specialist Routing:** Premium proprietary models (like gpt-5.2 or gemini-3-pro-preview) are reserved as "soloists" for the remaining **20% of tasks** that require absolute 100% fidelity or specialized reasoning.

By adopting this task-specific routing strategy, enterprises can move beyond expensive "science projects" and build a high-ROI, production-ready AI foundation that scales across the entire organization without exploding operational costs.

Stop overpaying for brand-name hype. Audit your workloads, adopt orchestration, and start reclaiming your AI ROI today.

# The High Cost of Brand Name Large Language Models (LLMs)

Many enterprises are falling into the "Premium Trap," with their AI deployments. Organizations default to expensive, well-known "celebrity" LLMs for every task to alleviate perceived hallucination risks, regardless of the actual complexity required. This lack of discernment leads to massive budget leakage and unnecessary costs. Using top-tier, proprietary models for routine tasks creates unsustainable operational expenses.

When you over-engineer simple workflows with high-cost models, you significantly erode your AI project ROI. You are effectively using a supercar to deliver mail across the street; it is impressive, but the economics do not scale.

## Celebrity AI is Killing Your ROI

The costs of brand-name AI are not just financial; they are operational.

### Hidden Latency Costs

High-scoring "celebrity" models often suffer elevated latency. These delays ripple through real-time business processes, slowing down outcomes.

### Diminishing Returns

Our testing shows that paying premium rates for marginal gains in accuracy on non-critical tasks is a fiscal failure. For many enterprise tasks, "good enough" is reached long before you hit peak proprietary pricing.

### Scalability Bottlenecks

High token costs consume too much of your AI deployment costs. If your model costs \$15.00 per million tokens for output, you cannot afford trial-and-error or to deploy it across the entire enterprise.

Without a disciplined approach, your AI strategy remains an expensive "science project" rather than a foundational business differentiator.

## Task-Specific Orchestration

The optimal AI project is not finding one "perfect" model to suit your needs, but using a system that orchestrates the right model for the right task. Krista provides this orchestration and decision-making capability to route work to models based on performance and cost requirements across your enterprise. This strategy allows you to optimize for speed, quality, and cost across many use cases simultaneously, rather than picking the most expensive LLM for each.

# Matching Requirements to Outcomes

Enterprise workloads are diverse. Therefore, you must match model capabilities to specific activities, as different tasks require unique model skills. To build an effective AI and automation strategy, identify your requirements first, then select the model that delivers the best outcome at the lowest cost.

We have identified **five core enterprise generative AI requirements** based on real-world customer deployments:

## Unstructured Data Processing Use Cases

Automatically extracting keywords, identifying customer sentiment in feedback, or performing basic arithmetic on invoice data to enhance processing speed.

## Structured Workflow Automation Use Cases

Generating precise JSON-structured outputs to trigger system events, execute decision logic, or route user requests to the correct department without manual intervention.

## Domain-Specific Knowledge Retrieval Use Cases

Providing accurate answers to complex employee or customer queries based on internal HR policies, utility FAQs, or property management documents.

## Natural Language Intent Conversion Use Cases

Converting free-form conversations into structured forms and workflows, or translating complex system objects back into human-readable language.

## High-Fidelity Dialogue Understanding Use Cases

Transforming transcripts from technical discussions, operations reviews, and planning sessions into accurate, actionable summaries that update CRMs or open project tickets.

# Methodology: How We Validated LLM Performance

Krista engineers used a standardized LLM Evaluation Pipeline designed to systematically compare models across the aforementioned enterprise requirements across four critical vectors: **accuracy, error metrics, latency, and cost**. This framework ensures that closed-, and open-source, and task-optimized models compete on a level playing field to identify the best fit for specific enterprise use cases.

## The Evaluation Framework

Our team fed the dataset into each model using a disciplined, two-part prompt structure:

- **System Prompt:** Provides high-level guidance or instructions to steer the model's behavior.
- **User Prompt:** Contains the specific Question we want the model to answer, along with the necessary Context or background information it needs to understand the query.
- **Chain of Thought (CoT):** We use CoT prompting during evaluation to help judge models break down problems into a series of logical steps, mimicking human reasoning and improving assessment consistency. [1]

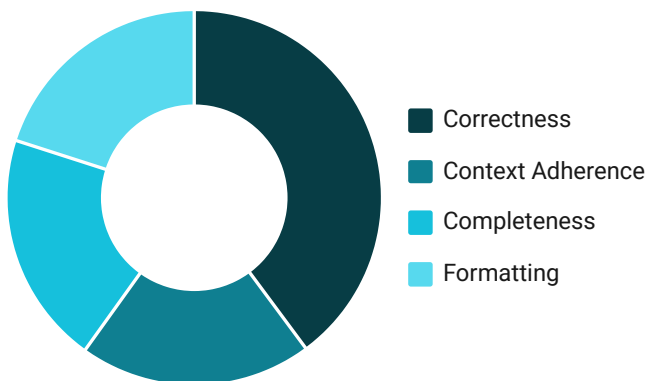
## The Dataset

We tested models against approximately **2,800 records** across five critical enterprise domains:

1. **Generic Data:** General benchmarks including attribute extraction, classification, entity recognition, and simple math.
2. **Krista AI Q&A:** Domain-specific data covering HR policies, Energy/Utilities FAQs, and Property Management.
3. **Agentic AI:** Conversions between natural language and structured form-based interactions.
4. **Krista Enterprise Understanding (KEU):** Summarization of actual enterprise meeting transcripts, technical discussions, and project updates.

## Scoring and Pass / Fail Thresholds

To remove human bias, we utilized LLM-based judges (specifically GPT 5.1 and Gemini 3-pro) to score every response on a scale of 0–100. [2,3]



- **Correctness (40 pts):** Evaluates factual accuracy and logical coherence.
- **Context Adherence (20 pts):** Measures how well the response aligns with the provided context.
- **Completeness (20 pts):** Assesses whether the response fully addresses the query.
- **Formatting (20 pts):** Examines the structure and clarity of the output.

📌 **The Enterprise Standard:** A model only "Passes" if the final averaged score from both judges is >80, or >70 with zero errors. This rigorous threshold ensures the results meet the demands of a high-stakes production environment.

# Benchmark Results: The ROI of Right-Sizing LLM Usage

Our testing proves that using a "Celebrity Model" is a choice, not a necessity. By measuring model performance across diverse enterprise tasks, we found that **task-specific models often match or exceed the performance of their proprietary counterparts at a fraction of the cost.** [5,6]

- **Closed-Source Soloists:** Premium models like **gpt-5.2** (KEU) and **gemini-3-pro-preview** (Q&A) are best reserved for high-stakes accuracy requirements.
- **The Krista LLM Engine:** For routine processing and agentic automation, **Krista LLM** delivers enterprise-grade quality with massive cost savings.
- **The Latency Advantage:** **Krista LLM** processes requests up to **4x faster** than reasoning-heavy closed models, which is critical for real-time user and customer experiences.
- **Peace of Mind:** Utilizing Krista LLM inside of your private Krista instance, ensures your data stays in your control.

## Ultimate Data Security & Privacy

While performance, speed, and cost are critical metrics for AI success, **Data Security** is the non-negotiable foundation of the Cognitive Enterprise. A primary concern for organizations adopting "Celebrity Models" is the risk of proprietary data being used by third-party providers to train their foundational models.

The Krista LLM eliminates this risk by offering a "Contained AI" architecture designed for maximum privacy:

- **Instance-Level Isolation:** When utilizing the Krista LLM, all processing and inference occur entirely within your private Krista instance.
- **Zero Third-Party Training:** Your sensitive data—including HR records, customer transcripts, and internal documentation—never leaves your environment to reach external AI companies.
- **Data Sovereignty:** Unlike proprietary closed models that require data to travel to external clouds, Krista LLM ensures your internal knowledge remains your intellectual property and is never used for third-party AI training.
- **Secure-by-Design:** By hosting the model locally within your Krista instance, Krista provides the highest level of security for compliance-heavy industries that cannot risk external data exposure.

In short, the Krista LLM doesn't just right-size your budget; it right-sizes your risk profile.

# Generic Data Test Results: The Strategic Efficiency Leader

The Generic Data dataset serves as the foundational workhorse for routine enterprise operations. This category evaluates a model's raw accuracy and instruction-following capabilities across the high-volume, mission-critical tasks that form the backbone of a modern Cognitive Enterprise.

In the modern Cognitive Enterprise, "Generic" refers to the fundamental nature of the task rather than its importance. These tasks power the essential "unstructured data processing" layer of a business, including:

- **Customer Experience (CX):** Automating initial support triage through Sentiment Analysis and Classification to route urgent issues effectively.
- **Data Entry Automation:** Utilizing Attribute Extraction and Entity Recognition to process invoices, receipts, or emails and format them into structured JSON for database entry.
- **Global Communication:** Maintaining contextual integrity during real-time Translation of internal or customer-facing communications.
- **Business Intelligence:** Converting natural language queries into SQL queries to provide non-technical staff with immediate database insights.

## Performance Comparison: The Strategic Efficiency Leader

Our evaluation pipeline systematically tested over 20 different LLMs to identify the best performers for these essential tasks. While proprietary "Celebrity Models" and open-source variants achieve high accuracy, the results show that the Krista LLM offers the best balance of speed and reliability for standard operations. [7,12]

# 91.4%

### o3-mini Pass Rate

**o3-mini**, achieved a **91.4% pass rate**. However, this marginal improvement in accuracy is often unnecessary for foundational tasks.

# 88.5%

### Krista LLM Pass Rate

**Krista LLM** can serve as the primary engine for the 80% of enterprise workloads that prioritize operational throughput and reliability.

# 4x

### Speed Advantage

**Krista LLM** is the clear winner for real-time applications, generating responses in just **1.49 seconds**.

- **The Strategic Efficiency Leader: Krista LLM** achieved a robust **88.5% pass rate** in this category. **Krista LLM** can serve as the primary engine for the 80% of enterprise workloads that prioritize operational throughput and reliability. [7,12]
- **The Closed-Source Leader: o3-mini (Reasoning)** The top-performing proprietary model, **o3-mini**, achieved a **91.4% pass rate**. However, this marginal improvement in accuracy is often unnecessary for foundational tasks like attribute extraction or sentiment analysis, where the "cost of perfection" results in significantly higher costs. [7,12]
- **Industry-Leading Latency: Krista LLM** is the clear winner for real-time applications, generating responses in just **1.49 seconds**. This makes it nearly **4x faster** than **o3-mini**, which requires **6.03 seconds** to complete the same task. For fluid customer experiences and high-volume batch processing, customer experience is paramount. [8,9]
- **Enterprise-Grade Reliability** Despite the discrepancy in cost, both models demonstrated identical structural precision with an exceptionally low **0.17% format error rate**. This confirms that **Krista LLM** can strictly adhere to complex JSON schemas, providing the same level of system-integration safety as its high-cost celebrity model competitors. [10,12]

# The Cost of Perfection

While the performance gap between the top models is statistically microscopic, the financial discrepancy is massive. o3-mini is significantly more expensive than Krista LLM and other open source models. [4]

Metric	Krista LLM	o3-mini
Input Cost (per 1M tokens)	\$0.10	\$1.10 (~11x higher)
Output Cost (per 1M tokens)	\$0.10	\$4.40 (~44x higher)
Data Security	100% Contained & Private	Third-Party Exposure Risk

## ROI Analysis: The Case for Right-Sizing

For foundational enterprise workflows, paying a 44x premium for a marginal gain in accuracy is a fiscal failure. Because **Krista LLM** handles the **80% of routine workloads** with near-parity accuracy and industry-leading speed, it is the superior choice for scalable AI deployment.

**The Verdict:** High-cost reasoning models like **o3-mini** are over-qualified and under-optimized for routine tasks. They are better reserved as "specialist soloists" for complex reasoning where their specific capabilities justify higher price points and slower response times.

**The Privacy Advantage:** By utilizing the **Krista LLM** for high-volume routine data, organizations ensure that sensitive PII and internal classifications remain entirely within their private instance, never reaching third-party providers.

## Krista AI Q&A Results: Precision vs. Speed Trade-off

The Krista AI Q&A dataset represents one of the most high-value applications for Large Language Models in the enterprise: Document-Based Question Answering. Unlike general knowledge tasks, this category specifically tests a model's ability to retrieve information accurately from authoritative, internal sources while strictly adhering to the provided context.

## Enterprise Use Cases: Precision and Retrieval

The Q&A dataset we used represents three enterprise verticals but applies to any vertical or content, reflecting critical business workflows where accurate information retrieval is paramount:

- **HR Services:** Automating an HR helpdesk where employees query internal documents regarding leave, benefits, policies, and handbooks.
- **Energy & Utilities:** Powering customer self-service chatbots that handle high-volume queries based on specific FAQ data (e.g., "How do I read my meter?").
- **Property Management:** Providing operational support to managers and field workers querying safety protocols, lease terms, or management documentation.

## Performance Comparison: The Precision vs. Speed Trade-off

The evaluation of over 20 models revealed a specialized performance landscape for document-heavy tasks. The results show a clear trade-off between the meticulous precision of closed-source "archivists" and the high-speed efficiency of efficiently-tuned models.

# 86.3%

### gemini-3-pro-preview Pass Rate

The closed-source leader with low factual error rate (4.65%) but slow response time of 7.04 seconds

# 80.8%

### Krista LLM Pass Rate

The strategic challenger delivering robust performance for 80% of routine queries

# 6x

### Speed Advantage

Krista LLM delivers answers in just 1.15 seconds—more than 6 times faster than the closed-source leader

## Cost Analysis: The ROI of Quality

While the performance gap between the top models is narrow, the cost discrepancy is the most significant differentiator for scalable deployment.[4]

Metric	Krista LLM	gemini-3-pro-preview
Input Cost (per 1M tokens)	\$0.10	\$2.00 (~20x more expensive)
Output Cost (per 1M tokens)	\$0.10	\$12.00 (~120x more expensive)
Data Security	100% Contained & Private	Third-Party Exposure Risk

**The Verdict:** For strictly compliance-heavy tasks where a single factual error results in liability, the premium for **gemini-3-pro-preview** may be justified. However, for the high-volume, standard internal knowledge bases and customer FAQs that make up 80% of your workload, **Krista LLM** is the superior business choice. It delivers roughly 94% of the quality at less than 8% of the cost and 600% of the speed.

**The Privacy Advantage:** Keep your proprietary policies, HR documents, and technical FAQs 100% private. Using **Krista LLM** ensures your internal "brain" is never used to train external foundational models.

# Agentic AI Results: The Reliability Convergence

The Agentic AI category represents a critical frontier in enterprise automation: the ability of Large Language Models (LLMs) to bridge the gap between human intent and rigid software systems. This dataset evaluates a model's capacity to act as an intermediary, converting free-form conversation into structured digital actions and vice versa.

## Enterprise Use Cases: Connecting Intent to Action

Agentic AI is the backbone of workflows where the AI doesn't just provide information but executes tasks. Typical enterprise applications include:

- **IT Service Management (ITSM):** Converting a request like "My laptop won't connect to the VPN" into a structured ticket by auto-populating specific fields such as Device ID and Urgency.
- **HR Automation:** Interpreting a request to "add a newborn to my insurance" and converting it into a precise database update or form submission that strictly adheres to a required schema.
- **Workflow Orchestration:** Executing complex commands, such as "approve all pending travel requests under \$500," across multiple database records by interpreting the underlying logic.

In these use cases, **Format Compliance is mission-critical**; if the AI generates invalid JSON, the entire automation fails.

## Performance Comparison: The Reliability Convergence

Our evaluation of over 20 models indicates that open-source and task-optimized models have effectively reached parity with closed-source leaders in the most vital metric for agents: Structural Integrity. [7,11]

# 96.8%

### claude-sonnet-4 Pass Rate

Its performance was virtually flawless in terms of structure, with 0.00% formatting errors and 0.00% hallucinations.

# 94.9%

### Krista LLM Pass Rate

Krista matched the closed-source leader's reliability with a **0.00% formatting error rate**

# 4x

### Speed Advantage

Krista LLM delivers answers in just 2.68 seconds, nearly 4 times faster than the closed-source leader

- **The Closed-Source Leader:** **claude-sonnet-4** achieved the highest Pass Rate in this category at **96.8%**. Its performance was virtually flawless in terms of structure, with 0.00% formatting errors and 0.00% hallucinations. However, this precision comes with a latency of **3.46 seconds**. [9,10,11]
- **The Specialist Model:** **Krista LLM** Built on a high-performing open-source foundation, **Krista LLM** reached a Pass Rate of **94.9%**. Critically, it matched the closed-source leader's reliability with a **0.00% formatting error rate**, meaning it effectively never breaks the software forms it interacts with. [7,11]
- **Speed and Cost Advantage:** **Krista LLM** is roughly **23% faster** than the leading proprietary model, responding in **2.68 seconds**. Because it is optimized for high-volume automation, it delivers enterprise-grade structural perfection at roughly **5% of the cost** of premium closed-source models. It is the logical choice for the 80% of routine enterprise transactions like ticket creation and form filling. [8,9]

# Cost Analysis: The ROI of Scalable Automation

For agents that may process thousands of daily transactions, cost is the primary differentiator. While the quality gap between the leaders is less than 2%, the price discrepancy is massive.[4,5,6]

Metric	Krista LLM	claude-sonnet-4
Input Cost (per 1M tokens)	\$0.10	\$3.00 (~30x more expensive)
Output Cost (per 1M tokens)	\$0.10	\$15.00 (~150x more expensive)
Data Security	100% Contained & Private	Third-Party Exposure Risk

**The Verdict:** For high-stakes financial transactions where even a minor error rate carries legal risk, the premium for **claude-sonnet-4** may be justified. However, for the high-volume daily transactions—like ticket creation, form filling, and workflow routing—that make up 80% of your workload, **Krista LLM** is the superior business choice. It delivers 98% of the quality at roughly 7% of the cost with faster execution.

**The Privacy Advantage:** Execute complex automated workflows across your enterprise systems with the peace of mind that your system-access logic and transaction data stay within your secure Krista instance.

## Putting Performance in Perspective





Using claude-sonnet-4 for routine agentic tasks is like hiring a Senior Diplomat to fill out DMV paperwork; they will do it with absolute perfection, but their high salary and deliberate pace are overkill for the task. Using Krista LLM is like hiring a Certified Paralegal; they produce perfect, error-free forms (0% format errors), work significantly faster, and cost a fraction of the price. For data processing, the paralegal is the high-ROI hire.

# Krista Enterprise Understanding (KEU): Perfection vs. Near-Parity

The Krista Enterprise Understanding (KEU) category represents a fundamental administrative function in the modern Cognitive Enterprise: transforming unstructured business dialogue into structured, actionable intelligence. This dataset specifically evaluates a model's ability to process meeting transcripts—real-world enterprise conversations that require a sophisticated understanding of multi-speaker dialogue, subtle nuances, and business context.

## Enterprise Use Cases: Summarization and Insight Extraction

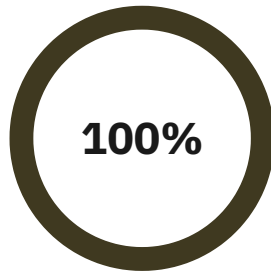
This capability drives internal productivity by converting hours of technical and operational dialogue into immediate, usable knowledge. Common enterprise applications include:

-  **Automated Minutes**  
Converting a 60-minute "Operations Review" transcript into a concise list of decisions made and action items assigned.
-  **Executive Briefings**  
Summarizing technical discussions for non-technical stakeholders, highlighting key risks and strategic milestones without technical jargon.
-  **Compliance Documentation**  
Extracting specific agreements or policy changes from decision-making sessions to create a reliable audit trail.
-  **Project Intelligence**  
Parsing planning meetings and project updates to automatically update CRMs or open project tickets based on actionable dialogue.

In these scenarios, **Factual Consistency** is required to ensure no fabricated action items are assigned, while **Formatting Accuracy** ensures the summary integrates seamlessly into downstream systems.

# Performance Comparison: Perfection vs. Near-Parity

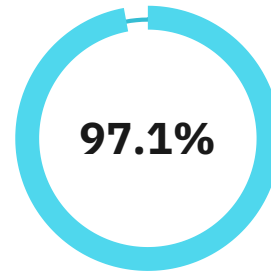
The evaluation results for KEU are unique as they feature the only instance of a "perfect" score in the entire report, highlighting a distinct trade-off between absolute perfection and high-value sufficiency. [7,13]



100%

## gpt-5.2 Pass Rate

The only model to achieve a flawless score with 0.00% error rate across all metrics



97.1%

## Krista LLM Pass Rate

Near-parity performance with 0.00% formatting errors and seamless system integration

### The Closed-Source Leader: gpt-5.2

- Undisputed champion: **100.0% Pass Rate**
- 0.00% error rate across all metrics
- No factual errors, formatting issues, hallucinations, or omissions
- Latency: 5.99 seconds [9,10,13]

### The Task-Optimized Challenger: Krista LLM

- Pass Rate: **97.1%**
- Within striking distance of perfection
- Minor 2.86% factual error rate
- 0.00% formatting errors
- Latency: 14.51 seconds [7,10,13]

For daily operational summaries and internal knowledge management, Krista LLM is the superior business choice, delivering enterprise-grade intelligence for roughly 5% of the price of proprietary "celebrity" models.

## Cost Analysis: The ROI of Scalability

While gpt-5.2 offers perfection, the premium for that final 2.9% of quality is significant.[4,5,6]

Metric	Krista LLM	gpt-5.2
Input Cost (per 1M tokens)	\$0.10	\$1.75 (~17x more expensive)
Output Cost (per 1M tokens)	\$0.10	\$14.00 (~140x more expensive)
Data Security	100% Contained & Private	Third-Party Exposure Risk

**The Verdict:** For legally binding transcripts or critical board meetings where absolute accuracy is required, the 100% accuracy of **gpt-5.2** justifies the premium. However, for the 80% of daily operational summaries and internal updates, **Krista LLM** is the superior business choice. It delivers enterprise-grade formatting and near-perfect accuracy for roughly 7% of the price of "celebrity" models.

**The Privacy Advantage:** Summarize sensitive board meetings, operational reviews, and technical discussions without the risk of your proprietary dialogue being leaked or used for third-party AI training.

# How Krista Solves the Multi-LLM Strategy

Organizations cannot achieve these ROI gains by simply picking one model. True efficiency requires **Strategic LLM Routing**, in which an orchestration and decision layer automatically selects the best model for each unique request.

Krista serves as the "conductor" for your orchestrated AI ecosystem, using a tiered logic to optimize ROI:

- **The 80/20 Efficiency Engine:** Krista is designed to route the vast majority of routine enterprise work (**roughly 80%**) to the **Krista LLM**. This ensures high speed and low cost for foundational tasks without sacrificing quality.
- **Strategic Specialist Routing:** For the remaining **20%** of tasks that require absolute 100% fidelity or specialized reasoning, the Krista conductor dynamically "fails up" to premium proprietary models.
- **Dynamic ROI Optimization:** By using **Krista LLM** as your default engine, you reclaim your AI budget while preserving premium reasoning for where it truly matters.
  - *Example:* Routing a routine HR FAQ to **Krista LLM** rather than **Gemini** saves **95% in costs** with nearly identical quality.
- **Graceful Failover:** If a primary model experiences high latency or fails, Krista's orchestration layer provides immediate failover to a secondary model, ensuring zero downtime for your business processes.

## The Mandate for a Cognitive Enterprise

The "Premium Trap" is the single greatest threat to a scalable AI budget. Our research proves that for over 90% of typical enterprise workloads, high-cost proprietary models are over-qualified and under-optimized. **Paying a 1,400% premium for a marginal 0.3% gain in accuracy is not a technical decision; it is a fiscal failure.**

The future of AI is not a single "supercar" model. It is a diverse fleet of specialized models coordinated by a single orchestration layer. By adopting a task-specific routing strategy, enterprises can finally move beyond "science projects" and build a high-ROI, production-ready AI foundation that scales without exploding costs.

## Stop Overpaying for Celebrity AI.

**Stop Overpaying for Celebrity AI.** Don't let brand-name hype dictate your operational expenses or compromise your data sovereignty. The "Premium Trap" is no longer a necessary cost of doing business.

### Audit Your Workloads

Identify where you are currently routing routine data processing—such as sentiment analysis, form filling, or meeting summaries—to high-cost external models.

### Secure Your Data

Shift from "public cloud" reasoning to **Contained AI**. Use the **Krista LLM** to ensure your proprietary documents, conversations, and workflows stay entirely within your private instance.

### Adopt Strategic Orchestration

Deploy Krista as your "Conductor" to gain the visibility needed to route work dynamically based on cost, quality, and security requirements.

### Schedule a Strategy Session

[Contact our team today](#) for a live demonstration of Krista's multi-LLM routing. Learn how to move beyond "science projects" and build a production-ready AI foundation that scales without exploding costs or exposing your data.

# About Krista

Krista is an AI-led agentic platform designed to transform complex business processes into seamless natural language interactions. By emphasizing a "nothing-like-code" approach, Krista empowers non-technical users to build and scale sophisticated automations through everyday language, reducing dependency on specialized developers and accelerating digital transformation.

As the **conductor** of your digital ecosystem, Krista functions as an NLP-enabled Integration Platform as a Service (iPaaS) that synchronizes people, enterprise systems, and intelligent agents. Its agentic architecture autonomously handles multi-step tasks across disparate applications—such as ERP and CRM systems—interpreting user intent to ensure efficient collaboration between human users and software.

Built on the principle of agentic autonomy, Krista enables AI agents to execute tasks independently within defined governance frameworks. This approach eliminates operational silos and democratizes automation by focusing on outcome-oriented interactions. Organizations in industries like IT, healthcare, and finance leverage Krista's conversational interface to streamline operations and foster innovation without extensive IT involvement. Visit [krista.ai](https://krista.ai) to learn how we are redefining the modern Cognitive Enterprise.

## [krista.ai](https://krista.ai)



# Appendix: Sources

1. [Deepeval's llm eval matrices](#)
2. <https://arxiv.org/abs/2303.16634D>
3. [LLM-as-Judge Huggingface](#)
4. [Open source model pricing](#)
5. Open-Source Models Overview
6. Closed-Source Models Overview
7. Model Performance Metrics
8. Pass Rate with Latency by Category (Open-Source Models)
9. Pass Rate with Latency by Category (Closed-Source Models)
10. Error Analysis by Category
11. Error Matrix: AgenticAI
12. Error Matrix: Generic Data
13. Error Matrix: KEU
14. Error Matrix: Q & A

# Open-Source Models Overview

Open-source models provide flexibility and control, allowing organizations to deploy and run models locally while maintaining data privacy. The following table presents the key specifications and pricing for each open-source model evaluated in this pipeline.

Open-Source Model	Context (I/O)	Vendor	Cost* (I/O)
Kimi-K2-Instruct	128k/16k	MoonshotAI	\$0.60 / \$2.50
gpt-oss-20b	128k/128k	OpenAI	\$0.07 / \$0.30
gpt-oss-120b	128k /128k	OpenAI	\$0.15 / \$0.60
Llama-4-Maverick-Instruct	1M / 4k	Meta	\$0.22 / \$0.88
Llama-3.3-70B-Instruct	128k / 2k	Meta	\$0.90 / \$0.90

\* Cost is in USD per million token for Input and Output. This is as per third party vendor ([fireworks.ai](https://fireworks.ai))

# Closed-Source Models Overview

Closed-source models are proprietary solutions accessed exclusively through APIs. While they often deliver strong performance, they require sending data to external servers, which may raise privacy and compliance concerns. The following table details the specifications and pricing for each closed-source model in our evaluation.

Closed-Source Model	Context (I/O)	Vendor	Cost* (I/O)
gpt-5.2	400k / 128k	OpenAI	\$1.75 / \$14.00
gpt-5.1	400k / 128k	OpenAI	\$1.25 / \$10.00
gpt-5	400k / 128k	OpenAI	\$1.25 / \$10.00
gpt-5-mini	400k / 128k	OpenAI	\$0.25 / \$2.00
gpt-4.1	1M / 32k	OpenAI	\$2.00 / \$8.00
gpt-4.1-mini	1M / 32k	OpenAI	\$0.40 / \$1.60
o3-mini (Reasoning)	200k / 100k	OpenAI	\$1.10 / \$4.40
gpt-4o	128k / 16k	OpenAI	\$2.50 / \$10.00
gpt-4o-mini	128k / 16k	OpenAI	\$0.15 / \$0.60
gpt-3.5-turbo	16k / 4k	OpenAI	\$0.50 / \$1.50
claude-sonnet-4.5	200k / 64k	Anthropic	\$3.00 / \$15.00
claude-sonnet-4	200k / 8k	Anthropic	\$3.00 / \$15.00
gemini-3-pro-preview	1M / 64k	Google	\$2.00 / \$12.00
grok-4-1-fast	2M / 256k	xAI	\$0.20 / \$0.50
Grok-4-fast	2M / 256k	xAI	\$0.20 / \$0.50
Krista LLM**	256k/32k	Krista	\$0.10 / \$0.10

\* Cost is in USD per million token for Input and Output. This is as per third party vendor ([fireworks.ai](https://fireworks.ai))

\*\*Krista Software

# Model Performance Metrics

## Pass Rate Matrix (Threshold: Avg Score > 80)

Model	AgenticAI	Generic Data	KEU	Q&A	Overall
gemini-3-pro-preview	92.4%	91.0%	97.1%	86.3%	<b>87.5%</b>
o3-mini (Reasoning)	94.3%	91.4%	90.0%	78.7%	<b>87.0%</b>
gpt-5.2	92.4%	88.6%	100.0%	84.4%	<b>85.7%</b>
gpt-4.1-mini	96.2%	88.3%	97.9%	84.1%	<b>85.7%</b>
claude-sonnet-4.5	95.5%	89.5%	90.7%	84.4%	<b>84.3%</b>
gpt-oss-120b	89.8%	91.5%	82.1%	83.3%	<b>84.2%</b>
gpt-4.1	92.4%	88.8%	87.9%	85.4%	<b>83.8%</b>
Krista LLM	94.9%	88.5%	97.1%	80.8%	<b>83.4%</b>
Kimi-K2-Instruct	90.5%	88.9%	94.3%	79.5%	<b>83.3%</b>
claude-sonnet-4	96.8%	87.5%	98.6%	83.9%	<b>83.3%</b>
gpt-oss-20b	90.5%	91.1%	82.9%	80.5%	<b>83.1%</b>
Llama-3.3-70B-Instruct	94.3%	85.6%	90.0%	77.8%	<b>81.7%</b>
gpt-5	93.0%	86.8%	97.9%	83.7%	<b>81.6%</b>
gpt-5-mini	93.6%	86.0%	89.3%	78.7%	<b>81.5%</b>
grok-4-1-fast	93.0%	84.3%	97.9%	79.5%	<b>81.2%</b>
gpt-5.1	94.3%	86.0%	99.3%	82.2%	<b>81.1%</b>
Grok-4-fast	93.0%	86.4%	92.9%	80.3%	<b>81.0%</b>
Llama-4-Maverick-Instruct	89.8%	86.9%	95.0%	77.4%	<b>80.6%</b>
gpt-4o	91.7%	86.2%	96.4%	79.5%	<b>80.2%</b>
gpt-4o-mini	88.5%	82.8%	87.9%	74.6%	<b>74.6%</b>
gpt-3.5-turbo	80.9%	79.9%	57.9%	61.7%	64.2%

# Pass Rate with Latency by Category

## Open-Source Models

Model	AgenticAI		Generic Data		KEU		Q & A	
	Pass%	Latency	Pass%	Latency	Pass%	Latency	Pass%	Latency
gpt-oss-120b	89.8%	3.17	91.5%	1.99	82.1%	5.65	83.3%	1.95
Kimi-K2-Instruct	90.5%	3.62	88.9%	1.45	94.3%	10.05	79.5%	0.76
gpt-oss-20b	90.5%	3.73	91.1%	1.65	82.9%	3.40	80.5%	1.51
Llama-3.3-70B-Instruct	94.3%	2.13	85.6%	1.61	90.0%	3.58	77.8%	2.02
Llama-4-Maverick-Instruct	89.8%	2.35	86.9%	1.37	95.0%	5.32	77.4%	1.01

# Pass Rate with Latency by Category

## Closed-Source Models

Model	AgenticAI		Generic_data		KEU		Q&A	
	Pass%	Latency	Pass%	Latency	Pass%	Latency	Pass%	Latency
Krista LLM	94.9%	2.68	88.5%	1.49	97.1%	14.51	80.8%	1.15
gemini-3-pro-preview	92.4%	9.51	91.0%	5.50	97.1%	11.19	86.3%	7.04
o3-mini (Reasoning)	94.3%	6.12	91.4%	6.03	90.0%	5.74	78.7%	4.82
gpt-4.1-mini	96.2%	2.02	88.3%	1.06	97.9%	3.96	84.1%	0.91
gpt-5.2	92.4%	2.10	88.6%	1.51	100.0%	5.99	84.4%	1.83
claude-sonnet-4.5	95.5%	4.39	89.5%	3.52	90.7%	14.45	84.4%	3.89
gpt-4.1	92.4%	2.04	88.8%	1.94	87.9%	5.75	85.4%	1.11
claude-sonnet-4	96.8%	3.46	87.5%	2.82	98.6%	11.63	83.9%	3.72
gpt-5	93.0%	3.52	86.8%	2.35	97.9%	7.76	83.7%	2.12
gpt-5-mini	93.6%	3.89	86.0%	2.19	89.3%	5.99	78.7%	1.68
grok-4-1-fast	93.0%	9.64	84.3%	2.05	97.9%	6.68	79.5%	1.58
gpt-5.1	94.3%	2.81	86.0%	2.02	99.3%	13.24	82.2%	3.10
Grok-4-fast	93.0%	3.54	86.4%	1.24	92.9%	4.40	80.3%	1.95
gpt-4o	91.7%	2.37	86.2%	2.11	96.4%	4.79	79.5%	1.18
gpt-4o-mini	88.5%	3.07	82.8%	1.44	87.9%	3.78	74.6%	0.85
gpt-3.5-turbo	80.9%	1.67	79.9%	0.90	57.9%	1.33	61.7%	0.57

# Error Analysis by Category

Model	AgenticAI	Generic Data	KEU	Q&A	Overall
gemini-3-pro-preview	7.64%	9.00%	2.86%	13.74%	<b>12.50%</b>
o3-mini (Reasoning)	5.73%	8.58%	10.00%	21.35%	<b>13.00%</b>
gpt-4.1-mini	3.82%	11.67%	2.14%	15.86%	<b>14.33%</b>
gpt-5.2	7.64%	11.42%	0.00%	15.64%	<b>14.33%</b>
claude-sonnet-4.5	4.46%	10.50%	9.29%	15.64%	<b>15.65%</b>
gpt-oss-120b	10.19%	8.50%	17.86%	16.70%	<b>15.83%</b>
gpt-4.1	7.64%	11.25%	12.14%	14.59%	<b>16.18%</b>
Krista LLM	5.10%	11.50%	2.86%	19.24%	<b>16.58%</b>
Kimi-K2-Instruct	9.55%	11.08%	5.71%	20.51%	<b>16.65%</b>
claude-sonnet-4	3.18%	12.50%	1.43%	16.07%	<b>16.68%</b>
gpt-oss-20b	9.55%	8.92%	17.14%	19.45%	<b>16.90%</b>
Llama-3.3-70B-Instruct	5.73%	14.42%	10.00%	22.20%	<b>18.33%</b>
gpt-5	7.01%	13.25%	2.14%	16.28%	<b>18.40%</b>
gpt-5-mini	6.37%	14.00%	10.71%	21.35%	<b>18.47%</b>
grok-4-1-fast	7.01%	15.67%	2.14%	20.51%	<b>18.79%</b>
gpt-5.1	5.73%	14.00%	0.71%	17.76%	<b>18.90%</b>
Grok-4-fast	7.01%	13.58%	7.14%	19.66%	<b>19.04%</b>
Llama-4-Maverick-Instruct	10.19%	13.08%	5.00%	22.62%	<b>19.44%</b>
gpt-4o	8.28%	13.83%	3.57%	20.51%	<b>19.83%</b>
gpt-4o-mini	11.46%	17.17%	12.14%	25.37%	<b>25.44%</b>
gpt-3.5-turbo	19.11%	20.08%	42.14%	38.27%	<b>35.83%</b>

# Error Matrix: **Agentic AI**

Model	Factual Error	Format Error	Hallucination	Omission	Low_Score	Overall Error
claude-sonnet-4	0.64%	0.00%	0.00%	1.27%	1.27%	3.18%
gpt-4.1-mini	0.64%	0.00%	0.00%	1.91%	1.27%	3.82%
claude-sonnet-4.5	0.00%	0.00%	0.00%	1.91%	2.55%	4.46%
Qwen3-235B-A22B-Instruct	1.91%	0.00%	0.64%	1.91%	0.64%	5.10%
o3-mini (Reasoning)	2.55%	0.00%	0.00%	1.27%	2.55%	5.73%
Llama-3.3-70B-Instruct	3.82%	0.00%	0.00%	0.64%	1.27%	5.73%
gpt-5.1	2.55%	0.00%	0.00%	3.18%	1.27%	5.73%
gpt-5-mini	3.82%	0.00%	0.00%	2.55%	1.27%	6.37%
Grok-4-fast	5.73%	0.64%	0.00%	2.55%	0.00%	7.01%
grok-4-1-fast	3.82%	0.00%	0.64%	1.91%	1.91%	7.01%
gpt-5	3.82%	0.00%	0.00%	1.91%	1.91%	7.01%
gpt-4.1	3.82%	0.00%	0.00%	4.46%	1.27%	7.64%
gpt-5.2	5.10%	0.00%	0.00%	3.18%	0.00%	7.64%
gemini-3-pro-preview	5.10%	0.00%	0.00%	3.18%	0.64%	7.64%
gpt-4o	4.46%	0.00%	0.64%	3.18%	1.91%	8.28%
Kimí-K2-Instruct	4.46%	0.00%	0.64%	3.82%	1.91%	9.55%
gpt-oss-20b	3.18%	1.27%	0.64%	5.10%	0.64%	9.55%
Llama-4-Maverick-Instruct	6.37%	0.64%	0.64%	3.82%	1.27%	10.19%
gpt-oss-120b	4.46%	0.00%	0.64%	5.10%	0.64%	10.19%
gpt-4o-mini	5.73%	0.00%	1.27%	4.46%	1.27%	11.46%
gpt-3.5-turbo	12.74%	0.00%	1.27%	7.01%	1.27%	19.11%

# Error Matrix: Generic Data

Model	Factual Error	Format Error	Hallucination	Omission	Low Score	Overall Error
gpt-oss-120b	3.58%	0.08%	0.83%	2.08%	3.00%	8.50%
o3-mini (Reasoning)	4.92%	0.17%	0.75%	1.42%	2.42%	8.58%
gpt-oss-20b	5.42%	0.17%	0.92%	1.33%	2.33%	8.92%
gemini-3-pro-preview	4.50%	0.08%	0.75%	2.42%	2.58%	9.00%
claude-sonnet-4.5	6.42%	0.42%	0.75%	2.00%	2.42%	10.50%
Kimí-K2-Instruct	7.17%	0.17%	0.42%	2.33%	2.00%	11.08%
gpt-4.1	7.50%	0.17%	0.42%	2.08%	2.25%	11.25%
gpt-5.2	6.00%	0.42%	1.00%	3.75%	2.00%	11.42%
Qwen3-235B-A22B-Instruct	7.08%	0.17%	1.08%	2.33%	2.00%	11.50%
gpt-4.1-mini	7.50%	0.17%	0.92%	2.00%	2.58%	11.67%
claude-sonnet-4	6.67%	0.42%	0.58%	3.17%	3.25%	12.50%
Llama-4-Maverick-Instruct	8.58%	0.42%	1.08%	2.42%	2.42%	13.08%
gpt-5	7.58%	0.67%	0.83%	3.42%	2.25%	13.25%
Grok-4-fast	9.33%	0.42%	1.92%	1.58%	2.00%	13.58%
gpt-4o	8.92%	0.25%	1.08%	2.83%	2.25%	13.83%
gpt-5-mini	9.58%	0.42%	0.67%	2.08%	2.42%	14.00%
gpt-5.1	8.33%	0.00%	0.92%	3.17%	2.67%	14.00%
Llama-3.3-70B-Instruct	10.00%	0.42%	1.17%	1.42%	2.67%	14.42%
grok-4-1-fast	10.83%	0.33%	1.00%	2.92%	2.00%	15.67%
gpt-4o-mini	12.92%	0.17%	0.75%	1.25%	2.83%	17.17%
gpt-3.5-turbo	12.58%	0.17%	0.92%	4.67%	3.00%	20.08%

In the Generic\_data dataset, **gpt-oss-120b** leads with the lowest overall error rate of 8.50%, with o3-mini (Reasoning) and gpt-oss-20b following at 8.58% and 8.92% respectively.

# Error Matrix: KEU

Model	Factual Error	Format Error	Hallucination	Omission	Low_Score	Overall Error
gpt-5.2	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
gpt-5.1	0.71%	0.00%	0.00%	0.00%	0.00%	0.71%
claude-sonnet-4	1.43%	0.00%	0.00%	0.00%	0.00%	1.43%
gpt-4-1-mini	1.43%	0.00%	0.00%	1.43%	0.00%	2.14%
grok-4-1-fast	2.14%	0.00%	1.43%	0.00%	0.00%	2.14%
gpt-5	2.14%	0.00%	0.71%	0.00%	0.00%	2.14%
Qwen3-235B-A22B-Instruct	2.86%	0.00%	0.71%	0.71%	0.00%	2.86%
gemin-3-pro-preview	2.86%	0.00%	0.71%	0.00%	0.00%	2.86%
gpt-4o	0.00%	0.71%	0.00%	3.57%	0.00%	3.57%
Llama-4-Maverick-Instruct	2.14%	0.00%	0.71%	5.00%	0.00%	5.00%
Kim-42-Instruct	3.57%	0.00%	2.14%	0.00%	0.71%	5.71%
Grok-4-fast	3.57%	0.00%	2.14%	4.29%	0.71%	7.14%
claude-sonnet-4.5	8.57%	0.00%	0.00%	6.43%	0.00%	9.29%
o3-mini (Reasoning)	3.57%	0.00%	0.71%	7.14%	0.00%	10.00%
Llama-3-3-70B-Instruct	5.00%	0.00%	1.43%	9.29%	0.00%	10.00%
gpt-5-mini	0.00%	3.57%	0.00%	10.71%	0.00%	10.71%
gpt-4o-mini	6.43%	0.00%	0.00%	7.14%	0.71%	12.14%
gpt-4.1	9.29%	0.00%	1.43%	10.71%	0.00%	12.14%
gpt-oss-20b	15.00%	0.00%	5.71%	2.14%	0.00%	17.14%
gpt-oss-120b	17.14%	0.00%	6.43%	12.14%	0.00%	17.86%
gpt-3.5-turbo	8.57%	7.14%	0.00%	36.43%	1.43%	42.14%

The KEU dataset reveals **gpt-5.2** achieving a perfect score with 0.00% overall error rate, demonstrating exceptional performance across all error categories.

# Error Matrix: Q&A

Model	Factual Error	Format Error	Hallucination	Omission	Low Score	Overall Error
gemin-3-pro-preview	4.65%	0.00%	2.33%	6.98%	1.27%	13.74%
gpt-4.1	7.82%	0.00%	2.96%	5.92%	0.63%	14.59%
claude-sonnet-4.5	7.61%	0.63%	1.90%	5.92%	0.85%	15.64%
gpt-5.2	6.55%	0.00%	2.11%	7.82%	0.85%	15.64%
gpt-4.1-mini	8.46%	0.00%	3.38%	5.29%	1.06%	15.86%
claude-sonnet-4	5.50%	0.00%	1.69%	9.94%	0.21%	16.07%
gpt-5	8.67%	0.00%	1.69%	6.77%	1.27%	16.28%
gpt-oss-120b	9.09%	0.00%	5.71%	4.65%	0.63%	16.70%
gpt-5.1	8.03%	0.00%	1.90%	9.51%	1.06%	17.76%
Qwen3-235B-A22B-Instruct	8.46%	0.00%	3.38%	8.67%	1.06%	19.24%
gpt-oss-20b	10.99%	0.00%	6.98%	4.65%	1.90%	19.45%
Grok-4-fast	11.21%	0.00%	3.81%	5.50%	1.27%	19.66%
gpt-4o	7.82%	0.00%	1.48%	12.47%	1.06%	20.51%
grok-4-1-fast	11.84%	0.00%	1.90%	7.40%	1.48%	20.51%
Kimi-K2-Instruct	7.19%	0.00%	1.48%	12.68%	1.27%	20.51%
o3-mini (Reasoning)	10.99%	0.00%	2.75%	8.67%	2.33%	21.35%
gpt-5-mini	9.30%	0.00%	2.75%	11.42%	0.85%	21.35%
Llama-3.3-70B-Instruct	9.51%	0.00%	2.11%	10.78%	1.90%	22.20%
Llama-4-Maverick-Instruct	12.90%	0.00%	4.23%	8.25%	2.11%	22.62%
gpt-4o-mini	11.42%	0.00%	2.54%	12.90%	1.27%	25.37%
gpt-3.5-turbo	23.47%	0.21%	4.02%	16.49%	1.06%	38.27%