THE FUTURE-READY CFO MASTERCLASS

MODULE 1 OF 20 | PILLAR I: AI, TECHNOLOGY AND DIGITAL INTELLIGENCE

The AI-Literate CFO

MODULE SUMMARY

This module prepares the chief financial officer to judge artificial intelligence with the discipline applied to any large investment. It explains generative models and their limits, agentic systems and human controls, and how machine learning, generative and agentic approaches differ. It examines compute and token costs, build versus buy, hidden recurring costs and vendor dependence, and the highest value uses across finance. A fictional thirty million dollar proposal is tested with a verified scenario model. The module separates demonstrations from production systems, sets data prerequisites, guides advice to the board, points to the NIST framework, and builds an investment scorecard.

MODULE AT A GLANCE

Core question	Can you explain what these systems do and fail to do, judge whether a proposal to spend millions on them will pay back after risk and hidden cost, and say so plainly to the CEO and the Board?
Why it matters to the CFO	Capital for AI is requested now, vendors present polished demonstrations, and the CFO is the officer who must separate a promise from a production result before the money leaves.
Key sources	Deloitte CFO Signals and State of AI in the Enterprise, KPMG Global AI in Finance 2026, EY Tax and Finance Operations Survey, RSM 2025 AI Survey, Grant Thornton 2026 AI Impact Survey, NIST AI Risk Management Framework.
Executive artifact	An AI investment scorecard, completed for the fictional thirty million dollar proposal, that converts the evidence into a weighted score and a funding decision.

Important notice. This material is provided for educational and informational purposes only and does not constitute accounting, tax, legal, investment, or other professional advice. Readers should consult qualified professionals in accounting, tax, legal, and other relevant services before acting on any information presented. Fictional companies and figures are used for teaching and are labeled as such. The maturity ratings are an educational framework and have not been validated against a broad population of finance executives.

1. Executive overview and learning objectives

The core question. A chief financial officer is now asked, with growing frequency, to approve spending on artificial intelligence, which this module will call **AI**: computer systems that perform tasks that ordinarily require human judgment, such as reading a contract, drafting an explanation or forecasting demand. The question is seldom whether the technology is impressive, because the demonstration almost always is. The question is whether the CFO can explain what the system does and does not do, judge whether the money will come back after risk and hidden cost, and say so plainly to the chief executive and the Board. Knowing how to use a chat assistant is a useful skill and an inadequate qualification for that decision. The source proposal for this program puts the contrast well: using a chat tool is one thing, and deciding whether to commit 30 million dollars to an AI transformation, assess its return, govern its risks and explain the decision to the Board is quite another.

Why it matters over the next five years. Deloitte reported in January 2026 that 87 percent of the 200 North American chief financial officers it surveyed, at companies with at least 1 billion dollars of revenue, expect AI to be extremely or very important to their finance department's operations in 2026, and that 50 percent named the digital transformation of finance as their top priority for the year (Deloitte CFO Signals, fourth quarter of 2025, fieldwork from November 14 to December 7, 2025). Attention of that kind brings requests for capital. It also brings a risk that is easy to overlook, which is that a company may spend a great deal and learn little. Deloitte's separate survey of 3,235 business and technology leaders found that only 25 percent had moved 40 percent or more of their AI pilots into production, and KPMG found that 71 percent of 1,013 senior finance leaders said AI was meeting or exceeding return expectations while only 23 percent said it was exceeding them. The CFO who can tell a pilot from a production system, and a claimed saving from a saving that appears in the bank account, becomes the person the rest of the executive team relies on. The CFO who cannot becomes a reviewer of presentations.

The module is organized around one decision. Section 2 builds the vocabulary and the economics. Section 3 gives the methods, including a worked evaluation of a fictional 30 million dollar proposal whose arithmetic was verified in Python, a programming language widely used for financial modeling. Section 4 places the CFO in a fictional company at a decision point, Section 5 prepares the boardroom conversation, and Section 6 produces the executive artifact, an AI investment scorecard. Governance and risk are treated here only as far as they bear on the investment decision, and Module 10 develops them fully. After completing the module the reader should be able to do the following.

1. Explain in plain terms how traditional machine learning, generative AI and agentic AI differ, and which kind of problem each suits.
2. Describe what large language models do well and badly, including hallucination and reasoning failures, and name the control that answers each weakness.
3. Describe an AI agent, its tool access and its workflow, and set the level of human oversight that matches the consequence of an error.
4. Compute the full cost of an AI system, including inference or token charges, data work, monitoring and vendor dependence, and compare building with buying.
5. Evaluate an AI proposal using net present value, payback, internal rate of return, productivity that converts to cash, and risk adjusted value across scenarios.
6. Separate a demonstration from a production system, state the data and integration prerequisites, and advise the CEO and the Board in language that leads to a decision.

2. Conceptual foundations and frameworks

The vocabulary. A **model** is a mathematical system that has learned patterns from data and applies them to new inputs. **Machine learning** is the practice of building models from examples and not from hand written rules. **Training** is the costly process of building the model, and **inference** is the act of using a trained model to produce an answer, which occurs every time someone asks a question or a process runs. A **large language model**, or **LLM**, is a model trained on very large quantities of text to predict the next fragment of text, and it is the engine inside the chat assistants most people know. The fragments are called **tokens**, and a token is roughly three quarters of an English word. The text supplied to the model is the **prompt**, and the amount of text the model can consider at once is its **context window**. **Generative AI** is the family of systems, including LLMs, that produce new content such as text, images or code. Providers charge by the token for both the text sent in and the text returned, which is why the economics of these systems depend on how much text a process consumes.

Three kinds of AI. The first distinction a CFO should hold firmly is that the three kinds of AI used in business are different in what they produce, how they fail and what they cost. Traditional machine learning produces a number or a class, such as a probability of default or a forecast of units. Generative AI produces language or other content. **Agentic AI** goes one step further: it is a system in which a model is given a goal, can call tools such as a database, a spreadsheet or a payment system, decides which step to take next, and repeats until it judges the goal met. Exhibit 1 sets the three side by side.

Exhibit 1 Traditional Machine Learning, Generative AI and Agentic AI Compared

Feature	Traditional machine learning	Generative AI	Agentic AI
What it produces	A score, a class or a forecast	Text, summaries, drafts, code or images	A completed multi step task, with actions taken in other systems
Finance example	Probability that an invoice will be paid late	A first draft of the variance commentary for the monthly package	A system that reads invoices, matches them to orders, books them and routes exceptions
Typical failure	Drift: the world changes and the model grows stale	Fluent but wrong text, stated with confidence	A chain of small errors, or an action taken that no person intended
Cost pattern	Costly to build and clean data for, cheap per prediction	Little build cost, with a charge per token every time it runs	Token charges multiplied by many steps, plus integration and monitoring
Control that fits	Back testing, performance monitoring and model validation	Human review of output, grounding in source documents, testing on real samples	Limits on tools and amounts, approval gates, logs and the ability to stop
Maturity in finance	Long established in credit, fraud and forecasting	Widely tried, uneven in production	Early, with governance often unfinished

Source: author's framework for teaching, drawn from general knowledge and not from a single publication. The statement about governance being unfinished for agents is supported by Deloitte (January 2026), which found that only 21 percent of companies planning agentic deployments reported a mature model for agent governance.

What large language models do well and badly. An LLM is strong at reading, summarizing, drafting, classifying and translating between forms, such as turning a contract into a table of terms. It is weaker where the answer must be exactly right, because the model produces the most plausible continuation of text and not a checked fact. **Hallucination** is the term for output that is fluent and confident and nonetheless false, for example a citation to a standard that does not exist or a total that does not add. **Reasoning failures** are the

related errors in multi step logic or arithmetic, where the model skips a step, mishandles a negative sign or loses track of a condition stated early in a long document. These failures are not rare accidents that the next release will remove completely. They are properties of how the systems work, and they are reduced by design, by supplying the source documents so that the answer is grounded in them, by letting the model call a calculator instead of computing in its head, by testing on the company's own samples and by routing consequential output to a person. Three further limits matter to a CFO: the model may know nothing of events after its training, it may give a different answer to the same question on a different day, and text that arrives from outside, such as an email, may contain instructions that trick the system, which security specialists call **prompt injection**. Exhibit 2 pairs each weakness with a finance consequence and a control.

Exhibit 2 Limits of Large Language Models, Finance Consequences and Controls

Limit	What it looks like in finance	Control the CFO should expect
Hallucination	An invented precedent in a technical memorandum, or a covenant term that is not in the agreement	Answers tied to quoted source passages, reviewed by a qualified person before use
Reasoning and arithmetic error	A flux explanation whose figures do not reconcile to the ledger	Calculations done by code or a spreadsheet, with the model limited to explanation
Inconsistency	Two runs of the same contract review that disagree on a termination right	Repeated testing, fixed instructions and sampling of production output
Stale knowledge	A tax or accounting answer that ignores a recent change	Current documents supplied to the model, and a dated source on every answer
Prompt injection and data exposure	A supplier email that tells the system to change a bank account, or confidential data sent to an outside service	Restricted tool permissions, separation of untrusted text, contract terms on data use, access logs
Overconfidence	No signal that an answer is uncertain	Confidence thresholds, exception routing and a clear record of what was reviewed

Source: author's synthesis of widely described properties of language models and common practice. No statistic is claimed. NIST's Generative AI Profile (NIST-AI-600-1, July 26, 2024) addresses risks specific to generative AI and is read as a source in Section 7, but its individual risk categories were not reviewed for this exhibit.

Agents, tools and human oversight. An agent is more than a chat assistant because it can act. The **tools** it is given define what it can touch, for example reading the general ledger, creating a journal entry in draft, sending an email or releasing a payment. **Workflow orchestration** is the software that sequences the steps, passes information between them, keeps a log and decides when to stop or ask for help. A **human in the loop** control places a person at a defined point, usually before an action that cannot easily be reversed. The central design question is how much autonomy to grant, and the sensible answer depends on the cost of a wrong action, the volume, and the evidence gathered so far. Exhibit 3 offers a five level ladder that the CFO can use to ask a vendor or an internal team where a proposed system sits.

Exhibit 3 A Five Level Ladder of Autonomy for AI in Finance

Level	Description	Finance example	Typical oversight
0. Assist	The model answers questions or drafts text. A person decides and acts	Draft of the board commentary on gross margin	Person edits and approves everything
1. Recommend	The system proposes an action with its reasoning. A person executes	Proposed account coding for an invoice	Person accepts or rejects each item

Level	Description	Finance example	Typical oversight
2. Act with approval	The system prepares the action in the system of record and waits	Draft journal entry posted to a holding queue	Named approver releases, with a log
3. Act within limits	The system acts alone below stated amounts and confidence, and escalates the rest	Auto approval of matched invoices under a set amount	Sampling, exception reports, limits set by the CFO
4. Act autonomously	The system pursues goals across systems with little review	Self directed collections and payment timing	Continuous monitoring, kill switch, independent testing

Source: author's framework for teaching. Grant Thornton (April 2026) reports that 95 percent of surveyed leaders do not allow agents to make fully autonomous, high stakes decisions without human review; the survey is self reported by nearly 1,000 senior leaders in the United States.

The economics of AI. The price a vendor quotes is rarely the cost of owning the system. **Compute** is the processing power needed to train and run models, usually rented from a cloud provider. **Inference cost** recurs with every use, and for an LLM it is the product of tokens consumed and the price per token, so a process that sends a long document to the model many times is far more expensive than one that sends a short question once. Beyond compute, the company must prepare its data, connect the system to the ledger and other applications, monitor accuracy, retrain staff, and carry the risk that a single provider changes its price, its terms or its model. Exhibit 4 separates the costs that usually appear in the proposal from those that usually do not. The pattern the CFO should remember is that build costs are visible and finite, whereas run costs are easy to understate and continue for as long as the system is used.

Exhibit 4 Visible and Hidden Costs of an AI System

Cost element	Usually in the proposal?	What the CFO should ask
Licences, platform fees and implementation services	Yes	Is the price fixed, capped or usage based, and what happens at renewal?
Inference and token charges	Often only a low estimate	What volume, how many model calls per task and what price per million tokens were assumed?
Data preparation, cleansing and integration	Seldom in full	Who pays to fix the source data, and who owns the interfaces afterward?
Internal team time and change management	Seldom	How many people from finance and technology are needed, and what do they stop doing?
Monitoring, testing, model risk review and audit support	Seldom	Who tests accuracy each month, and who reports results to the audit committee?
Price escalation, model changes and retraining	No	What does a change of model or a price rise cost in rework?
Exit and switching costs	No	Can data, prompts and workflows be moved, and at what cost?

Source: author's framework. EY (November 18, 2025) reports that 44 percent of 1,600 surveyed finance and tax leaders cited the inability to execute a sustainable plan for data, AI and technology as their biggest barrier, which is consistent with the data and team costs listed here. Module 4 develops technology investment economics in more detail.

A pointer to the NIST framework. The National Institute of Standards and Technology published version 1.0 of its voluntary **AI Risk Management Framework** on January 26, 2023, and a Generative AI Profile on July 26, 2024. The framework organizes the discipline around four functions named Govern, Map, Measure and

Manage, and the NIST page notes that the framework is being revised as part of the White House AI Action Plan. The reader should confirm the current text before relying on it. Exhibit 5 shows how each function appears as a question a CFO can ask during an investment review. The detailed treatment of governance, model risk and third party risk belongs to Module 10, and the CFO's responsibilities for controls and assurance over automated processes appear in Module 11.

Exhibit 5 The Four Functions of the NIST AI Risk Management Framework as Investment Questions

NIST function	Plain meaning	Question at the investment review
Govern	Policies, roles and accountability for AI across the organization	Who is accountable for this system, and which policy authorizes its use?
Map	Understanding the context, purpose, users and possible harms of a specific system	What decision does it support, who is affected, and what is the worst realistic error?
Measure	Testing and tracking how well the system performs and how it fails	What accuracy, bias and cost measures will we track, and who reports them?
Manage	Acting on the measured risks through controls, response plans and decommissioning	What do we do when it fails, and can we stop it quickly?

Source: NIST, AI Risk Management Framework page (version 1.0 released January 26, 2023; Generative AI Profile July 26, 2024). The names of the four functions were taken from the page; the plain meanings and the questions are the author's interpretation, since the page text did not describe the functions. Governance detail is in Module 10.

3. Practical applications and tools

Step one: price the tokens. The simplest piece of economics is also the one most often skipped. Take a process that handles 600,000 documents each year, and compare a single model call per document with an agent that works through eight calls. The prices used here, 3.00 dollars per million tokens sent in and 15.00 dollars per million tokens returned, are fictional and chosen only to make the arithmetic clear. Actual prices differ by provider and model, change often, and must be taken from the vendor's current price list. Exhibit 6 shows the calculation.

Exhibit 6 Token Arithmetic for a Document Process (fictional prices)

Item	Single call	Agent with eight calls
Input tokens per document	6,000	80,000 (8 calls of 10,000)
Output tokens per document	800	4,800 (8 calls of 600)
Cost of input at 3.00 per million	$0.018	$0.240
Cost of output at 15.00 per million	$0.012	$0.072
Cost per document	$0.030	$0.312
Annual cost at 600,000 documents	$18,000	$187,200
Annual cost if the price per token doubles	$36,000	$374,400

Fictional prices and volumes. Computed in Python: 6,000 x 3.00 / 1,000,000 = 0.018 and 800 x 15.00 / 1,000,000 = 0.012, a total of 0.030; 80,000 x 3.00 / 1,000,000 = 0.240 and 4,800 x 15.00 / 1,000,000 = 0.072, a total of 0.312, which is 10.4 times the single call. The lesson is the ratio and not the dollars: an agent that reconsiders the same context many times can cost ten times as much per task as a single call, so the number of calls belongs in the business case.

Step two: build, buy or partner. A company may build its own system on a general purpose model, buy a finished application from a vendor, or partner with a vendor that configures its platform for the company. The choice is a trade among control, speed, cost and dependence. Buying is faster and moves maintenance to the vendor, but it deepens dependence and may place company data in the vendor's hands. Building keeps control and a possible advantage, but it requires scarce skills and a team that must be paid every year. A sound rule is to buy what competitors also do, such as invoice capture, and to build only where the company's own data or process gives it something others cannot buy. Exhibit 7 summarizes the comparison.

Exhibit 7 Build, Buy or Partner: A Comparison for the CFO

Factor	Build	Buy	Partner
Time to first use	Slowest, often twelve months or more	Fastest, often weeks to a few months	Moderate
Up front cost	Highest, mainly people and data work	Lowest	Moderate, with services fees
Run cost	Team, cloud and model charges, all controllable	Subscription and usage fees set by the vendor	Both, with fees for configuration
Control over data and logic	Highest	Lowest	Moderate, depends on contract
Vendor dependence	Low for the application, but still reliant on a model provider	High	High to moderate
Best used when	The process and data are a source of advantage	The process is common to many companies	The company needs speed and lacks the skills to build

Source: author's framework, based on general practice and not on a cited publication. RSM (June 11, 2025) reports that 70 percent of 966 surveyed middle market respondents recognize the need for external support, which is why partnering is common.

Step three: choose use cases where value is highest. The highest value uses share three traits: the work is high in volume, the inputs are mostly text or documents, and a wrong answer can be caught cheaply. By that test the best early candidates in finance are the reading and matching tasks, and the weakest are decisions that are rare, consequential and irreversible. Exhibit 8 lists candidates by function, with an honest view of risk and readiness. The ratings are the author's judgment and each company should replace them with its own evidence.

Exhibit 8 Highest Value AI Use Cases Across Finance and Operations

Function	High value use cases	Main risk	Early readiness
Accounting	Invoice capture and coding, account reconciliations, flux commentary, lease and contract abstraction, intercompany matching	Wrong coding that passes review, weak audit trail	High for reading and matching, moderate for judgment
FP&A	Drafting variance commentary, scenario narratives, driver based forecasts that combine machine learning with finance input, board package preparation	Fluent but unsupported explanations, forecast bias	Moderate, depends on data quality
Treasury	Cash forecasting, bank statement parsing, covenant monitoring, payment anomaly detection	Wrong payment or forecast that drives a funding decision	Moderate, strong controls required

Function	High value use cases	Main risk	Early readiness
Order to cash	Collections prioritization, dispute resolution, cash application, pricing and billing leakage detection	Customer harm from wrong notices	High where records are clean
Procure to pay	Three way match, supplier onboarding checks, duplicate payment detection	Fraud through manipulated documents	High
Tax and legal	Research assistance, document comparison, drafting of first responses	Invented authority, loss of privilege	Moderate, qualified professionals must review
Operations	Demand forecasting, maintenance prediction, customer service assistance, supply chain exception handling	Operational error scaled by automation	Varies by data maturity

Source: author's judgment, informed by RSM (June 11, 2025), which reports that the most common use cases among its respondents were data analytics (58 percent), IT operations (57 percent) and customer service (48 percent), and KPMG (May 2026), which reports that 76 percent of surveyed organizations actively use AI in financial planning. Tax and legal items require review by qualified professionals.

Step four: tell a demonstration from a production system. A demonstration shows that a thing can work once, on chosen material, with a skilled person at the keyboard. A production system must work every day, on all material, with ordinary users, under controls, at a known cost. The distance between the two is where most of the cost and nearly all of the risk lie. Deloitte found that only 25 percent of respondents had moved 40 percent or more of their pilots into production, and the Grant Thornton survey found that only one in five organizations had tested a response plan for AI failures. Exhibit 9 lists the questions that separate the two conditions.

Exhibit 9 Demonstration Versus Production System: Eight Tests

Test	Demonstration	Production system
Data used	Selected, clean examples	All real data, including poor scans, exceptions and unusual formats
Accuracy	Reported on a handful of cases	Measured on a random sample, with an error rate and a cost per error
Volume and speed	A few items	Full daily volume and peak periods, with response time targets
Integration	Manual upload and copy and paste	Connected to the ledger and other systems, with access limited by role
Controls	None needed	Approvals, segregation of duties, logs and the ability to stop
Cost	Free or a pilot price	Full run cost including tokens, monitoring and support, at scale
Ownership	A project team	A named owner, an incident process and a retirement plan
Evidence	A presentation	Test results the auditor and the CFO can inspect

Source: author's framework. The statistics in the paragraph above come from Deloitte, State of AI in the Enterprise (January 21, 2026) and Grant Thornton, AI Impact Survey (April 13, 2026, self reported).

Step five: confirm the data and integration prerequisites. A model that reads poor data produces poor answers faster. EY reports that 51 percent of its 1,600 surveyed tax and finance functions were still in the early

stages of data management maturity, RSM reports that data quality was the most common rollout challenge among its generative AI users at 41 percent, and KPMG reports that 36 percent of its respondents named improving data quality, integration and interoperability as their greatest opportunity. Grant Thornton adds that 55 percent of the CIOs and chief technology officers it surveyed said most core applications were not ready for AI. Exhibit 10 turns these concerns into a short readiness checklist that should be completed before the main investment is released.

Exhibit 10 Data and Integration Prerequisites for AI Adoption

Prerequisite	Question to answer	Evidence the CFO should see
Data quality	Are the master data for customers, suppliers and accounts complete and consistent?	Error and duplicate rates by data set, with an owner for each
Integration	Can the system read from and write to the ledger and other applications through supported interfaces?	Architecture diagram and a tested interface, not a plan
Access control	Does the system see only what the user is allowed to see?	Role mapping, privileged access review, and log samples
Data rights and privacy	May the data be used with this model and this vendor, and where is it stored?	Contract terms, legal and privacy review by qualified professionals
Lineage and audit trail	Can each output be traced to the inputs and the version of the model that produced it?	Sample trace for ten transactions
Enterprise architecture	Does this fit the target architecture, or does it create another silo?	Sign off by the technology owner and a retirement plan for old tools

Source: author's checklist. Statistics from EY (November 18, 2025), RSM (June 11, 2025), KPMG (May 2026) and Grant Thornton (April 13, 2026), each as stated in the text. Data privacy and rights questions require advice from qualified legal and privacy professionals.

Step six: evaluate the 30 million dollar proposal. Northgate Industrial Group (fictional) has revenue of 1,200 million dollars. Its technology vendor and an internal sponsor propose an AI transformation across finance, procurement and customer operations for a one time cost of 30 million dollars, spread over three years. The sponsor expects that AI will save 600,000 hours of work each year, worth 65 dollars an hour fully loaded, or 39.0 million dollars, together with 9.6 million dollars a year of pricing and billing leakage, which is 0.8 percent of revenue, and a release of working capital from collecting two days faster. All amounts that follow are in millions of dollars, before tax, over a horizon of Year 0 and five further years, discounted at a 10 percent hurdle rate. Tax effects were left out for simplicity and would be considered with qualified tax advisers. **Net present value** is the sum of each year's cash flow discounted back to today at the hurdle rate, **payback** is the time until the cumulative cash flow turns positive, and **return on investment**, or **ROI**, is the net gain divided by the cost.The sponsor's case assumes that every hour saved becomes a dollar saved, that every dollar of leakage is recovered and that nothing costs more than quoted. The CFO's base case revises each of those assumptions, and the downside and upside cases bound the range. Exhibit 11 lists the assumptions.

Exhibit 11 Assumptions for the Fictional 30 Million Dollar Proposal (millions of dollars)

Assumption	Sponsor	Upside	Base	Downside
One time cost, Years 0 / 1 / 2	10.0 / 12.0 / 8.0	12.0 / 14.2 / 12.2	12.0 / 14.2 / 12.2	13.2 / 15.6 / 13.2
One time cost total	30.0	38.4	38.4	42.0
Run cost, Years 1 to 5 (total)	10.0	24.4	27.4	33.0

Assumption	Sponsor	Upside	Base	Downside
Share of hours saved that becomes cash	100%	55%	45%	25%
Share of leakage recovered	100%	80%	60%	30%
Ramp to full benefit, Years 1 to 5	30, 70, 100, 100, 100%	25, 60, 90, 100, 100%	20, 50, 80, 95, 100%	10, 30, 50, 60, 60%
Working capital released in Year 2	6.6 (2.0 days)	6.6 (2.0 days)	4.9 (1.5 days)	0.0
Probability weight	Not used	25%	50%	25%

Fictional company and figures. The base case one time cost of 38.4 is 28 percent above the 30.0 headline because it includes data remediation, internal team time and a contingency that the proposal omitted. Working capital: 1,200 / 365 x 2.0 days = 6.6 and x 1.5 days = 4.9. Gross hours value: 600,000 x 65 = 39.0. Leakage pool: 0.8 percent x 1,200 = 9.6.

Exhibit 12 **Base Case Cash Flows (millions of dollars, fictional)**

Year	Labor savings	Leakage recovered	Working capital	Total benefit	One time cost	Run cost	Net cash flow	Cumulative
0	0.0	0.0	0.0	0.0	12.0	0.0	(12.0)	(12.0)
1	3.5	1.2	0.0	4.7	14.2	3.2	(12.7)	(24.7)
2	8.8	2.9	4.9	16.6	12.2	5.0	(0.6)	(25.4)
3	14.0	4.6	0.0	18.6	0.0	6.0	12.6	(12.7)
4	16.7	5.5	0.0	22.1	0.0	6.4	15.7	3.0
5	17.6	5.8	0.0	23.3	0.0	6.8	16.5	19.5
Total	**60.5**	**19.9**	**4.9**	**85.3**	**38.4**	**27.4**	**19.5**	

Fictional. Computed in Python. Labor savings in Year 3, for example: 39.0 x 45 percent x 80 percent = 14.04; leakage 9.6 x 60 percent x 80 percent = 4.61; total benefit 18.65; run cost 6.0; net 12.65. Net present value at 10 percent is 6.4; the internal rate of return is 17.7 percent; payback occurs 3.8 years after the first outlay, when the cumulative column turns positive during Year 4.

The result for the base case is positive but thin. The net present value is 6.4 million dollars, the internal rate of return, which is the discount rate at which the net present value equals zero, is 17.7 percent against a 10 percent hurdle, and the investment is repaid only after 3.8 years. The sponsor's own case shows a net present value of 111.6 million dollars and payback in 1.3 years, which is the kind of result that should make a CFO suspicious and not delighted, since it requires every assumption to be right at once. Exhibit 13 compares all four cases, adds a probability weighted value, and shows what a stage gate adds.

Exhibit 13 **Scenario Results for the Fictional Proposal (millions of dollars)**

Measure	Sponsor	Upside	Base	Downside
Total benefits, Years 1 to 5	201.0	115.8	85.3	26.5
Total costs, Years 0 to 5	40.0	62.8	65.8	75.0
Net cash, undiscounted	161.0	53.0	19.5	(48.5)
Net present value at 10 percent	111.6	31.2	6.4	(43.6)
Internal rate of return	148%	46.0%	17.7%	Negative

Measure	Sponsor	Upside	Base	Downside
Payback, years from first outlay	1.3	2.7	3.8	Not within five years
Five year return on cost, (benefits less costs) / costs	403%	84%	30%	-65%
Probability weighted net present value (25, 50, 25)			0.1	
If downside is detected and stopped after the Year 1 gate				(29.5)
Probability weighted value with the stage gate			3.6	

Fictional. Computed in Python. Probability weighted: 0.25 x 31.2 + 0.50 x 6.4 + 0.25 x (43.6) = 0.1. With the gate, the downside case stops after Year 1, so only Years 0 and 1 are incurred: net present value (29.5), and the weighted value rises to 3.6. The probabilities are the author's illustration and are not empirical. Five year return on cost for the base case: (85.3 less 65.8) / 65.8 = 30 percent.

What the model teaches. Three lessons follow. First, the probability weighted value of the full commitment is about zero, 0.1 million dollars, so the proposal is a fair bet and not an attractive one when it is approved as a single block. Second, a stage gate improves the expected value to 3.6 million dollars, not because it changes the technology but because it limits the loss in the case where the benefits fail to appear. Third, the decisive variable is the share of saved hours that becomes cash. Exhibit 14 shows the base case net present value when that share and the one time cost vary together. The break even share, with other base assumptions unchanged, is 38.4 percent, which is the number the CFO should ask the sponsor to defend with named roles, attrition plans and contractor spend, because a saved hour that is not converted into reduced payroll, avoided hiring or avoided outside cost has produced capacity and no cash.

Exhibit 14 Base Case Net Present Value Sensitivity (millions of dollars, fictional)

Share of saved hours that becomes cash	One time cost as base	10% over base	20% over base	30% over base
25%	(12.8)	(16.3)	(19.8)	(23.3)
35%	(3.2)	(6.7)	(10.2)	(13.7)
45%	6.4	2.9	(0.6)	(4.1)
55%	16.0	12.5	9.0	5.5
65%	25.6	22.1	18.6	15.1

Fictional. Computed in Python by re-running the base case for each pair of values; the 45 percent row and zero overrun reproduce the base case value of 6.4. Break even share of saved hours that becomes cash, with all other base case assumptions unchanged: 38.4 percent.

A note on risk adjusted value. The phrase is used loosely, and the CFO should insist that the sponsor say which method is meant. There are three common ones. The first, used above, weights the outcomes by probability, which is simple and honest provided that the probabilities are discussed and not assumed. The second reduces each benefit by a haircut, for example by counting only 45 percent of labor savings, which is easy to explain but hides the fact that costs also carry uncertainty and tend to err in one direction. The third raises the discount rate to reflect risk, which is common in practice and has the drawback that a single higher rate penalizes distant years without saying why. None of the three is perfect, and each is better than ignoring the question. The strongest practice is to show the sponsor case, the base case and the downside side by side, to state what would have to be true for each, and to attach the stage gates that determine which of them

the company is living in after the first year.

Step seven: set adoption goals and track whether the benefit arrives. A proposal that is approved and never measured teaches the company nothing, and it leaves the next proposal to be judged on enthusiasm. KPMG reports that organizations that track key performance indicators for their AI report better return outcomes than those that do not, 68 percent against 58 percent, and that only 29 percent formally track where AI adoption fails. The CFO should therefore require that each approved investment carry a short list of measures with a baseline, a target at each gate, an owner and a source system, and that the first report is due before the second tranche of money is released. Exhibit 15 gives a starting list for the Northgate proposal, in which the gate after Year 1 is the point at which the downside case would be recognized and stopped.

Exhibit 15 Adoption and Benefit Tracking Measures for a Stage Gate (fictional)

Measure	Baseline and definition	Target at the Year 1 gate	Owner and source
Share of saved hours that becomes cash	Zero at the start. Hours released multiplied by the loaded rate, counted only if payroll, hiring or outside spend fell	At least 38.4 percent, the break even share, on the ramp of 20 percent	Controller, with human resources and procurement data
Share of eligible volume handled by the system	Documents processed by the system divided by eligible documents	20 percent of the full volume, matching the base ramp	Process owner, workflow log
Error rate on released output	Wrong amounts or wrong coding found by sampling, as a share of items	At or below the present rate, with no rise for two months	Internal audit sample, ledger
Cost per transaction, including tokens and review	Total run cost divided by items, compared with the present cost	Within 10 percent of the plan in the business case	FP&A, vendor invoices and cloud billing
Leakage recovered	Billing and pricing errors found and corrected, in dollars	12 percent of the 9.6 pool, which is 1.2 (60 percent recovery on a 20 percent ramp)	Revenue operations, credit memo records
One time cost against plan	Spend to date divided by the approved stage budget	No more than 10 percent over, with contingency use explained	Program office, project accounting
Incidents and near misses	Failures, stops and overrides of the system, with cause	All logged and reviewed. Response plan tested once	Technology risk, incident log

Fictional measures and targets for the Northgate example, computed in Python: the Year 1 base ramp is 20 percent, so 9.6 x 60 percent x 20 percent = 1.15, shown as about 1.2 in the table; the break even share of 38.4 percent comes from the sensitivity analysis. KPMG figures are from Global AI in Finance 2026 (May 2026).

4. Executive case study and decisions

Fictional case: the demonstration that did not survive contact with the invoices. Calder Fastener Corporation (fictional) is a manufacturer with 600,000 supplier invoices a year. The accounts payable team processes them at a fully loaded cost of 5.40 dollars each, 3,240,000 dollars a year. Its error rate on payments is 0.3 percent, or 1,800 wrong payments a year, with an average value of 1,150 dollars each, of which 20 percent is never recovered, a loss of 414,000 dollars. Total annual cost of the present process is therefore 3,654,000 dollars. A vendor has demonstrated an AI agent that reads invoices, matches them with purchase orders and receipts, and prepares the payment. In the demonstration, run on 200 invoices chosen by the vendor, 98 percent were handled correctly, and the vendor proposes a fee of 0.40 dollars per invoice,

240,000 dollars a year. The implied saving is 3,000,000 dollars a year, and the controller has asked the CFO to approve a full rollout.

The decision point. The CFO is wise to be skeptical and asks the vendor to run the agent on a random sample of 5,000 of last year's invoices, including damaged scans and unusual formats. The results are sobering. The agent handles 91 percent straight through, so that 9 percent, or 54,000 invoices, fall to exceptions that a person must resolve, at a cost of 9.00 dollars each because reviewing another party's error takes longer than doing the work from the start. In addition, the agent's own error rate on payments that go through unreviewed is 0.9 percent, three times the present rate. The agent also needs about eight model calls per invoice, which at the prices in Exhibit 6 costs 0.312 dollars per invoice, or 187,200 dollars a year, and the company would need two people at 120,000 dollars each to monitor accuracy and manage exceptions. The vendor's fee excludes the model charges. Exhibit 16 compares the demonstration with the production result under two operating designs, one with full autonomy and one with a person checking the 15 percent of invoices on which the agent is least confident, at 3.00 dollars each, which also lowers the payment error rate to 0.25 percent, below the present level.

Exhibit 16 Calder Fastener: Demonstration Claim Versus Production Economics (fictional)

Annual cost element (dollars)	Present process	Demonstration claim	Full autonomy in production	Human review of low confidence items
Processing labor at 5.40 per invoice	3,240,000	0	0	0
Vendor platform fee, 0.40 per invoice	0	240,000	240,000	240,000
Model charges, 0.312 per invoice	0	0	187,200	187,200
Exceptions, 54,000 at 9.00	0	0	486,000	486,000
Monitoring staff, two people	0	0	240,000	240,000
Review of 90,000 low confidence items at 3.00	0	0	0	270,000
Unrecovered payment errors	414,000	414,000	1,242,000	345,000
Total annual cost	**3,654,000**	**240,000**	**2,395,200**	**1,768,200**
Annual saving against the present process		3,000,000	1,258,800	1,885,800
Saving as a share of present cost		82.1%	34.4%	51.6%

Fictional. Computed in Python. The demonstration claim keeps the present error loss of 414,000 because the demonstration did not test payment errors: 240,000 + 414,000 = 654,000, a saving of 3,000,000 or 82.1 percent. Full autonomy: 240,000 + 187,200 + 486,000 + 240,000 + 1,242,000 = 2,395,200, a saving of 1,258,800 or 34.4 percent. Human review: 240,000 + 187,200 + 486,000 + 240,000 + 270,000 + 345,000 = 1,768,200, a saving of 1,885,800 or 51.6 percent.

The two production columns are worth studying together. Full autonomy saves 1,258,800 dollars a year and raises the payment error rate to three times the present level, so the company would be buying a saving partly with a rise in unrecovered payments, an exposure that the audit committee would not welcome. Adding a light human review of the least confident invoices costs 270,000 dollars and saves more overall, 1,885,800 dollars, because it avoids 897,000 dollars of unrecovered errors compared with the autonomous design. The gap between the demonstration claim of 3,000,000 dollars and the best production result is 1,114,200 dollars, or 37 percent of the claim.

The options. Exhibit 17 sets out four choices open to the CFO and the recommended path.

Exhibit 17 Calder Fastener: Options and Judgment (fictional)

Option	Advantages	Disadvantages	Judgment
A. Do nothing	No risk, no spend	Leaves a repeatable, high volume process unexamined, while competitors reduce cost	Weak. Fails the duty to test a promising tool
B. Approve the vendor's rollout with full autonomy	Highest headline saving if the demonstration held	Payment errors triple, saving is 1,258,800 and not 3,000,000, no stop rule	Reject
C. Approve a staged production rollout with human review of low confidence items	Saving of 1,885,800 with errors below the present rate, evidence gathered before expansion	Needs monitoring staff and a clear gate, saving is below the claim	Recommended
D. Decline and wait for better products	Avoids integration cost now	Delays learning and may leave the company behind in capability	Acceptable only if data readiness fails the checklist

Fictional. The recommendation rests on the arithmetic in Exhibit 16 and on the principle that autonomy should be earned by measured accuracy. The numbered options follow the ladder in Exhibit 3, with option C at level 3 for small and matched invoices and level 2 for the remainder.

The recommended path and the reasoning. The CFO should approve option C, as a three stage commitment. In the first stage the agent runs in shadow beside the team for one quarter, so that its answers are compared with the people's answers and nothing is paid on its authority. In the second stage it handles matched invoices under a set amount, with the low confidence 15 percent reviewed, and the team reports payment error rate, exception rate, cost per invoice including model charges and reviewer time. The third stage extends the agent to higher amounts only if the measured error rate is at or below the present 0.3 percent and the cost per invoice is at or below the plan in Exhibit 16. The contract with the vendor should state a price per invoice that includes model charges or caps them, allow export of the data and the workflow, and provide logs that the auditors may inspect. The CFO should also record the stop rule in advance: if the error rate on released payments exceeds 0.5 percent for two consecutive months, the agent returns to review of every item.

What a weaker answer misses. A weaker answer accepts the demonstration at face value and approves the rollout, or declines everything because the vendor was optimistic. The first mistake confuses a chosen sample with the population, and it ignores the cost of exceptions, which grows with the very errors that the agent introduces. The second mistake discards a real saving of more than half the cost of the process and ignores the information that a measured pilot can supply. A weaker answer also fails to ask who pays for the model charges, treats the monitoring staff as free, and lacks a rule for stopping. The stronger answer is not more pessimistic or more optimistic. It converts the proposal into a number the company can verify, and it designs the first steps so that the company learns cheaply whether the number is right.

5. Boardroom discussion and key takeaways

Directors do not need to understand how a model is trained. They need to know that the person who asks them to fund the investment has tested it, has counted its cost, and can say when it will be stopped. Grant Thornton reports that roughly three in four boards have approved major AI investments, yet only 52 percent have set clear expectations for AI governance, which suggests that the questions in Exhibit 18 are timely. The CFO should be ready to answer each of them from evidence and not from a vendor brochure.

Exhibit 18 Ten Questions Directors Should Ask About an AI Investment

Question a director should ask	What a ready CFO can say
What business problem does this solve, and how will we know it is solved?	The target process, the baseline cost and error rate, and the measure that will move
Is the benefit cash or capacity?	The share of saved hours converted to payroll, avoided hiring or outside spend, with named owners
What is the full cost over five years?	One time cost, run cost including model charges, data work, monitoring and exit, in a table such as Exhibit 12
What happens if the benefits are half of the plan?	The downside and break even results, and the stage gates that limit the loss
Have we tested on our own data, at full volume?	Results from a random sample and not from a vendor selected demonstration
Where does a person remain responsible?	The autonomy level for each process and the approver for each action
How dependent are we on one vendor or one model?	Contract terms for price, data export and notice of model changes, and a tested alternative
Who is accountable, and who reports results?	A named executive owner and a reporting date to the audit committee
How does this fit our policies and the NIST functions?	Mapping to Govern, Map, Measure and Manage, with detail from Module 10

Source: author's synthesis. The statistics on boards are from Grant Thornton, AI Impact Survey (April 13, 2026), self reported by nearly 1,000 senior leaders; the survey also reports that 54 percent have integrated AI risk and opportunity into ongoing board or committee oversight.

Advising the CEO and the Board on how AI changes the company's economics. The CFO should speak to three things. The first is cost structure: where AI will reduce unit cost, where it will add a new recurring cost, and how the savings will be shared with customers through price if competitors adopt the same tools. The second is competitive position: whether the company's data and customer relationships give it a lasting advantage in using AI, or whether it is buying what every rival can buy. The third is strategic choice: which investments to make first, which to defer, and what evidence would cause the company to change course. Advice delivered in these terms is more useful than a forecast of what the technology will do, because forecasts of technology are unreliable and the company's own choices are within its control.

Common errors in judging AI proposals. Experience with technology investments suggests a short list of mistakes that recur. The first is to count capacity as cash, as the sponsor did, and to treat every hour released as an hour removed from the payroll. The second is to accept the cost in the proposal as the cost of the program, when data work, internal staff and model charges are borne by the company. The third is to rely on a demonstration, which reflects the best case, in place of a random sample, which reflects the ordinary case. The fourth is to grant autonomy before accuracy has been measured, and to discover errors through a customer complaint or an audit finding. The fifth is to sign a long contract with one vendor before the market has settled, without terms for price, data export or a change of model. The sixth is to launch without an owner, so that nobody is responsible when the numbers disappoint. Each of these mistakes is avoidable, and each is cheaper to prevent in the proposal than to repair in the ledger.

Key takeaways. First, the three kinds of AI differ in what they produce, how they fail and what they cost, and the CFO should ask which kind is proposed. Second, the weaknesses of language models are reduced by design, by grounding, calculation, testing and human review, and not by hoping that the next version is better.

Third, the price in a proposal is not the cost of owning the system, and token charges, data work and monitoring belong in every business case. Fourth, a demonstration is not a production system, and the difference should be measured on a random sample at full volume. Fifth, productivity is not value until it becomes cash, and the share of hours saved that converts to cash is the variable that most often decides the answer. Sixth, stage gates and stop rules convert an uncertain commitment into a series of smaller ones, and they raise expected value without changing the technology.

6. The executive artifact: an AI investment scorecard

The scorecard in Exhibit 19 converts the evidence for a proposal into a single weighted score and a funding decision. Each of eight criteria is rated from 0 to 5, where 0 means no evidence and 5 means strong, inspected evidence, and the weighted points are the weight multiplied by the rating and divided by 5. The weights are the author's proposal and a company should adjust them to its own priorities. The bands that follow are proposed program standards and not empirical benchmarks: 75 points or more supports approval, 45 to 74 points supports a staged and conditional approval, and below 45 points means the proposal should be returned. The example completes the card for the fictional 30 million dollar proposal of Section 3, using the base case results, and the reader should complete a blank copy for a real proposal.

Exhibit 19 AI Investment Scorecard, Completed for the Fictional 30 Million Dollar Proposal

Criterion	Weig ht	Rating (0 to 5)	Points	Evidence in the example
Strategic fit and value at stake	10	4	8.0	Revenue 1,200 and a cost base that supports the thesis, but the leakage pool is estimated and not measured
Verified and cash realizable benefit	20	2	8.0	Sponsor assumes 100 percent conversion. Base case 45 percent. Break even 38.4 percent
Completeness of cost	15	2	6.0	Headline 30.0 against 38.4 one time cost and 27.4 run cost in the base case
Data and integration readiness	15	2	6.0	No data audit supplied. Industry surveys show data quality is the leading barrier
Production readiness and controls	10	2	4.0	No random sample test supplied. Only a vendor demonstration
Risk, governance and human oversight	15	3	9.0	Policy exists. Autonomy level and stop rules are not defined for the agents proposed
Vendor dependence and exit	10	2	4.0	Single vendor, no export clause, token charges not capped
Measurement and accountability plan	5	3	3.0	Benefit owners named for two of five lines. No reporting date to the audit committee
Total score	**100**		**48.0**	**Staged and conditional approval, not approval of the full 30.0**

Fictional. Computed in Python: 10 x 4 / 5 = 8.0; 20 x 2 / 5 = 8.0; 15 x 2 / 5 = 6.0; 15 x 2 / 5 = 6.0; 10 x 2 / 5 = 4.0; 15 x 3 / 5 = 9.0; 10 x 2 / 5 = 4.0; 5 x 3 / 5 = 3.0; total 48.0 of 100. The weights and bands are proposed program standards, not validated benchmarks.

Using the scorecard in the capital process. The card works best when the sponsor and an independent reviewer, such as the controller or internal audit, score it separately, and the two sets of ratings are compared before the meeting. A gap of two or more points on any criterion is a sign that the evidence is thin or that the sponsor has more information than has been shared. The card should travel with the proposal to the

investment committee and, for large amounts, to the Board, and the same card should be rescored at each stage gate so that directors can see whether the weakest ratings have improved. Over several proposals the company will learn which criteria predicted trouble, and the weights can then be adjusted on evidence.

The completed card reaches 48 points, which falls in the band for a staged and conditional approval. The sensible decision is therefore not to refuse the proposal and not to approve 30 million dollars, but to release the first stage only, 12.0 million dollars in Year 0 in the base case, on conditions that raise the weakest ratings. Those conditions are a data audit, a random sample test at full volume, a defined autonomy level and stop rule for each process, a contract that caps token charges and permits export of data, and named owners for every benefit line. When these are met the same card should be scored again before the next release of funds.

7. Research and Perspectives from the Leading Firms

The publications below were found by search, and each was opened before it was described. The descriptions state only what the page returned, and several pages gave only part of their content. A reader should confirm the current edition before relying on any figure, since this field changes quickly.

Exhibit 20 Publications of the Leading Firms and Primary Sources

Source	Title and URL	What it says
Deloitte	CFO Signals, 4Q 2025 (published January 13, 2026). https://www.deloitte.com/us/en/insights/topics/business-strategy-growth/4q-2025-cfo-signals-survey.html and press release https://www.deloitte.com/us/en/about/press-room/deloitte-q4-2025-cfo-signals-survey	Reports that 87 percent of 200 North American CFOs expect AI to be extremely or very important to finance operations in 2026 and that 50 percent named digital transformation of finance as the top priority. Fieldwork ran from November 14 to December 7, 2025
Deloitte	From Ambition to Activation: State of AI in the Enterprise (January 21, 2026). https://www.deloitte.com/us/en/about/press-room/state-of-ai-report-2026	Survey of 3,235 leaders in 24 countries. Only 25 percent had moved 40 percent or more of pilots into production. Only 30 percent were redesigning key processes around AI. Only 21 percent of those planning agentic deployments had a mature agent governance model
KPMG	The Decision Advantage: How AI is producing value across the finance function, Global AI in Finance 2026 (May 2026). https://kpmg.com/kpmg-us/content/dam/kpmg/pdf/2026/kpmg-ai-in-finance.pdf	Survey of 1,013 senior finance leaders. Active AI use in finance rose from 30 to 75 percent in two years. 71 percent said AI meets or exceeds ROI expectations and 23 percent said it exceeds them. Only 29 percent formally track where adoption fails
EY	Tax and Finance Operations Survey press release (November 18, 2025). https://ey.com/en_gl/newsroom/2025/11/ey-report-indicates-cfos-and-tax-leaders-innovating-with-ai-to-manage-mounting-geopolitical-turbulence-talent-shortages-and-regulatory-developments	Survey of 1,600 finance and tax leaders. 51 percent were in the early stages of data management maturity. 44 percent cited the inability to execute a sustainable plan for data, AI and technology as the biggest barrier
RSM US	Middle Market Firms Rapidly Embracing Generative AI, But Expertise Gaps Pose Risks, 2025 AI Survey (June 11, 2025). https://rsmus.com/newsroom/2025/middle-market-firms-rapidly-embracing-generative-ai-but-expertise-gaps-pose-risks-rsm-2025-ai-survey.html	Survey of 966 respondents. 91 percent used generative AI and one in four had fully integrated it. 92 percent of users met challenges, led by data quality (41 percent), privacy and security (39 percent) and skills (35 percent)

Source	Title and URL	What it says
Grant Thornton	A widening AI proof gap is emerging, but well governed AI is showing results (April 13, 2026). https://www.grantthornton.com/insights/press-releases/2026/april/grant-thornton-survey-on-ai-proof-gap	Self reported survey of nearly 1,000 leaders. 78 percent lack full confidence they could pass an independent AI governance audit within 90 days. Fully integrated AI users were nearly four times as likely to report AI driven revenue growth (58 versus 15 percent)
NIST	AI Risk Management Framework and Generative AI Profile (NIST-AI-600-1). https://www.nist.gov/itl/ai-risk-management-framework	Voluntary framework, version 1.0 released January 26, 2023, with a Generative AI Profile released July 26, 2024. Four functions: Govern, Map, Measure, Manage. The page notes that the framework is being revised
PwC	2026 AI Business Predictions webcast page (December 9, 2025). https://www.pwc.com/us/en/library/webcasts/2026-ai-predictions.html	Says AI is moving from potential to proof and that organizations are moving from experimentation to measurable outcomes through agentic AI, governance and enterprise transformation. The page lists no statistics and not the predictions themselves

Each page was opened and the descriptions follow what the tool returned. No BDO, Baker Tilly, Crowe, Forvis Mazars, CBIZ, Moss Adams, CohnReznick, CLA or EisnerAmper publication was read for this module, and none is cited. Survey statistics are self reported by respondents and describe perceptions and not audited outcomes.

8. The ten question self assessment

Rate each statement from 1 to 5 using the scale below, and mark one box in each row. The ratings must rest on actual experience or documented evidence, such as a proposal you evaluated, a policy you wrote or a result you reported, and not on confidence. A rating of 1 means Unprepared: limited understanding and no meaningful application. A rating of 2 means Developing: conceptual familiarity, relying heavily on others. A rating of 3 means Competent: able to apply the capability independently in conventional situations. A rating of 4 means Advanced: the capability is integrated into strategy and delivers measurable results. A rating of 5 means Future-ready: enterprise wide impact, anticipation of change and the development of others. The five point scale and the ten statements are those of the source proposal for this program, lightly polished.

No.	Statement of demonstrated capability	1	2	3	4	5
1	I can explain the differences among traditional machine learning, generative AI and agentic AI.	☐	☐	☐	☐	☐
2	I understand the capabilities and limitations of large language models, including hallucinations and reasoning failures.	☐	☐	☐	☐	☐
3	I can identify the highest value AI use cases across accounting, FP&A, treasury and business operations.	☐	☐	☐	☐	☐
4	I can evaluate AI initiatives using financial measures such as ROI, payback, productivity gains and risk adjusted value.	☐	☐	☐	☐	☐
5	I understand AI agents, their tool access, workflow orchestration and human in the loop controls.	☐	☐	☐	☐	☐

No.	Statement of demonstrated capability	1	2	3	4	5
6	I can distinguish between AI demonstrations and production ready enterprise AI systems.	☐	☐	☐	☐	☐
7	I understand the importance of data quality, system integration, access controls and enterprise architecture for AI adoption.	☐	☐	☐	☐	☐
8	I can assess AI vendor proposals, implementation costs, recurring expenses and vendor dependency risks.	☐	☐	☐	☐	☐
9	I can establish measurable AI adoption goals and determine whether expected financial benefits are being realized.	☐	☐	☐	☐	☐
10	I can advise the CEO and the Board on how AI could change the company's economics, competitive position and strategic choices.	☐	☐	☐	☐	☐

Total score: _____ out of 50. Ten statements rated from 1 to 5 give a minimum of 10 and a maximum of 50 points. Exhibit 21 gives the proposed interpretation.

Exhibit 21 Proposed Interpretation of the Module 1 Score

Score	Reading	Suggested use
10 to 19	Unprepared	Study the module in full, and shadow a colleague who evaluates AI proposals
20 to 29	Developing	Complete the worked model and the scorecard on a real proposal, with review by a more experienced peer
30 to 39	Competent	Lead an evaluation and a post investment review, then present results to the Board
40 to 45	Advanced	Embed the scorecard and stage gates in the capital process and teach others
46 to 50	Future-ready	Share the method across the enterprise and test it against outside review

These bands are proposed program standards and have not been empirically validated against a broad population of finance executives. The scale is from the source proposal; the band boundaries are the author's.

Critical module note

Module 1 is one of the five critical modules in the program, together with capital allocation (Module 5), governance (Module 9), risk (Module 10) and enterprise leadership (Module 13). A score below 30 out of 50 on a critical module would block the highest designation, even when the total across all twenty modules is high. The reason is that excellence in one area should not conceal a consequential weakness in another.

Development actions by pattern. A single total can hide the pattern behind it, so look at which statements received the lowest ratings. Exhibit 22 suggests three actions for each pattern of low scores.

Exhibit 22 Development Actions for Low Score Patterns

Low scores on	Pattern	Three suggested development actions
Statements 1, 2 and 5	Conceptual gap: the reader cannot yet explain what the systems are and how they fail	1. Test a language model on ten of your own documents and record every error. 2. Draw the autonomy ladder for one process in your company. 3. Ask your technology lead to walk you through one agent's tools and logs

Low scores on	Pattern	Three suggested development actions
Statements 4, 8 and 9	Economic gap: the reader cannot yet price and judge a proposal	1. Rebuild the model in Exhibit 12 in a spreadsheet and change each assumption. 2. Review an actual vendor quote for hidden and recurring costs. 3. Set up a benefits tracking report for one live project
Statements 6 and 7	Production gap: the reader cannot yet tell a demonstration from a system that works	1. Require a random sample test before the next pilot is extended. 2. Complete the data checklist for one process. 3. Meet the owner of the ledger interfaces and list what is not integrated
Statements 3 and 10	Strategy and advisory gap: the reader cannot yet turn knowledge into advice	1. Rank five use cases by volume, risk and readiness. 2. Draft a one page memorandum to the Board on how AI may alter unit cost. 3. Rehearse the answers to the questions in Exhibit 18 with a peer

Source: author's suggestions. The mapping of statements to patterns is the author's proposal, and ratings should rest on experience or documented evidence.

A closing thought. Every generation of finance leaders has faced a technology that arrived with large promises, from the spreadsheet to the enterprise system, and in each case the companies that gained most were not the ones that adopted first. They were the ones that measured honestly, fixed their data, kept a person responsible for each important decision, and stopped what did not work. Artificial intelligence is likely to follow the same pattern, though it may move faster and fail in less familiar ways. The CFO who learns the vocabulary, counts the full cost, insists on evidence from the company's own records and builds the stage gates before the money is spent will not need to predict the technology. That CFO will need only to remain curious, candid and careful, which are the habits that have always served the office well. The assessment in Section 8 is a way to find out, with evidence and without flattery, how far along that road the reader has traveled.

9. Currency check and sources

Currency check: this field changes quickly, and prices, products and survey results shown here may already have changed. The survey statistics were read on the pages listed in Exhibit 20 in October 2026 and describe the perceptions of respondents, largely self reported, and not audited outcomes. The following were confirmed on pages read: the Deloitte CFO Signals results of 87 percent and 50 percent, the sample of 200 CFOs at companies with at least 1 billion dollars of revenue, the fieldwork dates and the publication date; the Deloitte State of AI figures; the KPMG figures; the EY figures; the RSM figures; the Grant Thornton figures; and the NIST dates, the names of the four functions and the note on revision. The following come from the author's own knowledge and were not confirmed in a source read: the descriptions of hallucination, reasoning failure, prompt injection and the cost structure of language models; the definitions of tokens, context windows and inference; the five level autonomy ladder and the build versus buy comparison; the claim that a token is roughly three quarters of a word; and the use case ratings. The token prices, the 30 million dollar proposal, the companies, the probabilities, the scorecard weights and all dollar figures are fictional and chosen for teaching. All sums, net present values, internal rates of return, paybacks, sensitivities and shares were computed in Python and checked by assertion. This module is educational and is not accounting, tax, legal, privacy, securities or investment advice. Readers must consult qualified professionals before acting, and data rights, privacy, contract and regulatory questions in particular require legal advice.

Sources: Deloitte, CFO Signals 4Q 2025 and press release (January 13, 2026); Deloitte, State of AI in the Enterprise (January 21, 2026); KPMG, Global AI in Finance 2026, The Decision Advantage (May 2026); EY, Tax and Finance Operations Survey press release (November 18, 2025); RSM US, 2025 AI Survey press release (June 11, 2025); Grant Thornton, AI proof gap survey release (April 13, 2026); NIST, AI Risk Management Framework page; PwC, 2026 AI Business Predictions webcast page. URLs appear in Exhibit 20. The ten assessment statements and the five point scale are from the source proposal for this program.