WHITEPAPER

IntelePeer®

Best Practices for Evaluating Large Language Models (LLMs) for Use by Analysis Agents

CONTENTS:

EXECUTIVE SUMMARY

» Overview of the agentic AI framework and its role in conversational AI

» Introduction to analysis agents and how they use LLMs to perform various tasks

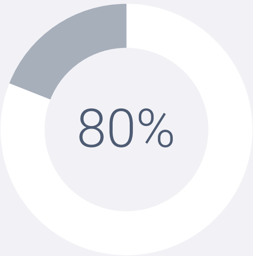
» Approach to evaluating LLMs used by AI agents and measuring their performance against specific use cases

» Analysis agents and their benefits — IntelePeer’s approach to analysis agents leveraging commercial LLMs to breakdown tasks and delivering insights at scale

Introduction

The rise of LLMs in enterprise customer experience (CX)

The emergence of commercial large language models (LLMs) is reshaping the customer experience (CX) landscape, enabling businesses to engage with customers through intelligent, human-like conversations at scale, in real-time, and across channels. These models serve as the foundation for next-generation conversational AI systems that go beyond reactive support to deliver proactive, context-aware interactions. By automating routine tasks such as inquiries, appointment scheduling, and outbound communications, LLMs free human agents to focus on complex, high-value engagements. This shift is accelerating the transition from traditional contact centers to AI-driven service models. According to Gartner, by 2030[1], **AI is expected to handle 80% of all customer interactions** — underscoring the urgency for enterprises to evaluate and adopt the right LLMs to stay competitive.



AI is expected to handle 80% of all customer interactions by 2030.

LLMs like OpenAI's GPT4, Anthropic's Claude, or xAI's Grok are optimized for general-purpose use and offer strong out-of-the-box performance, particularly in language fluency and reasoning. As the adoption of AI agents grows, observability and trust are becoming critical. Enterprises are investing in tools to monitor model behavior, benchmark performance, and trace outputs to ensure reliability and transparency. These capabilities help mitigate risks such as hallucinations and bias, enabling the responsible deployment of AI at scale. Meanwhile, innovations like context-augmented generation and agentic frameworks are gaining traction, with analysts emphasizing the importance of real-world utility and developer ecosystem maturity.

Beyond operational efficiency, AI agents are redefining how enterprises approach customer engagement. By enabling real-time personalization, these AI agents adapt responses based on customer history, sentiment, and context, creating more human-like and satisfying interactions. This not only boosts customer satisfaction but also drives loyalty and long-term value. As organizations explore agentic AI and conversational frameworks, the focus is shifting toward building systems that are not only intelligent but also empathetic and brand-aligned.

Selecting the right commercial model for domain-specific AI agent use cases

While general-purpose commercial LLMs deliver powerful, out-of-the-box performance across a wide range of tasks, they also provide flexible tools and ecosystems that make them ideally suited — even for highly specialized industries like healthcare, finance, or telecommunications. Rather than reinventing the wheel, organizations can leverage built-in features such as prompt templates, retrieval-augmented generation (RAG) pipelines, and turnkey fine-tuning services to quickly inject domain knowledge, dramatically reducing both cost and complexity. With mature SDKs, pre-built connectors to common enterprise data sources, and ongoing model improvements from leading AI vendors, enterprises can achieve rapid time-to-

[1] Source: Gartner, Gartner Predicts Chatbots Will Become a Primary Customer Service Channel Within Five Years, July 27, 2022.

value without building or sourcing large domain-specific datasets or orchestration layers from scratch — unlocking meaningful outcomes right from launch.

Another major challenge is navigating the complexity of costs, compliance, and integration. Commercial LLMs often come with opaque pricing structures, making it challenging to forecast total cost of ownership, especially when factoring in usage-based fees, infrastructure, and customization. In regulated industries, data governance and explainability are crucial; however, many commercial models lack the necessary transparency and deployment flexibility (e.g., on-premises options) required to meet compliance standards. Integration into legacy systems or fragmented workflows also adds friction, requiring robust middleware and observability tools to ensure reliability, traceability, and trust. These factors make vendor selection a strategic decision that must align with both technical and business priorities.

Selecting the right commercial LLM for enterprise use cases and workflows is a strategic decision that hinges on aligning model capabilities with specific business objectives, such as reducing resolution times, enhancing personalization, or ensuring compliance. While domain-specific models offer contextual precision, they often lack flexibility and require significant upkeep. In contrast, commercial LLMs provide a more adaptable foundation. With techniques such as retrieval-augmented generation (RAG), prompt engineering, and lightweight fine-tuning, enterprises can tailor general-purpose models to fit their workflows without having to build them from scratch.

This adaptability is especially valuable in dynamic environments where customer expectations, regulatory requirements, and internal processes are constantly evolving. Commercial models can be integrated with internal systems, customized for industry-specific language, and deployed securely in the cloud. Their transparency, auditability, and growing ecosystem also support long-term innovation and governance. For customer service teams, this means a commercial model can be tuned to understand domain-specific terminology, access internal knowledge bases, and comply with handling policies — all while remaining flexible as business needs to shift.

Another key consideration is architectural agility. Enterprises benefit from adopting a model-agnostic framework that allows them to experiment with different LLMs, benchmark performance, and swap components as needed. This flexibility not only reduces vendor lock-in but also empowers teams to rapidly adopt emerging innovations in the LLM ecosystem, unlocking strategic advantages through continuous improvement and differentiation. Ultimately, selecting the right commercial LLM is not just about technical fit — it's about building a scalable, secure, and future-ready foundation for AI-driven transformation.

Why you need a flexible, secure, and transparent evaluation framework

As enterprises increasingly adopt AI agents driven by commercial LLMs to power customer experience, analytics, and automation, the need for a robust and practical LLM evaluation framework becomes essential. Traditional academic benchmarks often fail to reflect how these models perform in real-world, domain-specific workflows. Enterprises require a framework that goes beyond token-level accuracy to assess how well a model aligns with business goals, integrates with existing systems, and handles sensitive data. A flexible evaluation approach allows teams to test commercial models across varied use cases — such as customer service, compliance, or marketing — without being locked into a single vendor or architecture.

Security and governance are equally critical. Commercial LLMs must be evaluated not only for performance but also for their ability to operate within enterprise-grade security protocols. This includes compliance with standards like HIPAA, PCI DSS, and SOC 2, as well as the ability to enforce ethical behavior through prompt constraints, content filtering, and fallback mechanisms. Internal architectures like IntelePeer's five-layered guardrail system — spanning orchestration, governance, embedded ethics, real-time filtering, and analytics — demonstrate how layered defenses can ensure responsible AI behavior at scale.

Transparency is the third pillar of a strong evaluation framework. Enterprises need visibility into how commercial models make decisions, especially in regulated industries where auditability is non-negotiable. Transparent evaluation frameworks incorporate both automated metrics and human-in-the-loop assessments to ensure that LLM outputs are not only accurate but also explainable and aligned with brand values. Tools like transcript scoring, semantic clustering, and proactive dialog analysis help teams understand model behavior in context and refine it over time.

> “Selecting the right commercial LLM is not just about performance — it's about building a foundation for sustainable innovation.”

Ultimately, **selecting the right commercial LLM is not just about performance — it's about building a foundation for sustainable innovation.** A well-structured evaluation framework empowers organizations to iterate quickly, integrate modularly, and maintain trust across stakeholders. As the commercial LLM landscape evolves, this approach ensures enterprises can remain agile, secure, and competitive in deploying AI at scale.

The agentic AI framework and analysis agents

Definition and principles of agentic AI

Agentic AI is a new approach to enterprise AI, where systems are built to be autonomous, flexible, and focused on specific goals. These systems comprise specialized agents that act independently to achieve specific objectives. These agents understand context, adjust to new information, and work together with other agents and tools to consistently execute complex workflows with minimal human input. This is especially useful in CX environments where speed, accuracy, and the ability to scale are critical.

At the core of agentic AI is modularity. Each agent is built for a specific function or objective — such as interacting with customers for scheduling an appointment or interacting with data for data analysis — and multiple agents using task specific LLMs can be orchestrated through frameworks that support memory, task routing, and contextual awareness. For example, a customer service workflow might involve one agent authenticating a user, another retrieving account data, and a third escalating unresolved issues — each powered by the most appropriate commercial LLM and working in concert. Embed these models into a secure, enterprise-grade environment governed through layered guardrails that ensure compliance, ethical behavior, and observability makes them especially suitable for regulated industries like healthcare, finance, and legal services.

Ultimately, **agentic AI powered by commercial LLMs offers a scalable, flexible, and future-ready approach to enterprise automation,** empowering organizations to build intelligent systems that not only respond to tasks but proactively manage them, driving efficiency, improving customer experience, and enabling continuous innovation. As enterprises evolve, a modular architecture provides the ability to adapt and scale with control.

What are analysis agents, and how do they leverage the agentic AI framework?

Analysis agents powered by commercial LLMs have the power to change how enterprises evaluate and interpret large volumes of customer interactions, such as customer service calls, chat transcripts, and sales conversations. These agents are designed to autonomously extract deep insights from structured and unstructured interaction data such as assessing the quality and effectiveness of both human and AI agents for resolution accuracy, compliance, and agent follow-through or identifying operational insights such as cancellation drivers or dissatisfaction with services. Analysis agents allow organizations to maintain consistent quality standards at a scale without relying on resource and time-intensive manual data analysis processes.

Agentic AI, powered by commercial LLMs, offers a scalable, flexible, and future-ready approach to enterprise automation.

Built on the agentic AI framework, analysis agents are modular, and goal driven. Each agent is responsible for a specific set of tasks, such as detecting sentiment and outcome, verifying policy adherence, or identifying escalation triggers, and can operate independently or as part of a coordinated workflow. Commercial LLMs provide the natural language understanding backbone for these agents to understand the unstructured data and generate useful insights at scale. These insights help improve AI agent automation, support human interactions, and drive operational change to deliver the service and experience customers expect.

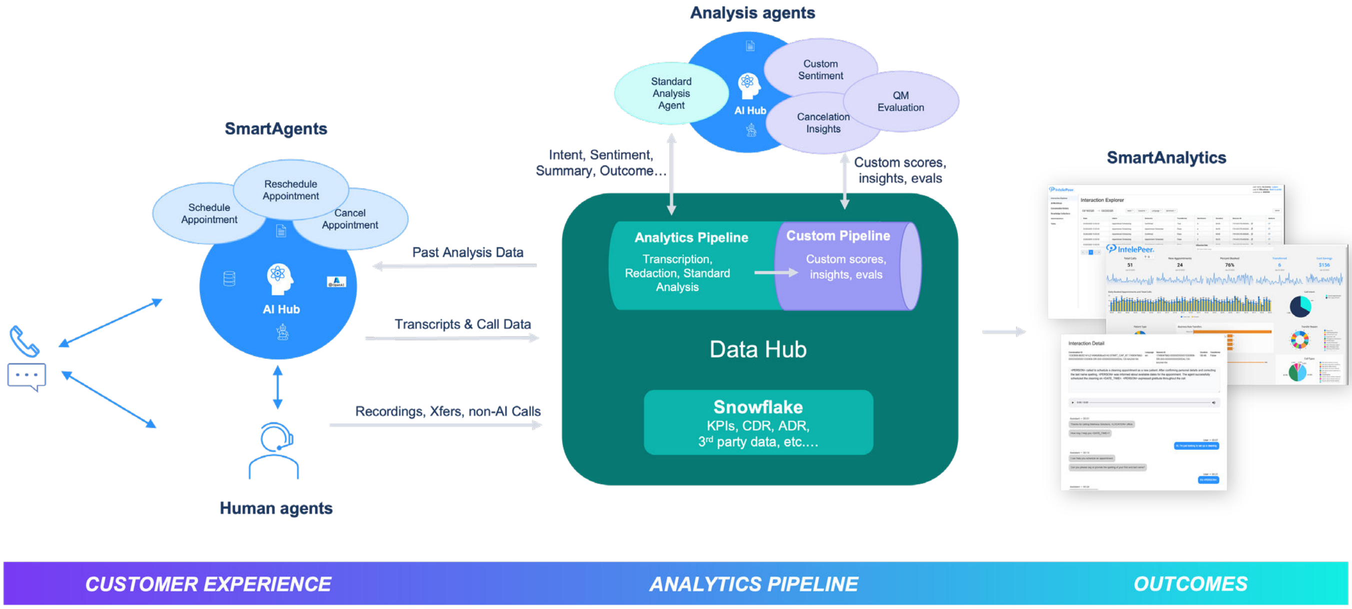


FIGURE 1. SmartAnalytics™. **End-to-end insights from every interaction with SmartAgents and analysis agents.**

For example, one analysis agent will evaluate the caller sentiment, while a second evaluates the interactions based on a quality rubric and a third analysis agent seeks to understand the reason the caller was canceling their service. The ability to orchestrate tasks specific analysis agents across all interaction data enables enterprises to automate performance evaluations including surface coaching opportunities and ensure compliance with internal policies as well as gain insights into customer experience across the business all while maintaining transparency and auditability. The modularity of analysis agents also allows enterprises to deploy, update, and scale agents across different business units without disrupting existing systems.

Structured evaluation of commercial LLMs

A four-stage evaluation framework

To ensure a rigorous and enterprise-relevant assessment of commercial large language models (LLMs), we developed a four-stage evaluation framework designed to balance automation, human judgment, and privacy. This framework enables organizations to benchmark models like OpenAI’s GPT-4, Anthropic’s Claude, and xAI’s Grok across real-world enterprise use cases — such as customer service analysis, compliance monitoring, and operational insight generation.

Stage 1: Data preparation and privacy protection

We begin by preparing a standardized dataset of enterprise transcripts for the target use case, ensuring all personally identifiable information (PII) is redacted. This step is critical for maintaining compliance with data protection regulations and for enabling fair, reproducible evaluations across providers.

Stage 2: Parallel model execution

Each transcript is processed in parallel through multiple commercial LLMs. This allows for direct comparison of model behavior under identical conditions, ensuring consistency in input and enabling meaningful output analysis.

Stage 3: Automated performance assessment

We apply a suite of quantitative metrics to evaluate functional and semantic performance:

- Tool invocation rate: Measures how often the model correctly triggers expected analysis of functions.
- Schema compliance: Assesses the accuracy of attribute identification and population.
- Categorical consistency: Analyzes scoring distribution across an 11-point quality scale.
- Inter-provider similarity: Uses cosine similarity to compare semantic content across models.
- Response length variability: Evaluates verbosity patterns to detect over- or under-generation.
- Contextual relevance: Scores how well responses align with the original conversation context.

Stage 4: Human evaluation and preference mapping

To complement automated metrics, we conduct blind human evaluations. Using in-house developed processes and tools, reviewers rank anonymized outputs based on clarity, relevance, and utility. Open-ended feedback is collected to capture qualitative insights, and bias detection is performed to identify systematic provider-specific patterns.

This multi-dimensional framework provides enterprises with a transparent, scalable, and vendor-agnostic method for evaluating commercial LLMs for use in CX agentic AI applications and informed decision-making by combining statistical rigor with human judgment ensuring that selected models align with business goals.

Evaluation results

Test setup

To evaluate the performance of commercial LLMs in the customer experience domain, we generated synthetic transcripts and used them to assess how different models performed. We then conducted human evaluations to compare the models, involving ten participants. While these evaluations offered directional insights, the results were not statistically significant. In total, we analyzed approximately 205 transcripts of typical conversations in the insurance industry, using OpenAI GPT-4o-mini, xAI Grok-3 mini, and Claude 3 Haiku.

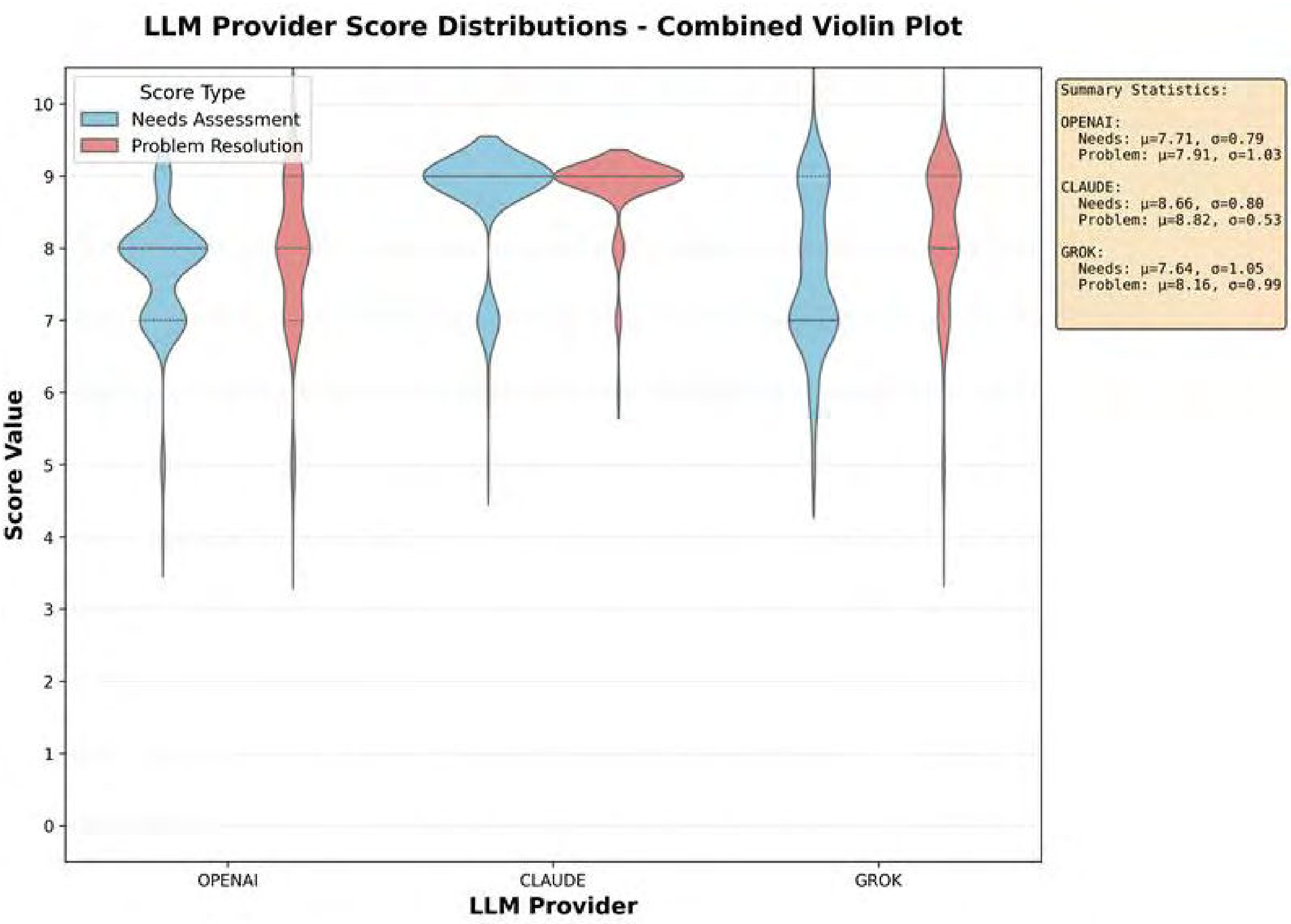


FIGURE 2. Commercial LLM provider performances. **These data analyze approximately 205 transcripts of typical conversations in the insurance industry.**

Test results

Both GPT4o-mini and Grok3-mini performed well for the use cases that were tested. However, these findings are directional and context-dependent. Additionally, the use of real-world data and dimensions such as prompting techniques can significantly impact the model's performance. Rather than pointing to a single "best" model, our experience underscores the importance of having a systematic framework for evaluating not only LLM performance in isolation, but as part of a comprehensive evaluation of the end-to-end agentic AI platform — grounded in real-world use cases, clear success metrics, and iterative testing — can help organizations align model capabilities with business goals, reduce risk, and accelerate time to value.

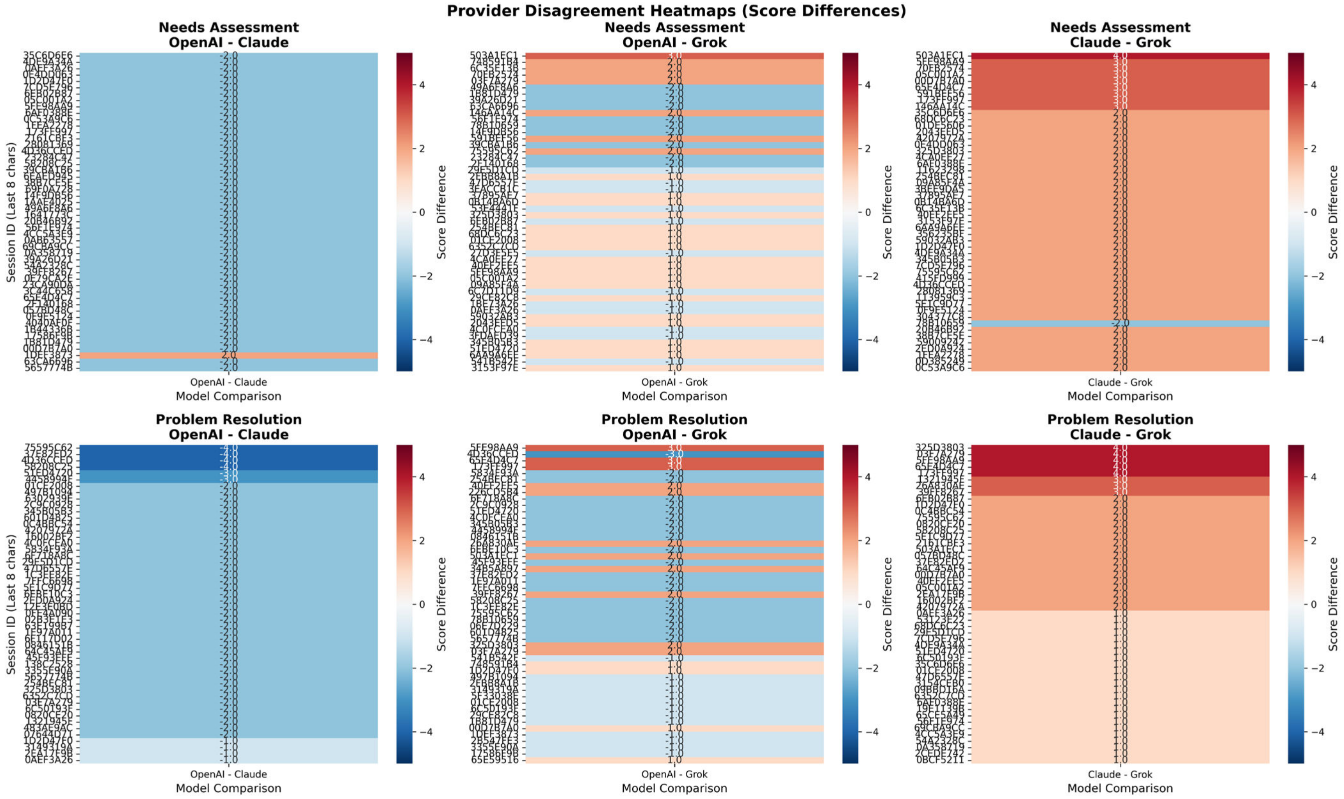


FIGURE 3. Commercial LLM provider disagreements. Model comparison reveals the importance of having systemic frameworks in place for evaluating LLM performance in isolation as well as the agentic AI platform itself.

Conclusion

Next steps for organizations building analysis agents

It is important to note that some models are strong generalists while others really shine in specific use cases and tasks like real-time conversations. However, the most important aspect is to have a comprehensive evaluation framework that can be applied continuously to real-world data to obtain the most informative results for decision-making related to model selection.

As businesses navigate the rapidly evolving landscape of large language models and agentic AI, adopting robust evaluation frameworks becomes essential for successful implementation. By prioritizing flexibility, security, and transparency, organizations can ensure that their AI systems not only meet operational and compliance requirements but also align with broader business objectives. **The agentic AI framework, with its modular and goal-oriented approach, provides a pathway for building intelligent, adaptable systems that deliver meaningful outcomes across various domains.**

> “The agentic AI framework, with its modular and goal-oriented approach, provides a pathway for building intelligent, adaptable systems that deliver meaningful outcomes across various domains.”

This comprehensive approach empowers modular agents, enables dynamic workflows, and ensures ethical AI practices, allowing organizations to strike a balance between autonomy and control. Additionally, analysis agents illustrate how specialized components can drive performance insights and maintain quality standards at scale. Through innovative evaluation methods like multi-dimensional assessment pipelines and human-in-the-loop reviews, enterprises can establish trust and accountability when deploying these advanced technologies.

Ultimately, integrating frameworks that support iterative testing and a modular design positions organizations to achieve immediate operational gains and long-term sustainability. As the capabilities of LLMs and agentic AI continue to advance, the strategic application of these tools will define the next era of enterprise transformation, fostering innovation, efficiency, and resilience in an increasingly complex digital economy.

AUTHOR(S) INFORMATION

Arvind Rangarajan is Director of AI Product Marketing at IntelePeer with deep expertise in translating complex AI technologies into real-world business value. He brings years of experience in product marketing and management, driving thought leadership and customer-centric innovation.

Matthew Caraway, Senior AI Product Manager at IntelePeer, is an accomplished leader in developing innovative and trustworthy enterprise AI agents. Over the past five years, Matthew has pioneered AI product initiatives across various industries, leveraging advanced large language models and traditional machine learning techniques. His passion for innovation, coupled with a disciplined approach to product management and observability, positions him as a respected voice in AI product strategy and responsible AI implementation.

ABOUT INTELEPEER

IntelePeer streamlines customer interactions, enabling businesses and contact centers to lower costs, improve the customer experience, and accelerate return on investment. Harnessing the power of agentic AI, IntelePeer’s Conversational AI Platform delivers speed, observability, visibility, and flexibility — all built on top of a global, secure communications network. Producing human-like interactions, the platform automates voice and digital customer service capabilities and provides industry-leading time-to-value with solutions that work seamlessly with existing enterprise software and infrastructure, and easy-to-use tools that can be utilized by anyone.

Get in touch with IntelePeer:

Contact us

www.intelepeer.ai

(877) 336-9171