# MASTERING AI AGENTS

A comprehensive guide to evaluating AI agents

Galileo

# Preface

In our previous e-book, "Mastering RAG," our goal was clear: building enterprise-grade RAG systems, productionizing them, monitoring their performance, and improving them. At the core of it, we understood how RAG systems enhance an LLM's ability to work with specific knowledge by providing relevant context.

In this e-book, we're taking a step further and asking, "How do we use LLMs to accomplish end-to-end tasks?" This singular question opens up a door: AI agents. A RAG system helps an LLM provide accurate answers based on given context. An AI agent takes that answer and actually does something with it — makes decisions, executes tasks, or coordinates multiple steps to achieve a goal.

A RAG-enhanced LLM could help answer questions about policy details by pulling relevant information. But an AI agent could actually process the claim end-to-end by analyzing the documentation, checking policy compliance, calculating payments, and even coordinating with other systems or agents when needed.

The ideas behind agents has existed for years. It can be a software program or another computational entity that can accept input from its environment and take actions based on rules. With AI agents, you're getting what has never been there before: the ability to understand the context without predefined rules, the capacity to tune decisions based on context, and learning from every interaction. What you're getting is not just a bot working with a fixed set of rules but a system capable of making advanced decisions in real-time.

Companies have quickly adapted, adopted, and integrated AI agents into their workflows. Capgemini's research found that "10% of organizations already use AI agents, more than half plan to use them in 2025 and 82% plan to integrate them within the next three years."

This e-book aims to be your go-to guide for all things AI agents. If you're a leader looking to guide your company to build successful agentic applications, this e-book can serve as a great guide to get you started. We also explore approaches to measuring how well your AI agents perform, as well as common pitfalls you may encounter when designing, measuring, and improving them.

**The book is divided into five chapters:**

**Chapter 1** introduces AI agents, their optimal applications, and scenarios where they might be excessive. It covers various agent types and includes three real-world use cases to illustrate their potential.

**Chapter 2** details three frameworks—LangGraph, Autogen, and CrewAI—with evaluation criteria to help choose the best fit. It ends with case studies of companies using these frameworks for specific AI tasks.

**Chapter 3** explores the evaluation of an AI agent through a step-by-step example of a finance research agent.

**Chapter 4** explores how to measure agent performance across systems, task completion, quality control, and tool interaction, supported by five detailed use cases.

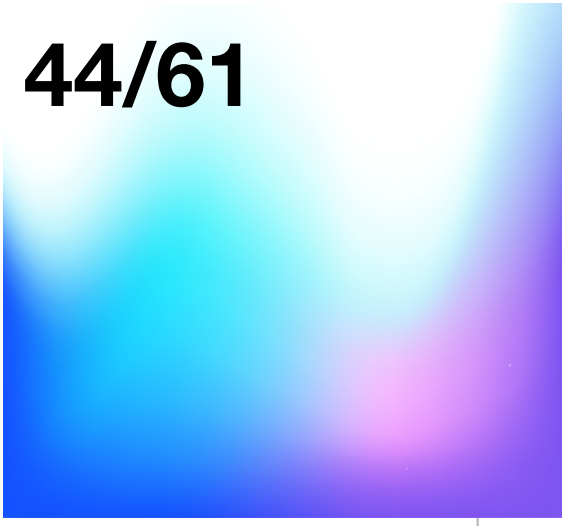
**Chapter 5** addresses why many AI agents fail and offers practical solutions for successful AI deployment.

We hope this book will be a great stepping stone in your journey to build trustworthy agentic systems.

**- Pratik Bhavsar**

# Contents

## Chapter 3:
## How to Evaluate Agents






## Chapter 4:
## Metrics for Evaluating AI Agents

## Chapter 5:
## Why Most AI Agents Fail & How to Fix Them

# 04
# CHAPTER

# METRICS FOR EVALUATING AI AGENTS

# Metrics for Evaluating AI Agents

Before we explore metrics for evaluating AI, let's recall our key insights into agent evaluation. Using LLM-based judges (like GPT-4o) and robust metrics (such as context adherence), we effectively measured an agent's performance across various dimensions, including accuracy, speed, and cost efficiency. We then set up Galileo's evaluation callback to track and record the agent's performance.

This next chapter will explore various metrics for evaluating AI agents using five solid case studies.

Let's consider a document processing agent. While it might initially demonstrate strong performance metrics, we may have to probe into several questions:

- Is it maintaining optimal processing speeds and resource usage?
- How consistently does it complete assigned tasks without human intervention?
- Does it reliably adhere to specified formatting and accuracy requirements?
- Is it selecting and applying the most appropriate tools for each task?

Through a series of hypothetical case studies, we'll explore how organizations may transform their AI agents into reliable digital colleagues using key metrics. These examples will demonstrate practical approaches to:

- Improving task completion rates and reducing human oversight
- Enhancing output quality and consistency
- Maximizing effective tool utilization and selection

You should remember that the goal isn't perfection but establishing reliable, measurable, and continuously improving AI agents that deliver consistent value across all four key performance dimensions. See **Fig 4.1**


**Fig 4.1:** Four key performance dimensions to evaluate AI agents

# Case Study 1: Advancing the Claims Processing Agent

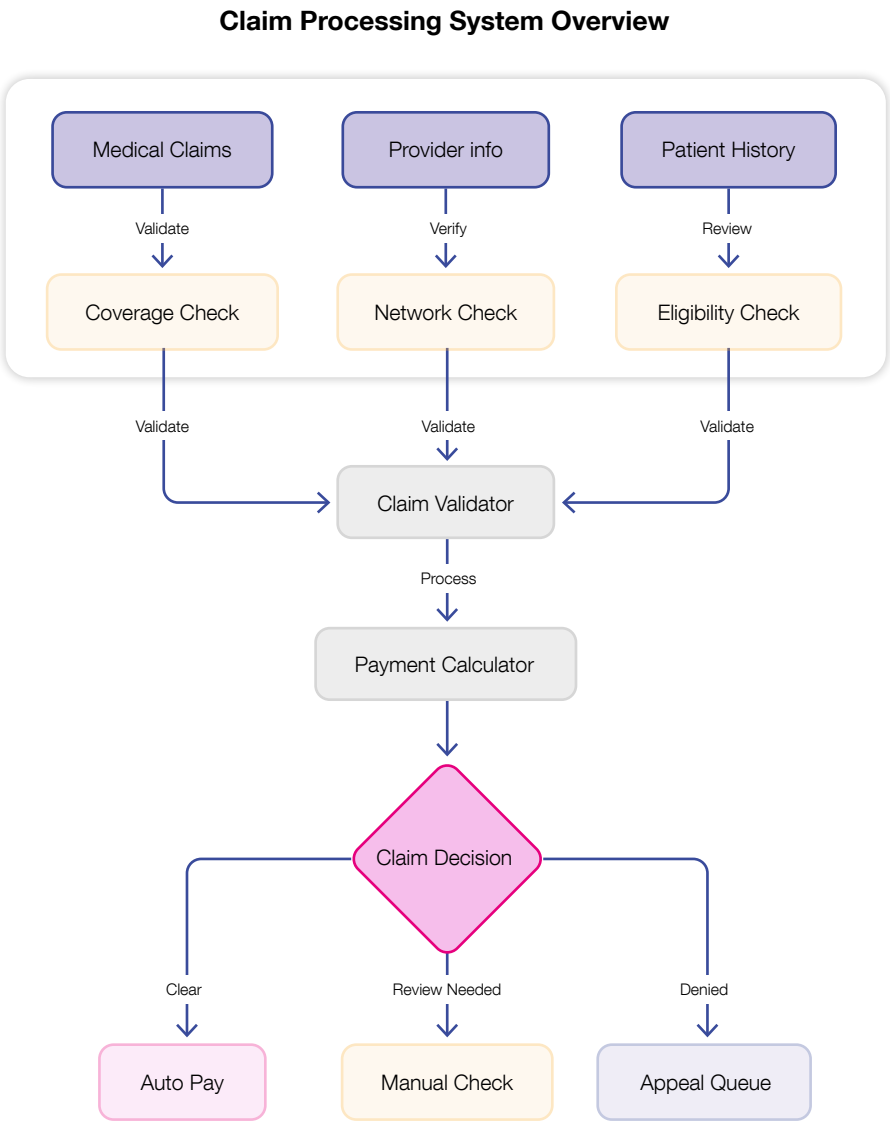



**Fig 4.2:** An overview of the Claims Processing System

A healthcare network implemented an AI agent to automate insurance claims processing, aiming to enhance efficiency and accuracy. However, this initiative inadvertently introduced compliance risks, highlighted by several key issues:

- The AI agent struggled with complex claims, leading to payment delays and provider frustration. Because of the inconsistency in handling these claims, claims processors spent more time verifying the AI's work than processing new claims.
- The error rate in complex cases raised alarms with the compliance team, especially critical given the stringent regulatory demands of healthcare claims processing.

## Functionality

The AI was designed to:

- Analyze medical codes
- Verify insurance coverage
- Check policy compliance
- Validate provider information
- Automatically assess claim completeness and compliance
- Calculate expected payments and generate preliminary approvals for straightforward claims

## Challenges

To counter these issues, the network focused on three key performance indicators to transform their AI agent's capabilities:

1. **LLM Call Error Rate**
   - **Problem:** API failures during claims analysis led to incomplete processing and incorrect approvals.
   - **Solution:** Implementing robust error recovery protocols and strict state management ensured accurate rollbacks and reprocessing.

2. **Task Completion Rate**
   - **Problem:** The agent incorrectly marked claims as 'complete' without conducting all necessary verifications.
   - **Solution:** Mandatory verification checklists and completion criteria were introduced to meet all regulatory requirements before finalizing claims.

3. **Number of Human Requests**
    - **Problem:** The agent took on complex cases beyond its capability, such as experimental procedures or cases requiring coordination of benefits across multiple policies.
    - **Solution:** Stricter escalation protocols automatically route high-risk cases to human experts based on claim complexity and regulatory requirements.

4. **Token Usage per Interaction**
    - **Problem:** Unnecessary inclusion of patient details in processing routine claims heightened privacy risks.
    - **Solution:** Strict data minimization protocols and context-cleaning practices were adopted to ensure that only essential protected health information is used

## Outcomes

The enhanced agent delivered:

- Faster claims processing
- Higher compliance accuracy
- Improved resource utilization
- Reduced rejection rates

# Case Study 2: Optimizing the Tax Audit Agent

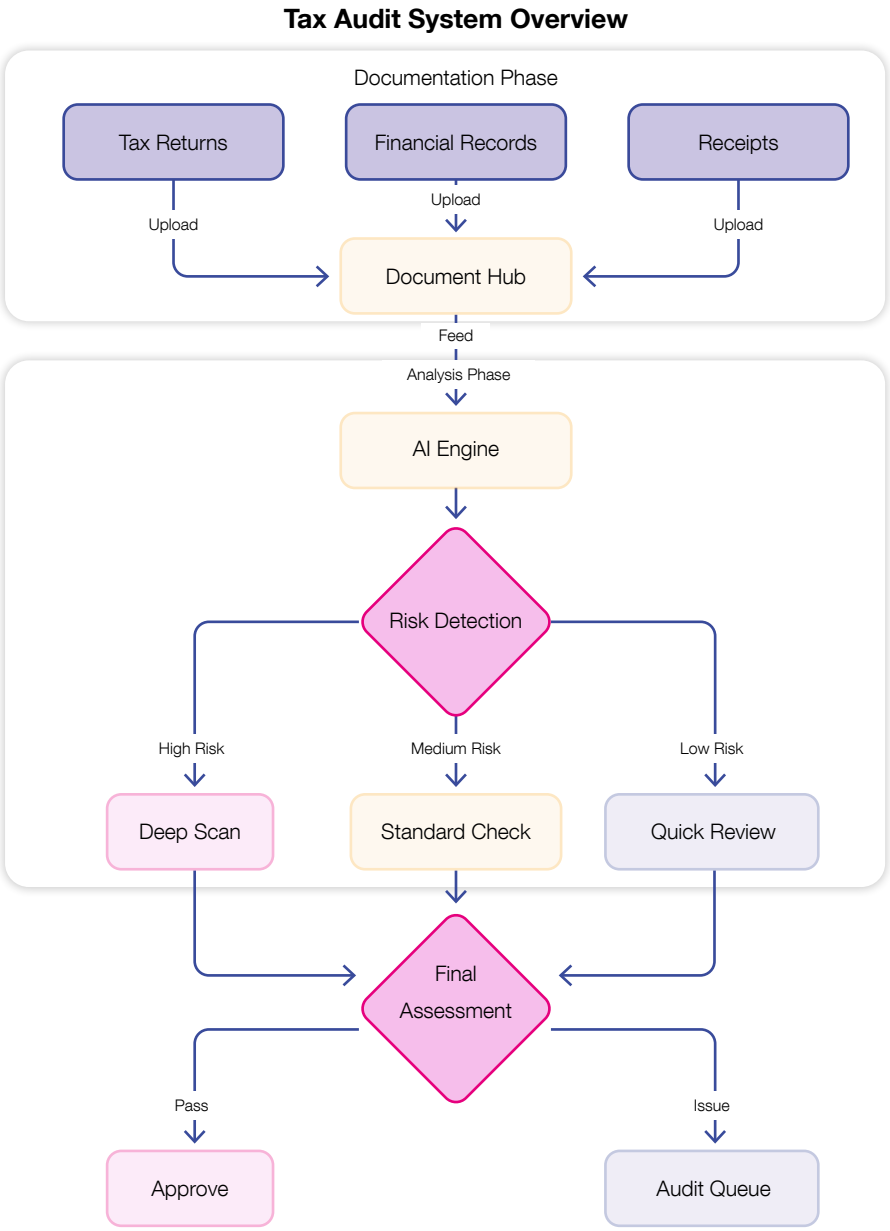



**Fig 4.3:** An overview of the Tax Auditing System

At a mid-sized accounting firm, their deployed AI audit agent created unexpected workflow bottlenecks. While the agent effectively handled routine tax document processing, the firm was concerned about three critical issues:

- Lengthy turnaround times for complex corporate audits
- Excessive computing costs from inefficient processing
- A growing backlog of partially completed audits requiring manual review

What should have streamlined their operations was instead causing senior auditors to spend more time supervising the AI's work than doing their specialized analysis. The firm needed to understand why its significant investment in AI wasn't delivering the anticipated productivity gains.

## Functionality

The AI audit agent was designed to:

- Process various tax documents, from basic expense receipts to complex corporate financial statements.
- Automatically extract and cross-reference key financial data in corporate tax returns.
- Systematically verify compliance across multiple tax years.
- Validate deduction claims against established rules and flag discrepancies for review.
- For simpler cases, it could generate preliminary audit findings and reports.
- The system was integrated with the firm's tax software and document management systems to access historical records and precedents.

## Challenges

The team focused on three critical metrics to reshape their agent's capabilities:

1. **Tool Success Rate**
   - **Problem:** The agent struggled with document processing efficiency, especially with complex document hierarchies.
   - **Solution:** Implementation of structured document classification protocols and validation frameworks improved handling of complex documents.

2. **Context Window Utilization**
   - **Problem:** The agent's processing of tax histories in their entirety was suboptimal, often missing connections between related transactions.

- **Solution:** Smart context segmentation was introduced, allowing the agent to focus on relevant time periods and maintain historical context. This enhanced the detection of subtle tax patterns.

3. **Steps per Task**
   - **Problem:** The agent applied the same level of analysis intensity to all tasks, regardless of complexity.
   - **Solution:** Adaptive workflows were implemented to adjust analytical depth based on the complexity of the task.

## Outcomes

The refined capabilities of the AI agent led to:

- Decreased audit completion times
- Improved accuracy in discrepancy detection
- More efficient utilization of processing resources

# Case Study 3: Elevating the Stock Analysis Agent

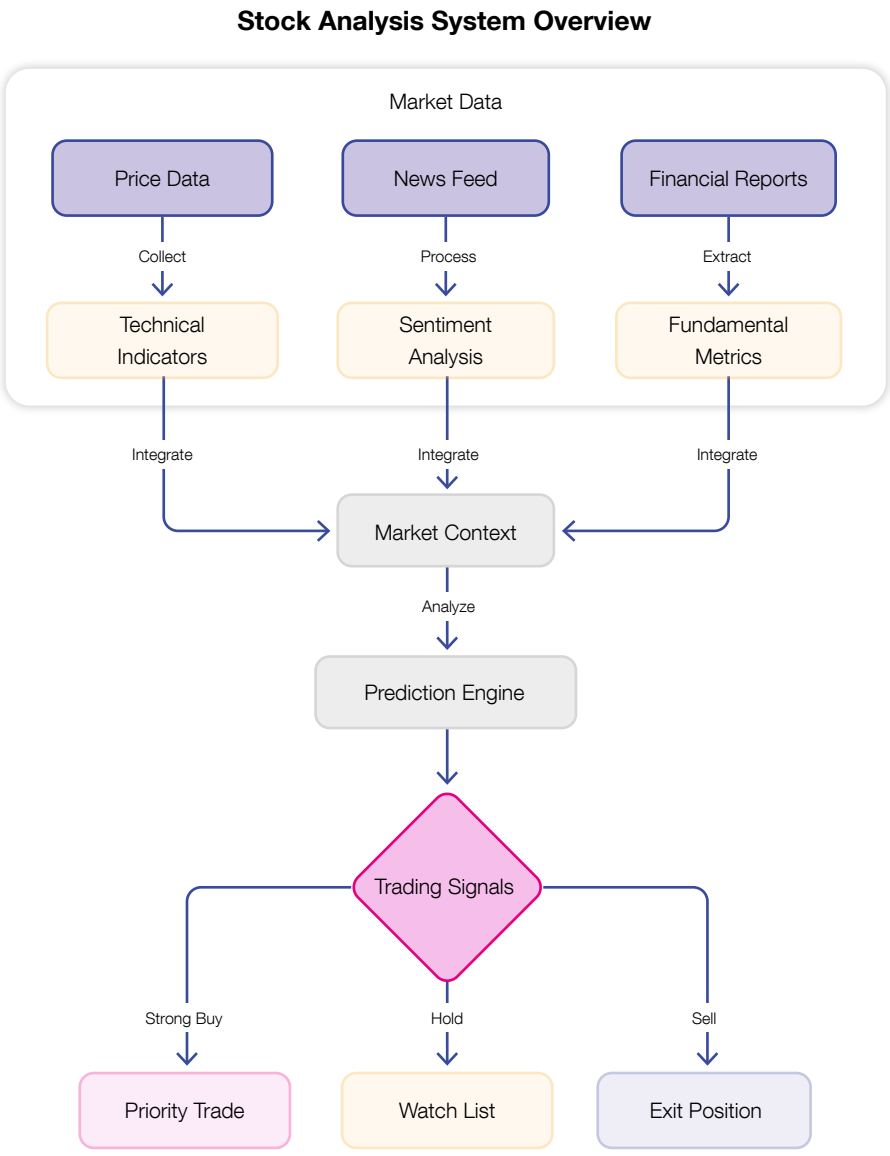
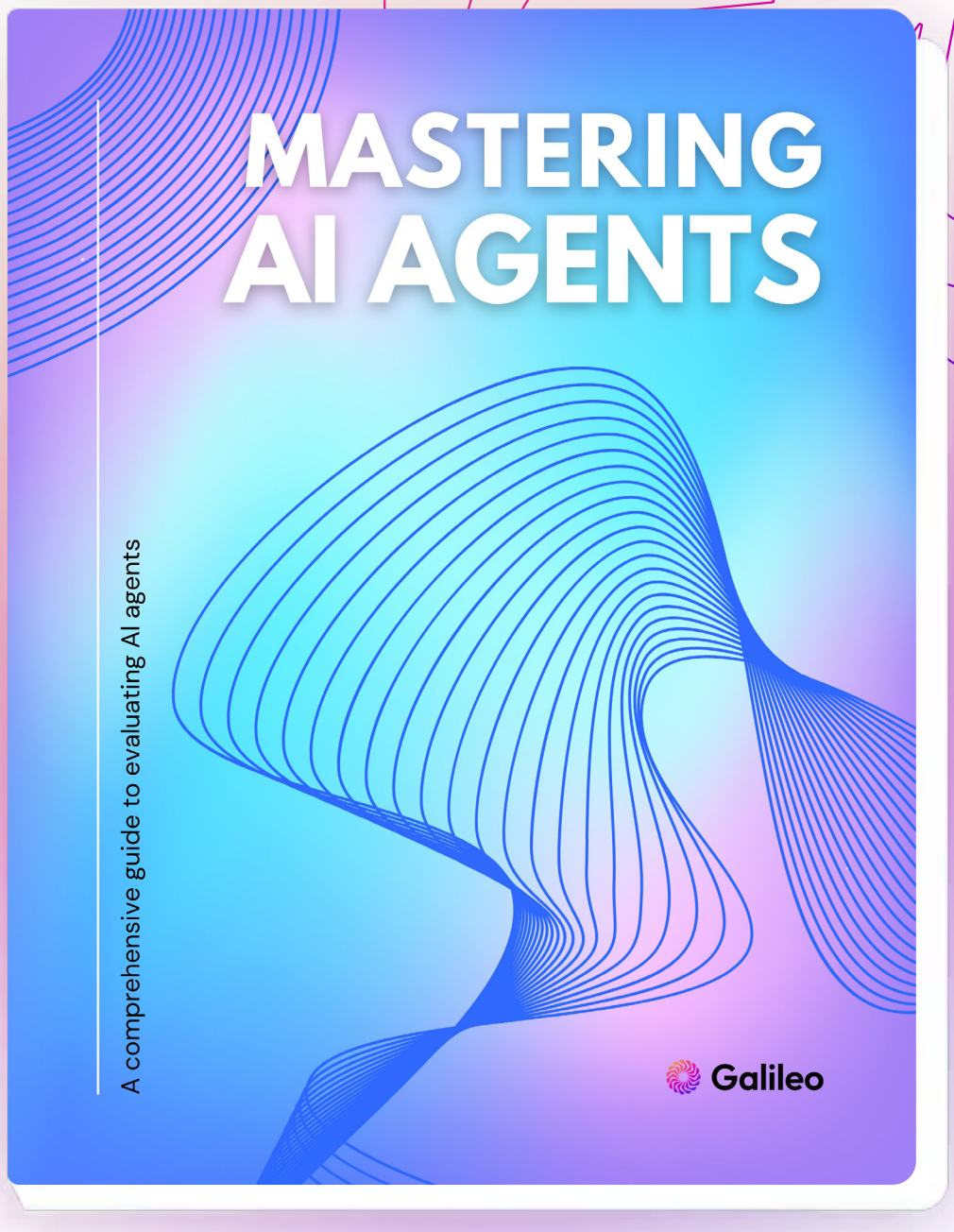



**Fig 4.4:** An overview of the Stock Analysis System

## Download the Full ebook




Galileo