AI Agents in Finance: When Algorithms Act Autonomously

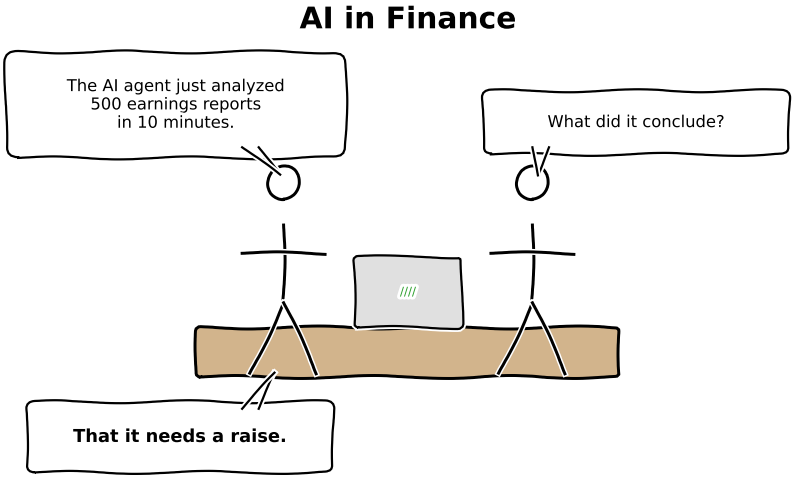
Module 5: The Automation Problem

Prof. Dr. Joerg Osterrieder

Digital Finance — BSc Course

Standalone lecture — explores what happens when AI moves from advising humans to acting on their behalf in financial markets.

When the Algorithm Takes the Wheel



The leap from "AI recommends" to "AI acts" is the defining shift in financial automation — and the source of both opportunity and risk.

Learning Objectives

After completing this lecture, you will be able to:

1. **Distinguish** between language models, copilots, and autonomous agents in terms of capability and autonomy [Understand]
2. **Explain** the ReAct architecture — how tools, memory, and reasoning enable agents to act [Understand]
3. **Analyze** four application domains where AI agents are transforming finance: trading, compliance, customer service, and research [Analyze]
4. **Evaluate** agent-specific risks including hallucination, liability gaps, and systemic correlation [Evaluate]
5. **Assess** how the EU AI Act classifies and governs autonomous AI systems in financial services [Evaluate]

Bloom's levels covered: Understand, Analyze, Evaluate

Objectives follow Bloom's taxonomy: Understand → Analyze → Evaluate. We build from definitions to critical assessment.

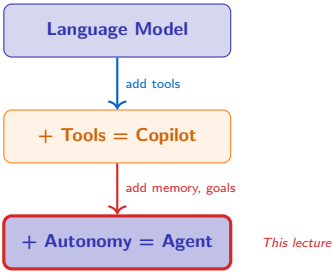
Bridge: From Language Models to Agents

Module 5 Lesson 2 taught us:

- Large Language Models (LLMs) can read, summarize, and generate financial text
- They operate as powerful **tools** — answering questions when asked
- But they have no **memory** between conversations and cannot **act** on the world

This lecture asks:

What happens when we give LLMs tools, memory, and the authority to act?



An LLM answers questions. A copilot helps with tasks. An agent pursues goals autonomously — that is the progression we explore today.

The Central Question

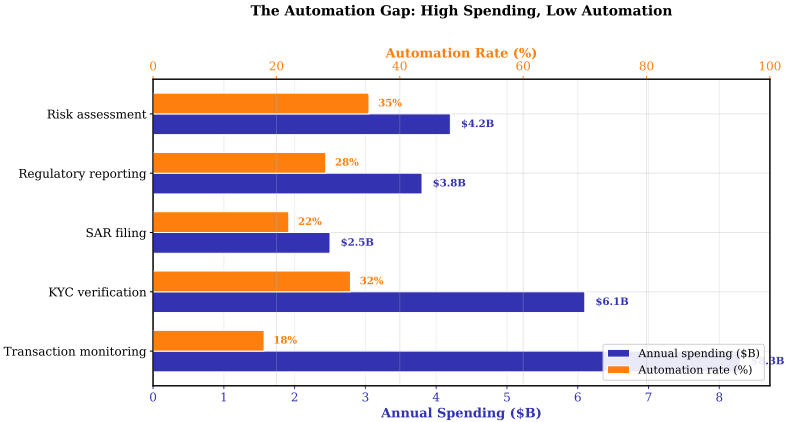
"What happens when AI doesn't just advise — it acts?"

Three guiding sub-questions:

1. Which financial tasks can be **safely delegated** to autonomous agents — and which cannot?
2. What **architecture** enables an AI system to reason, use tools, and learn from experience?
3. When an agent makes a mistake that costs money, **who is responsible**?

These three questions structure the entire lecture — we return to each in the summary on Slide 43.

The Automation Gap



What you see: The share of manual versus automated work in major banking functions — compliance, client onboarding, trade settlement, and reporting. Despite decades of IT investment, a majority of compliance work remains manual (Source: IIF/McKinsey, 2023).
Key pattern: The most regulated functions are the least automated.

Takeaway: Banks spend an estimated $270B/year (*IIF Financial Services Regulatory Outlook*, 2023) on compliance globally; agents promise to close this gap.

The automation gap is largest where the work is most judgmental — exactly the territory where AI agents, not simple scripts, are needed.

The Paradox of AI Advice

What LLMs Can Already Do:

- Read a 200-page 10-K filing in 12 seconds
- Summarize earnings calls with 95%+ accuracy
- Compare financial ratios across 500 companies simultaneously
- Draft investment memos in the style of a senior analyst
- Flag anomalies in transaction data

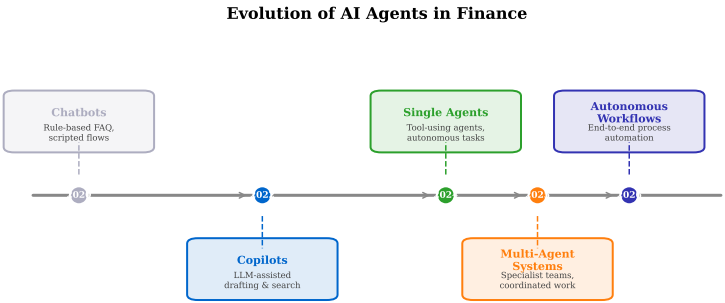
What They Cannot Do:

- Click "buy" on a trading platform
- Submit a Suspicious Activity Report (SAR) to regulators
- Transfer funds between accounts
- Update a client's risk profile in the CRM
- Escalate a fraud case to the compliance officer

The paradox: AI can analyze faster than any human, but it cannot *do* anything with its analysis — unless we give it tools and authority.

The gap between "knows the answer" and "can act on the answer" is what separates a language model from an agent.

From Chatbot to Agent: An Evolution



What you see: The evolution from rule-based chatbots (2016) through transformer-based assistants (2020) to tool-using copilots (2023) and autonomous agents (2024–2026). Each generation adds capability: understanding, reasoning, tool use, and finally autonomy.
Key pattern: Each step roughly doubled the scope of tasks AI could handle in finance.

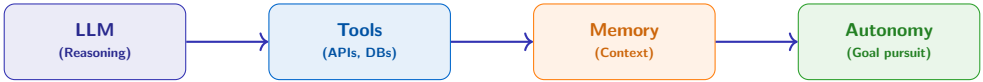
Takeaway: We are at the "agent" inflection point — AI is transitioning from a tool humans use to a system that uses tools itself.

The chatbot-to-agent evolution mirrors the progression from calculators (tools humans use) to autopilots (systems that act within boundaries).

What Is an Agent? (Not a Chatbot)

Definition: AI Agent

An **AI agent** is a software system that uses a language model to **reason** about tasks, **use tools** to interact with the world, **remember** past interactions, and **act autonomously** toward a goal — with or without human supervision.



Agent = all four components working together

Key insight: A chatbot has only the first component (LLM). A copilot adds tools. An agent adds memory and autonomy — it can pursue goals across multiple steps without waiting for human instructions at each step.

Agent = LLM + Tools + Memory + Autonomy. Remove any one component and you have something less capable: a chatbot, a copilot, or a script.

The New Employee Analogy

Think of an AI agent like a **new hire** at a bank:

Capability	New Employee	AI Agent
Reads procedures	Studies the compliance manual	Ingests policy documents into memory
Uses systems	Logs into Bloomberg, SAP, email	Calls APIs for market data, CRM, email
Makes decisions	"This transaction looks suspicious"	Classifies transactions by risk score
Asks for help	Escalates to manager when unsure	Routes to human when confidence is low
Learns over time	Gets faster after 100 reviews	Stores outcomes in episodic memory

Key insight: The analogy breaks down in two important ways:
(1) Agents scale instantly — you can run 1,000 copies simultaneously.
(2) Agents do not get tired, bored, or distracted — but they also lack common sense and ethical intuition.

The new-employee analogy helps build intuition, but agents lack the judgment, ethics, and contextual awareness that humans develop over years.

Think-Pair-Share

"What financial tasks would you trust an AI agent to handle autonomously?"

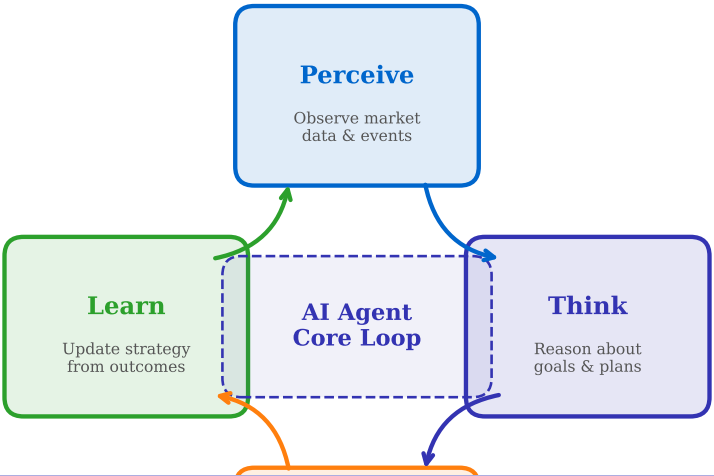
Instructions:

1. **Think** (2 minutes): List 3 financial tasks you would delegate to an agent and 3 you would not
2. **Pair** (3 minutes): Compare your lists with a neighbor — where do you disagree?
3. **Share** (2 minutes): One pair shares their most interesting disagreement with the class

Timer: 2 min think → 3 min pair → 2 min share.

Your gut reaction about which tasks to delegate reveals your assumptions about trust, risk, and the limits of automation — we will revisit this at the end.

The Agent Loop: Perceive-Think-Act-Learn



Definition: Chain-of-Thought (CoT)

Chain-of-Thought prompting is a technique where the language model is instructed to "think step by step" before producing an answer. This improves accuracy on multi-step reasoning tasks by 20–40 percentage points on benchmarks such as GSM8K (Source: Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models", NeurIPS 2022).

Worked Example — Loan Approval Reasoning:

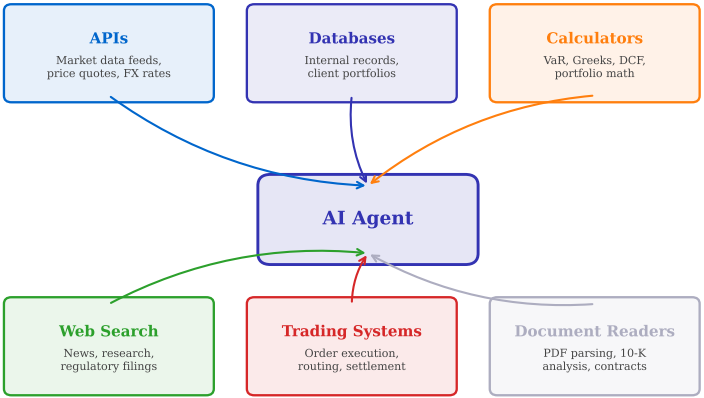
1. **Step 1:** Check credit score — 720 (meets threshold of 680)
2. **Step 2:** Calculate Debt-to-Income (DTI) ratio — Monthly debt \$2,100 / Monthly income \$7,000 = 30% (under 36% limit)
3. **Step 3:** Verify employment — 3 years at current employer (meets 2-year minimum)
4. **Step 4:** Check collateral — Property appraisal \$320,000 vs. loan \$250,000, Loan-to-Value (LTV) = 78% (under 80%)
5. **Conclusion:** All four criteria met → recommend approval

Key insight: Without Chain-of-Thought, the model might jump to "approve" without checking each criterion — and miss the one that fails.

Chain-of-Thought turns the LLM from a pattern-matcher into a step-by-step reasoner — essential for financial decisions that require multi-criteria evaluation.

Component 2: Tools

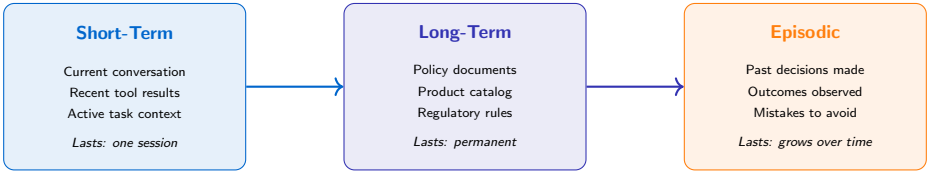
Tools Available to Financial AI Agents



What you see: The ecosystem of tools an AI agent can access — market data APIs (Bloomberg, Reuters), databases (SQL, vector stores), calculators (risk models, pricing engines), web search, email and messaging, and trading platforms (order management systems).
Key pattern: Each tool extends the agent beyond what the LLM can do alone — the LLM reasons about *which* tool to use and *how* to use it.

Takeaway: An agent is only as capable as its tool set. Adding a new tool (e.g., a regulatory filing API) instantly expands what the agent can do.

Component 3: Memory

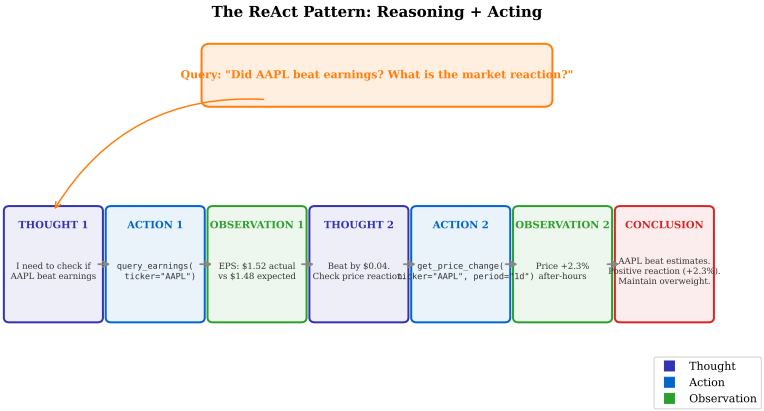


Together, these three memory types let an agent learn and improve over time

Key insight: Without memory, every conversation starts from zero. With memory, an agent can say: "Last time we discussed this client, they were concerned about interest rate risk — let me factor that in."

Memory is what separates an agent from a stateless chatbot — it enables context, learning, and personalization across interactions.

The ReAct Pattern: Reason + Act



What you see: The ReAct (Reasoning + Acting) loop introduced by Yao et al. (2023). The agent alternates between (1) Thought — reasoning about what information is needed, (2) Action — calling a tool to get that information, and (3) Observation — reading the tool's response. This cycle repeats until the agent has enough information to produce a final answer.
Key pattern: ReAct interleaves thinking and acting, rather than thinking first and acting later.
Takeaway: ReAct is the dominant architecture for financial AI agents because it produces auditable reasoning traces.

2026 frontier note: Since 2024, the frontier moved beyond pure ReAct toward *agentic reinforcement learning* and *computer-use* capability. Production deployments include Anthropic Claude Code, Cognition AI Devin, and OpenAI's ChatGPT Operator — agents that plan over long horizons and directly operate software interfaces.

Claude Code, Devin, ChatGPT Operator, computer-use agents (*Anthropic, Cognition AI, OpenAI product launches 2024–2025*, 2025–2026)

ReAct (Yao et al. 2023) remains the auditable baseline, but by 2026 the frontier is agentic RL and computer-use — Claude Code, Devin, Operator.

Worked Example: ReAct for Earnings Analysis

Task: "Should we increase our European bank stock position?"

THOUGHT 1: I need to check the European Central Bank (ECB) interest rate policy — bank profitability depends on it.

ACTION 1: `Search(''ECB latest rate decision 2026'')`

OBSERVE 1: ECB held rates at 3.5% in March 2026, signaling stability.
THOUGHT 2: Good for net interest margins. Now I need sector valuation — are European banks cheap?

ACTION 2: `Query(Bloomberg, ''EU bank sector P/E ratio'')`

OBSERVE 2: Current P/E = 8.2 vs. 10-year average of 10.5.
THOUGHT 3: Undervalued. But I should check credit risk — are spreads widening?

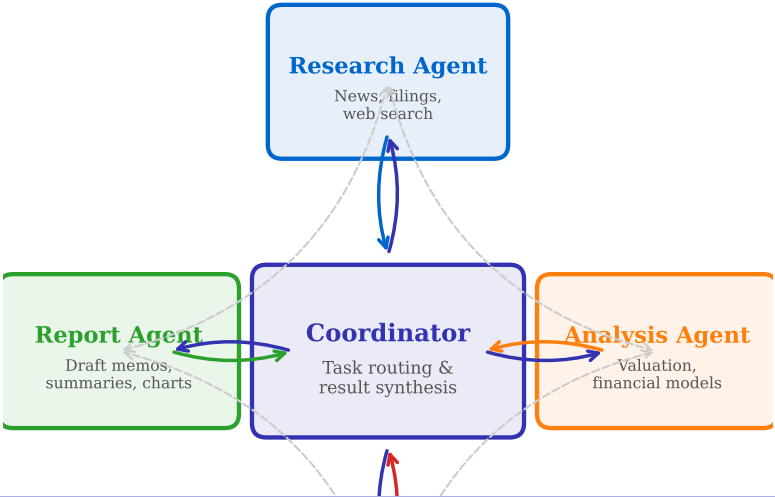
ACTION 3: `Query(Bloomberg, ''EU bank CDS spreads 3-month trend'')`

OBSERVE 3: CDS spreads narrowed 15 basis points over 3 months — improving credit outlook.

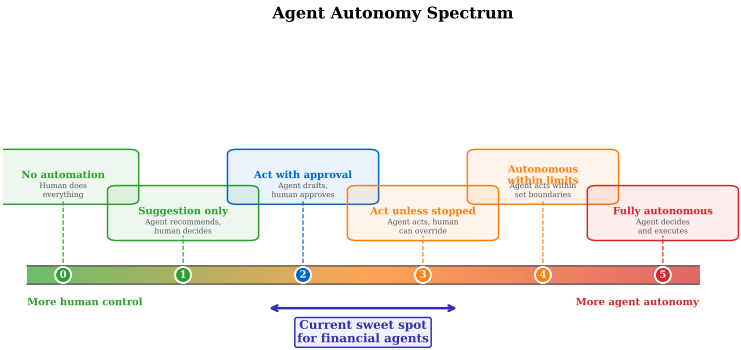
CONCLUSION: Undervalued + stable rates + improving credit risk → recommend moderate position increase, subject to portfolio risk limits.

Every step is auditable: a compliance officer can trace exactly why the agent reached its conclusion — this is why ReAct matters for finance.

Multi-Agent System for Financial Analysis



The Autonomy Spectrum: From Manual to Fully Autonomous



Level	Description	Finance Example	Human Role
0	Fully manual	Analyst reads 10-K, writes memo	Does everything
1	AI assists	LLM summarizes 10-K	Human decides and acts
2	AI recommends	Agent proposes trades with rationale	Human approves each action
3	AI acts, human vetoes	Agent executes within pre-set limits	Human monitors, can override
4	AI acts, human audits	Agent operates autonomously, logs all decisions	Human reviews after the fact
5	Fully autonomous	Agent manages portfolio end-to-end	Human sets goals only

Most financial agents today operate at Level 2–3. Level 5 remains aspirational and raises unresolved regulatory and ethical questions.

Four Application Domains

1. Trading

Portfolio rebalancing
Order execution
Market analysis

2. Compliance

Transaction monitoring
SAR generation
Regulatory reporting

3. Customer Service

Account management
Product recommendations
Dispute resolution

4. Research

Earnings analysis
Market intelligence
Report generation

Key insight: These four domains represent a spectrum from high-frequency/low-judgment (trading execution) to low-frequency/high-judgment (research). Agents are deployed differently in each.

We examine each domain in detail — the opportunities are real, but so are the domain-specific risks that come with autonomous action.

Trading Agents: Three Generations

Generation 1
Rule-Based (1990s)

- IF price drops 5% THEN buy
- Fixed rules, no learning
- Fast but brittle
- Fail in novel conditions

Generation 2
ML-Powered (2010s)

- Learn patterns from historical data
- Adapt to changing markets
- Random forests, neural nets
- Still react to numbers only

Generation 3
LLM-Powered (2024+)

- Read news, filings, reports
- Reason about context and causality
- Use tools (APIs, calculators)
- Explain their decisions in English

The shift: Generation 3 agents do not just respond to price data — they can read an earnings call transcript, assess management tone, check macro indicators, and *then* decide whether to trade.

Each generation expanded the information set: Gen 1 = price only, Gen 2 = price + features, Gen 3 = price + text + context + reasoning.

Case: Autonomous Portfolio Rebalancing

Scenario: A robo-advisor agent monitors a client's 60/40 stock/bond portfolio.

Step	Agent Action	What Happens
1. Monitor	Checks portfolio weights daily at 9:00 AM	Detects stocks drifted to 67% (threshold: 5%)
2. Analyze	Pulls market data, assesses volatility	VIX at 18, no macro event — safe to rebalance
3. Plan	Calculates trades to restore 60/40 split	Sell $7,000 equities, buy $7,000 bonds
4. Check	Verifies against risk limits and tax rules	Trade is within daily limit; no wash-sale issue
5. Execute	Submits orders via broker API	Market orders filled at 9:32 AM
6. Report	Sends client a rebalancing notification	"Portfolio rebalanced: 67% → 60% equities"

Key insight:

This agent operates at **Level 3** autonomy — it acts within pre-set rules (5% drift threshold, daily trade limits) but does not ask permission for each trade.

Autonomous rebalancing is one of the most mature agent applications — the rules are clear, the actions are bounded, and the downside is limited.

Compliance Agents: Real-Time Watchdogs

Compliance is the most labor-intensive function in banking. Agents can automate three core tasks:

Transaction Monitoring

- Scans every transaction in real time
- Flags patterns matching Anti-Money Laundering (AML) typologies
- Current systems: 95%+ false positive rate
- Agent advantage: reads context (client history, news) to reduce false positives

SAR Generation

- Drafts Suspicious Activity Reports automatically
- Pulls relevant transaction data, client profile, prior alerts
- Human compliance officer reviews and submits
- Typical time savings: $\sim$4 hours $\rightarrow$ 20–30 minutes per SAR (Source: Thomson Reuters / NICE Actimize compliance-automation case studies)

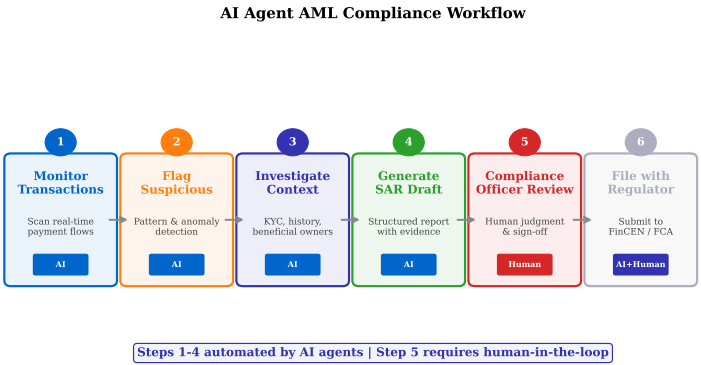
Regulatory Reporting

- Reads new regulations and maps to internal controls
- Generates quarterly reports (MiFID II, Basel III)
- Cross-checks data across systems
- Flags inconsistencies before submission

Key insight: Compliance agents operate at Level 2–3: they draft and recommend, but a human officer must approve submissions to regulators.

Compliance is the "killer app" for financial agents — high volume, rule-heavy, and currently drowning in manual work and false positives.

Case: An AML Agent at Work



What you see: The workflow of an AML compliance agent: (1) ingests transaction stream, (2) flags suspicious patterns using ML scoring, (3) enriches alerts with client history, news, and prior SARs, (4) drafts a Suspicious Activity Report with supporting evidence, (5) routes to human compliance officer for review and filing.
Key pattern: The agent does the heavy lifting (data gathering, pattern matching, drafting) but a human makes the final filing decision.

Takeaway: This "human-in-the-loop" design reduces analyst workload by 70–80% while maintaining regulatory accountability (illustrative).

AML agents do not replace compliance officers — they free them from data gathering so they can focus on judgment calls about genuine suspicious activity.

Customer Service Agents: Beyond the FAQ Bot

Modern customer service agents go far beyond answering questions — they can **act** on accounts:

Capability	Traditional Chatbot	AI Agent
Account inquiry	"Your balance is $5,230"	"Your balance is $5,230. I notice a $200 recurring charge — would you like me to analyze your subscriptions?"
Product suggestion	Generic banner ad	"Based on your saving pattern and upcoming trip, a travel rewards card could save you $180/year"
Dispute handling	"I'll transfer you to a specialist"	Reviews transaction, pulls merchant data, drafts resolution — escalates only complex cases
Risk alert	(none)	"Your card was used in two countries within 3 hours — I've temporarily frozen it. Shall I keep the freeze?"

Key insight:

The agent does not just inform — it takes initiative. This raises a design question: how proactive should a financial agent be before it feels intrusive?

Proactive service agents improve customer satisfaction but must balance helpfulness with privacy — knowing too much about spending feels invasive.

Research Agents: Earnings Analysis at Scale

What a Research Agent Does:

1. **Reads** the full 10-K filing (200+ pages) in seconds
2. **Extracts** key metrics: revenue, margins, guidance, risk factors
3. **Compares** current quarter to prior quarters and analyst consensus
4. **Identifies** tone shifts in management commentary (sentiment analysis)
5. **Generates** a structured memo with data tables, key quotes, and a risk summary

Human analyst time: 4–6 hours per company
Agent time: 10–15 minutes per company

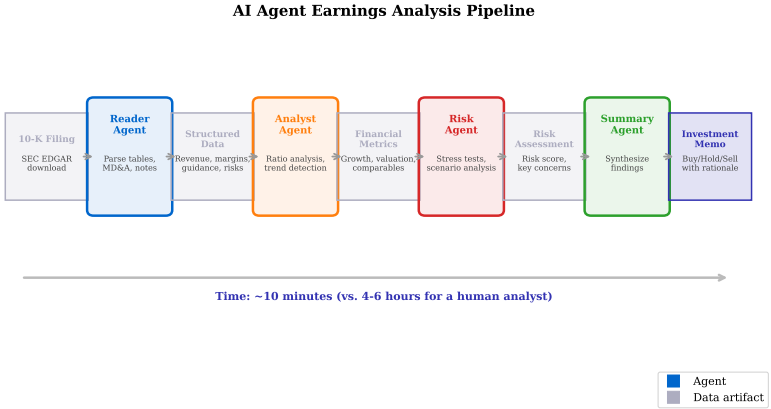
What It Cannot (Yet) Do:

- Detect management deception that is not in the text
- Understand industry context built over years of coverage
- Make judgment calls about "soft" factors (culture, leadership quality)
- Network with company insiders for off-the-record context
- Take responsibility for investment recommendations

Key insight: Research agents are excellent *first-pass analysts* — they do the data work so human analysts can focus on judgment and insight.

Research agents compress the data-gathering phase from hours to minutes — but the interpretation phase still requires human expertise and accountability.

Case: Multi-Agent Earnings Pipeline



What you see: A multi-agent pipeline for earnings analysis: (1) Ingestion Agent downloads the 10-K filing and earnings call transcript, (2) Extraction Agent pulls key financial metrics and management quotes, (3) Comparison Agent benchmarks against consensus estimates and prior periods, (4) Writing Agent generates a structured research memo, (5) Review Agent checks for errors, hallucinations, and internal consistency.
Key pattern: Each agent is specialized — the pipeline is modular and auditable at every step.

Takeaway: A 5-agent pipeline can process 50 earnings reports overnight — work that would take an analyst team a week.

Multi-agent pipelines mirror how analyst teams work: one person gathers data, another compares, another writes — the agents just do it 100 times faster.

What Banks Are Actually Deploying (2026)

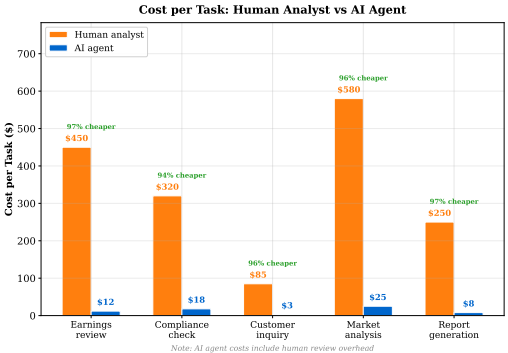
Bank	Initiative	What It Does	Autonomy Level
Goldman Sachs	Internal AI assistant	Helps developers write code, analysts draft memos	Level 1 (assists)
Morgan Stanley	AI @ Morgan Stanley	Wealth advisors query research library via chat	Level 1 (assists)
JPMorgan Chase	LLM Suite + IndexGPT	Contract analysis, investment recommendations	Level 1–2 (recommends)
Bloomberg	BloombergGPT	Financial NLP for terminal queries	Level 1 (assists)
Klarna	Customer service agent	Handles 2/3 of customer chats autonomously	Level 3 (acts within limits)
Ant Group	Compliance agent	Real-time transaction monitoring for Alipay	Level 3 (acts within limits)

Pattern: Most deployments are at Level 1–2 (assist and recommend). Only customer service and compliance have reached Level 3 (act within limits). No bank has deployed Level 4–5 agents in production. **Key insight:** The gap between "demo" and "deployment" is enormous. Regulation, liability, and trust are the bottlenecks — not technology.

Sources: Company press releases and earnings calls, 2024–2026. Specific implementations are illustrative summaries.

The industry is moving cautiously: assist first, automate later. Klarna's customer service agent is the most aggressive deployment to date.

The Cost Equation



What you see: Cost comparison between human-only and agent-assisted workflows for earnings analysis.

Key pattern: Agent costs are dominated by human review, not compute.

Worked Example: 10 Earnings Reports

Human-Only:

- 10 reports $\times$ 4 hours each = 40 hours
- Analyst cost: \$75/hour (hypothetical rate)
- **Total: \$3,000**

Agent-Assisted:

- Agent processing: 2.5 hours compute
- LLM token cost: \$1.50 (order-of-magnitude estimate)
- Human review: 5 hours $\times$ \$75 = \$375
- **Total: \$377**

Savings: 87%

Figures are order-of-magnitude BSc-level estimates; actual costs vary by firm, model, and complexity.

The 87% cost reduction is compelling — but only if the agent's output quality is high enough that human review time stays at 30 minutes per report.

Discussion: The Agent's Dilemma

An AI agent discovers evidence of fraud at a major client.

Reporting it would crash the client's stock price.

What should the agent do?

Consider these perspectives:

1. **Legal obligation:** Financial institutions are required to report suspected fraud (SAR requirements)
2. **Client relationship:** The bank earns $10M/year in fees from this client
3. **Market impact:** 50,000 retail investors hold this stock
4. **Agent autonomy:** Should the agent file the report automatically, or escalate to a human?

Group discussion: 5 minutes. Each group presents a recommendation with reasoning.

This dilemma has no clean answer — it illustrates why fully autonomous agents in high-stakes finance remain controversial. The "right" action depends on values, not just rules.

The Hallucination Problem

Definition: Hallucination

A **hallucination** occurs when a language model generates information that is fluent, confident, and completely false. The model does not "know" it is wrong — it has no internal fact-checking mechanism.

Why hallucinations are especially dangerous in agents:

LLM without tools:

- Hallucinates a wrong number
- Human reads it, notices it looks odd
- Human checks the source
- No real-world consequence

Agent with tools:

- Hallucinates a wrong number
- Uses that number in a calculation
- Passes calculation to trading API
- Executes a trade based on false data
- **Real money lost**

Key insight: When an LLM hallucinates, it produces a wrong answer. When an *agent* hallucinates, it produces a wrong *action*. The stakes are fundamentally different.

Hallucination risk is amplified by autonomy: the more an agent can do without human oversight, the more damage a hallucination can cause.

Case: Hallucinated Earnings

Scenario (hypothetical): A research agent analyzes Q3 earnings for a major tech company.

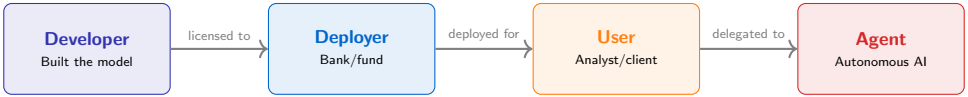
Event	What Happened	Consequence
Agent reads 10-K	Correctly extracts revenue: $4.1B	—
Agent extracts earnings	**Hallucination:** reports EPS $2.30	Actual EPS was $1.80
Agent calculates P/E	Uses hallucinated EPS → P/E looks cheap	Stock appears undervalued
Agent recommends	"Buy — undervalued relative to sector"	Recommendation is based on false data
Trading agent acts	Executes $500K purchase order	Portfolio now holds overpriced stock
Discovery	Human analyst spots error 3 days later	Loss: $45K (stock dropped 9%)

Root cause: The EPS figure appeared in a footnote about a *different* quarter. The agent lacked the ability to cross-reference the number against the correct time period.

Lesson: Agents need **verification steps** built into their pipeline — a second agent or tool call that cross-checks extracted data against known sources.

This scenario is hypothetical but realistic — hallucinated financial figures are one of the most commonly reported failure modes in LLM-based finance tools.

The Liability Question: Who Pays When the Agent Is Wrong?



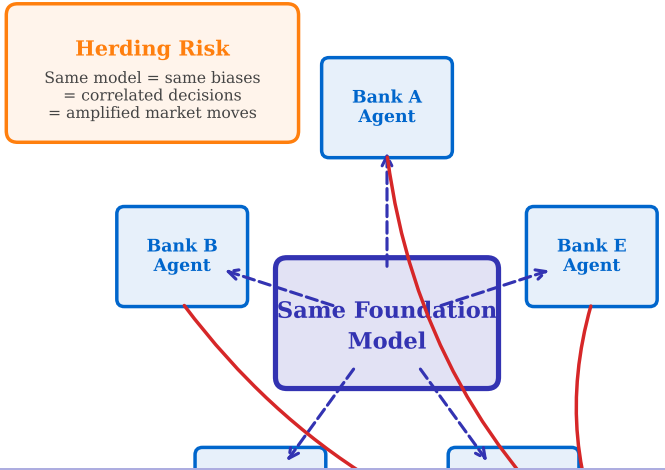
Agent makes a costly error. Who is liable?

- **Developer?** "We provided the model, not the deployment decisions"
- **Deployer (bank)?** "We followed best practices and set guardrails"
- **User?** "I trusted the system the bank provided me"
- **Agent?** An agent has no legal personhood — it cannot be sued

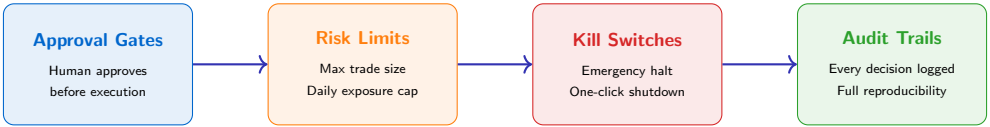
Key insight: Current law has no clear answer. The EU AI Act assigns primary responsibility to the **deployer** (the bank), but case law is still developing.

The liability chain is the single biggest unresolved legal question for AI agents in finance — until courts decide, banks deploy conservatively.

Systemic Risk: When All Agents Think Alike



Guardrails: Human-in-the-Loop Design



Defense in depth: multiple layers, not a single control

Best practice for financial agents:

- **Graduated autonomy:** Start at Level 1, earn trust over time with a track record
- **Confidence thresholds:** Agent only acts autonomously when confidence exceeds 95%; below that, it asks a human
- **Scope boundaries:** Agent can rebalance within 5% bands but cannot change asset allocation
- **Time limits:** Agent pauses and reports if a task takes longer than expected

---

No single guardrail is sufficient — financial agents need layered controls: approval gates + risk limits + kill switches + audit trails working together.

The EU AI Act and Financial AI Agents

The EU AI Act (Regulation 2024/1689, effective August 2025) classifies AI systems by risk level. Most financial AI agents fall under **high-risk**:

Risk Level	Finance Examples	Requirements
Unacceptable	Social scoring for credit	Banned outright
High	Credit scoring, insurance pricing, fraud detection, trading agents	Risk management system, human oversight, transparency, data governance, conformity assessment
Limited	Customer chatbots	Must disclose that user is interacting with AI
Minimal	Spam filters, basic analytics	No specific requirements

Key requirements for high-risk financial agents:

1. **Human oversight** — a qualified person must be able to override or stop the agent
2. **Transparency** — users must know they are interacting with an AI, and the system's capabilities and limitations must be documented
3. **Technical documentation** — the agent's architecture, training data, and decision logic must be recorded
4. **Record-keeping** — all agent actions must be logged for a minimum of 5 years

The EU AI Act is the world's first comprehensive AI regulation — financial agents must comply by August 2026. Source: EU Regulation 2024/1689.

AI Agents in Finance: Benefits vs. Risks

Benefits

1. **Speed** — processes 10-K filings in seconds, not hours
2. **Scale** — monitors millions of transactions simultaneously
3. **Consistency** — applies rules uniformly without fatigue or bias
4. **Cost** — reduces analyst workload substantially (specific per-firm ranges published by McKinsey AI 2024 and JPMorgan LOXM/COIN case studies)
5. **Availability** — operates 24/7, no sick days or holidays
6. **Auditability** — every reasoning step is logged and traceable

Risks

1. **Hallucination** — confidently generates false financial data and acts on it
2. **Liability gap** — unclear who is responsible when agents err
3. **Systemic correlation** — agents using the same model herd together
4. **Over-reliance** — humans stop checking the agent's work
5. **Opacity** — reasoning traces can be misleading (post-hoc rationalization)
6. **Data privacy** — agents process sensitive client data at scale

Bottom line: The benefits are real and measurable. The risks are also real but harder to quantify — which is why regulation lags deployment.

Benefits are immediate and quantifiable; risks are probabilistic and hard to price — this asymmetry explains why adoption moves faster than governance.

What AI Agents Should **Not** Do (Yet)

Five red lines for autonomous AI in finance (as of 2026):

1. **Unsupervised lending decisions** — Approving or denying loans without human review violates fair-lending regulations and creates discrimination risk
2. **Large autonomous trades** — Executing trades above a material threshold (e.g., ¿$1M) without human approval creates unacceptable downside risk
3. **Client data access without consent** — Agents must not access or combine client data beyond what the client explicitly authorized
4. **Regulatory filings without human sign-off** — SARs, tax filings, and Basel reports require a named responsible officer — an agent cannot be that officer
5. **Irreversible decisions under uncertainty** — Closing accounts, liquidating portfolios, or terminating client relationships should never be autonomous

The principle: Agents should handle the **volume** so humans can handle the **judgment**. Reversible, bounded, rule-based tasks are safe to automate. Irreversible, high-stakes, judgment-heavy tasks are not.

This is not a permanent list — as agents prove reliability and regulation catches up, some red lines will move. But in 2026, these boundaries protect clients, firms, and markets.

2026 State of Play: Works, Experimental, Hype

Works Today

- Customer service agents (Klarna, bank chatbots)
- Document summarization and extraction
- Code generation assistants for developers
- Basic compliance alert triage
- Portfolio drift monitoring and rebalancing

Experimental

- Multi-agent research pipelines
- Autonomous SAR drafting
- Real-time regulatory change monitoring
- AI-assisted credit decisioning
- Personalized financial planning agents

Hype (Not Yet Real)

- Fully autonomous trading funds
- AI replacing portfolio managers
- "One agent runs the entire bank"
- Self-improving agents without oversight
- AI-to-AI negotiation of financial contracts

Key insight: Most of what you read in tech media about AI agents in finance falls in the "Hype" column. What actually works today is narrower but genuinely useful.

Separating hype from reality is a critical skill — most "AI agent" announcements describe Level 1–2 systems, not the autonomous agents the headlines imply.

The Next 3 Years: What Is Coming

Timeline	Development	Impact	Probability
2026–2027	Agent-to-agent communication protocols	Agents at different banks negotiate terms directly	Medium
2026–2027	Regulatory sandboxes for agent testing	Controlled environments to test Level 3–4 agents	High
2027–2028	DeFi agents managing on-chain portfolios	Agents that interact with smart contracts autonomously	Medium
2027–2028	Agent-to-agent payment settlement	Automated B2B payments without human intervention	Medium
2028–2029	Standardized agent audit frameworks	Industry-wide standards for agent logging, testing, certification	High

Key insight:

The bottleneck is not technology — it is governance. The agents already work. What is missing is the regulatory and organizational infrastructure to deploy them responsibly.

All probabilities and timelines are conceptual estimates for educational purposes.

Technology leads, regulation follows — the next 3 years will be defined by governance frameworks catching up to agent capabilities.

New Job Roles: AI Agent Supervisors

As agents take over routine tasks, new roles emerge to supervise them:

New Role	What They Do	Skills Required
AI Agent Manager	Monitors agent performance, sets autonomy levels, handles escalations	Finance domain expertise + AI literacy
Prompt Engineer (Finance)	Designs agent instructions, tunes reasoning chains, tests edge cases	NLP understanding + financial regulation
Agent Compliance Officer	Ensures agents meet EU AI Act requirements, maintains audit trails	Legal/compliance + technical documentation
Human-AI Team Lead	Designs workflows where humans and agents collaborate effectively	Process design + change management

Key insight: AI agents do not eliminate financial jobs — they transform them. The analyst who spent 4 hours reading a 10-K now spends 30 minutes reviewing the agent's analysis and 3.5 hours on client strategy. The work shifts from **data processing** to **judgment and supervision**.

The job market is not "humans vs. agents" — it is "humans who work with agents vs. humans who do not." Adaptability is the career advantage.

The Centaur Model: Human + AI Teams

The Chess Analogy:

- 1997: IBM Deep Blue beats Kasparov
- 2005: "Freestyle" chess tournaments allow human-AI teams
- Result: **Human + AI teams beat both the best humans and the best AI playing alone**
- These teams were called "centaurs" — half human, half machine

Key insight: The winning combination is not the strongest AI or the smartest human — it is the best *process* for combining human judgment with AI speed.

Applied to Finance:

- **AI strength:** Speed, scale, consistency, data processing, pattern recognition
- **Human strength:** Judgment, ethics, context, creativity, client relationships, accountability
- **Centaur approach:** Agent handles data and drafts; human reviews, decides, and takes responsibility

Example: An analyst team with an AI agent produces significantly more research output at higher quality than a team without one — not because the AI is smarter, but because it frees humans to do what they do best (Source: Brynjolfsson et al., "Generative AI at Work", NBER 2023, which reports 14% average productivity gains with larger gains for junior staff).

The centaur model is the most practical framework for deploying AI agents in finance: combine AI speed with human judgment, not replace one with the other.

Key Takeaways

1. **An agent is an LLM + tools + memory + autonomy** — it does not just answer questions, it pursues goals by reasoning, using tools, and learning from outcomes.

2. **ReAct (Reason-Act-Observe) is the dominant architecture** — it produces auditable reasoning traces that regulators and compliance officers can review.

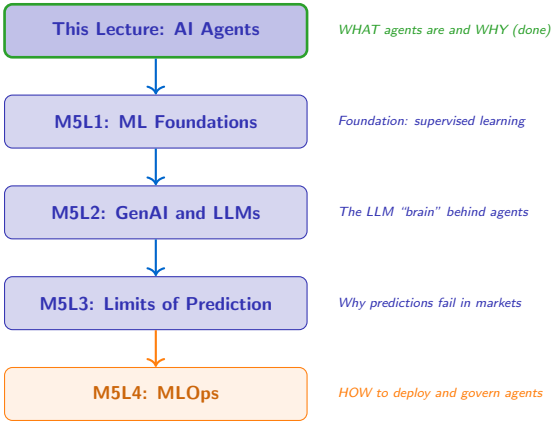
3. **Four domains lead adoption: trading, compliance, customer service, research** — but most real deployments are at Level 1–2 autonomy (assist and recommend), not Level 4–5 (autonomous).

4. **Agent-specific risks are real: hallucination, liability gaps, systemic correlation** — these are not theoretical concerns but active deployment challenges.

5. **The EU AI Act classifies most financial agents as high-risk** — requiring human oversight, transparency, documentation, and record-keeping. Compliance is mandatory by 2026.

These five takeaways map directly to the five learning objectives on Slide 3 — revisit those objectives to check your understanding.

Bridge: Agents Meet MLOps



You now understand WHAT AI agents are and WHY they matter for finance.

Next: Module 5 Lesson 4 (MLOps) will teach you **HOW** to deploy, monitor, and govern these agents in production — including model versioning, monitoring, and rollback procedures.

This lecture provides the "what and why" of agents; M5L4 (MLOps) provides the "how" of deploying and governing them safely in production.

Further Reading

Recommended resources for deeper study:

1. **Yao, S. et al. (2023).** "ReAct: Synergizing Reasoning and Acting in Language Models." *ICLR 2023.*
   The foundational paper on the ReAct architecture used by most financial agents.

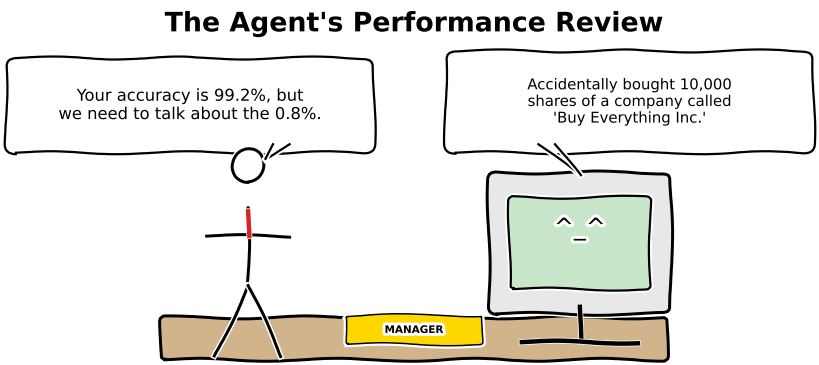
2. **European Parliament (2024).** "Regulation (EU) 2024/1689 — The AI Act." *Official Journal of the European Union.*
   Full text of the EU AI Act — read Articles 6–9 (high-risk classification) and Annex III (finance-specific rules).

3. **Wei, J. et al. (2022).** "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *NeurIPS 2022.*
   The paper that introduced Chain-of-Thought (CoT) prompting — the reasoning engine behind agent decision-making.

4. **Anthropic (2025).** "Building Effective Agents." *Anthropic Research Blog.*
   Practical guide to agent architecture: tool use, memory, and safety patterns.

5. **Bank for International Settlements (2024).** "Artificial Intelligence in Financial Services." *BIS Papers No. 145.*
   Central banker perspective on AI agent risks, with focus on systemic implications.

Start with the Anthropic guide for practical architecture, then the ReAct paper for theory, then the EU AI Act for regulatory context.



The question is not whether AI agents will act in finance — it is whether we will build the guardrails before or after the first major failure.

Appendix: The Klarna Reversal (2023 → 2025)

The cleanest public case study in productivity-narrative reversal. Klarna was the industry's most-cited AI-agent success story — until its own CEO revised the claim.

The "Level 3 customer service agent" story — original version

- **Feb 27 2024:** Klarna + OpenAI press release — AI assistant handling 2.3M conversations, equivalent to 700 FTEs, in first month (*Klarna press release, Feb 27 2024*, 2024)
- **Dec 2023 – Aug 2024:** Klarna imposes an external-hiring freeze (*Sebastian Siemiatkowski, Bloomberg interview, Dec 2023*, 2023); headcount falls from ~4500 to ~3500 (*Klarna Q2 2024 IPO filing F-1, revenue per employee disclosure*, 2024)
- CEO Sebastian Siemiatkowski's public positioning: "AI can already do all of the jobs that we as humans do" (*Siemiatkowski, World Economic Forum Davos, Jan 2024*, 2024)
- This became the canonical "agents replace staff" anecdote in 2024 finance-AI decks (including earlier versions of this lecture)

The reversal — May 2025

- **May 8 2025:** Siemiatkowski to Bloomberg: "Really investing in the quality of the human support is the way of the future for us" (*Bloomberg, "Klarna CEO Walks Back AI Replacing Humans," May 8 2025*, 2025)
- Klarna announces re-hiring human customer-service staff (*FT, "Klarna rehires humans after AI service push," May 9 2025*, 2025) — initially as a pilot gig-worker model (remote, per-hour), later formalized
- Siemiatkowski's concession: "Cost unfortunately seems to have been a too predominant evaluation factor" (*Bloomberg interview, May 8 2025*, 2025) — i.e., they cut humans for price, not quality
- Broader pattern: Duolingo contractor rollback (2024), DPD-chatbot public-shaming incident (Jan 2024), Air Canada chatbot liability ruling (Feb 2024) — see M5L4 MLOps

Pedagogical lesson: productivity claims about AI agents require a 24–36 month observation window before asserting. A