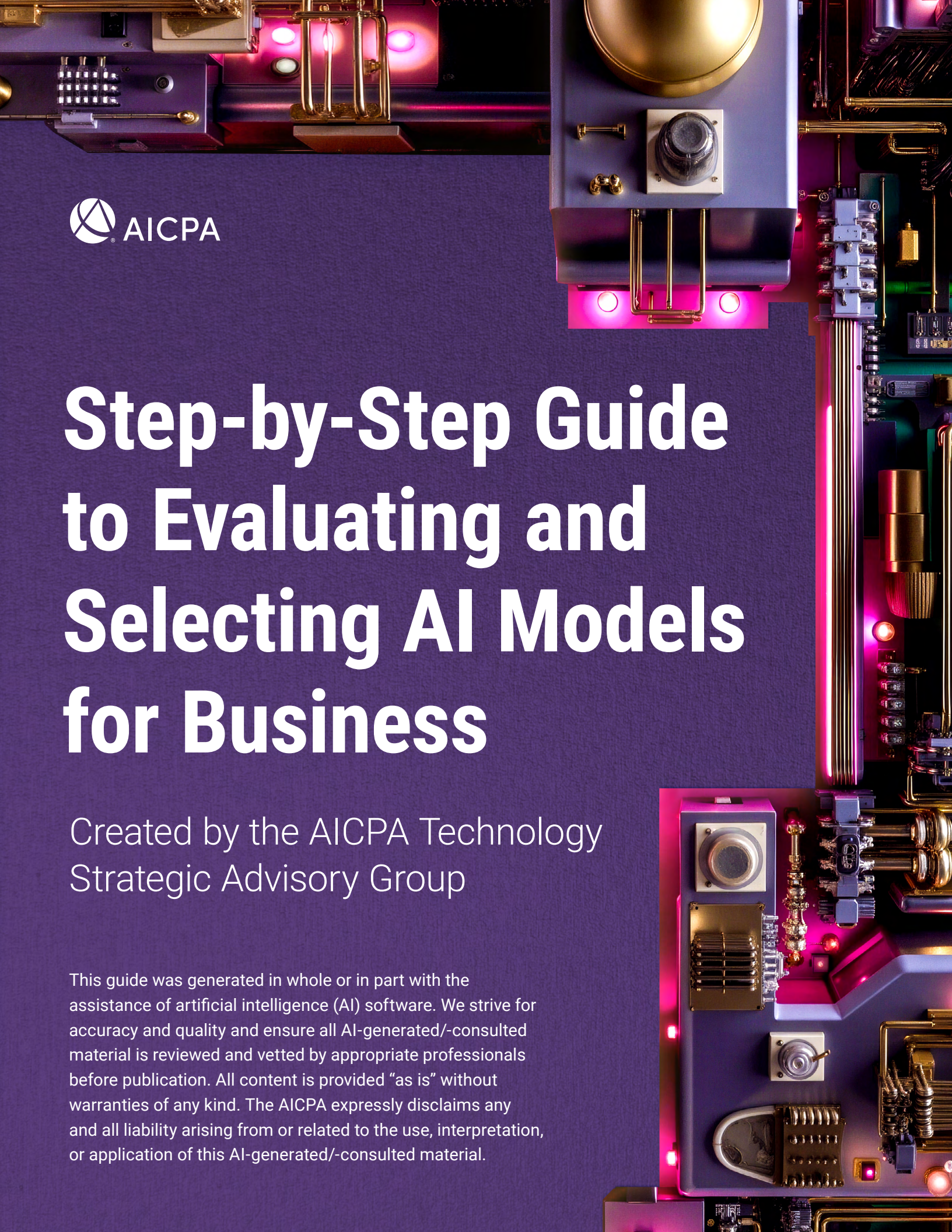


AICPA

Step-by-Step Guide to Evaluating and Selecting AI Models for Business

Created by the AICPA Technology Strategic Advisory Group

This guide was generated in whole or in part with the assistance of artificial intelligence (AI) software. We strive for accuracy and quality and ensure all AI-generated/-consulted material is reviewed and vetted by appropriate professionals before publication. All content is provided “as is” without warranties of any kind. The AICPA expressly disclaims any and all liability arising from or related to the use, interpretation, or application of this AI-generated/-consulted material.

Scope of This Guide

This guide covers the evaluation and selection of AI solutions broadly, including the following:

- Large language models (LLMs) and generative AI (gen AI) for text generation, analysis, and conversation
- Traditional machine learning (ML) models for predictive analytics, classification, and forecasting
- Specialized AI systems, including computer vision, natural language processing, and recommendation engines
- Follows a structured, phase-based evaluation framework, with a visual flowchart included at the end of the guide for reference

Why Use This Guide?

The Challenge

Organizations today face an overwhelming array of AI solutions, each promising transformative results. However, choosing the wrong AI model can lead to significant setbacks. Research from the RAND Corporation estimates that more than 80% of AI projects fail — twice the failure rate for traditional IT projects. According to Gartner, on average only 48% of AI projects make it from pilot to production, taking approximately eight months to move from prototype to deployment. S&P Global's 2025 analysis found that 42% of companies abandoned most of their AI initiatives, up from 17% in 2024.[1]

The consequences of failed AI projects include wasted investment in both direct costs and lost opportunity, months of development time with no usable outcome, technical debt from poorly integrated systems requiring costly rework, and missed competitive opportunities.

[1] James Ryseff, Brandon De Bruhl, and Sydne J. Newberry, *The Root Causes of Failure for Artificial Intelligence Projects and How They Can Succeed: Avoiding the Anti-Patterns of AI*, Research Report RR-A2680-1 (RAND Corporation, 2024). rand.org/content/dam/rand/pubs/research_reports/RRA2600/RRA2680-1/RAND_RRA2680-1.pdf; Gartner, Inc., "Gartner Survey Finds Generative AI Is Now the Most Frequently Deployed AI Solution in Organizations," press release, May 7, 2024, gartner.com/en/newsroom/press-releases/2024-05-07-gartner-survey-finds-generative-ai-is-now-the-most-frequently-deployed-ai-solution-in-organizations; Lindsey Wilkinson, "AI Project Failure Rates Are on the Rise: Report," CIO Dive, March 14, 2025, ciodive.com/news/AI-project-fail-data-SPGlobal/742590/.

Who Should Use This Guide

This guide is designed for

- **business leaders** evaluating AI investments without deep technical expertise,
- **IT directors** investigating which enterprise AI solution to select,
- **project managers** leading digital transformation initiatives,
- **consultants** helping clients navigate AI adoption, and
- **procurement teams** assessing AI vendor proposals.

What Makes This Guide Different

Unlike vendor-specific materials or purely technical frameworks, this guide

- **balances business and technical considerations** equally,
- **provides concrete evaluation criteria** with scoring templates,
- **includes real-world timelines** based on actual implementations,
- **offers vendor-neutral comparisons** across major providers, and
- **addresses common pitfalls** from numerous AI deployments.

Expected Outcomes

By following this systematic approach, you will

- **significantly reduce selection risk** through structured evaluation;
- **accelerate decision-making** from months to weeks;
- **align stakeholders** with clear, objective criteria;
- **avoid common mistakes** that derail many AI projects; and
- **build confidence** in your AI investment decisions.

When to Use This Guide

Apply this framework when

- your organization is considering its first AI implementation,
- you're expanding AI use to new business areas,
- your current AI solutions aren't delivering expected value,
- you need to justify AI investments to leadership, and
- you're comparing multiple vendors or approaches.

Investment Context

Gartner forecast that worldwide AI spending would total nearly $1.5 trillion in 2025, and expected spending to top $2 trillion in 2026. International Data Corporation projects that AI solutions and services will generate a cumulative global economic impact of $22.3 trillion by 2030, representing approximately 3.7% of global GDP. Organizations that successfully implement AI are reporting measurable results: according to a 2023 Gartner survey, respondents reported an average 15.8% revenue increase, 15.2% cost savings, and 22.6% productivity improvement from gen AI initiatives. However, realizing these benefits requires careful selection and planning — the focus of this guide.[2]

Phase 1: Define Business Requirements

Why This Phase Matters

Starting with technology rather than business needs is the most common reason AI projects fail. The RAND Corporation's 2024 research identified that industry stakeholders "often misunderstand — or miscommunicate — what problem needs to be solved using AI" as a leading cause of failure. McKinsey's March 2025 AI survey found that organizations that redesign workflows have a larger overall impact on the "ability to see [earnings before interest and taxes] impact from its use of gen AI." This phase ensures you're solving real problems, not just implementing trendy technology. Organizations that clearly define requirements upfront are far more likely to achieve positive return on investment (ROI) within the first year.[3]

[2] Gartner, Inc., "Gartner Says Worldwide AI Spending Will Total $1.5 Trillion in 2025," press release, September 17, 2025, gartner.com/en/newsroom/press-releases/2025-09-17-gartner-says-worldwide-ai-spending-will-total-1-point-5-trillion-in-2025; International Data Corporation (IDC), "IDC Predicts AI Solutions & Services will Generate Global Impact of $22.3 Trillion by 2030," press release, April 1, 2025, https://my.idc.com/getdoc.jsp?containerId=prUS53290725; Gartner, Inc., "Gartner Predicts 30% of Generative AI Projects Will Be Abandoned After Proof of Concept By End of 2025," press release, July 29, 2024, gartner.com/en/newsroom/press-releases/2024-07-29-gartner-predicts-30-percent-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025.

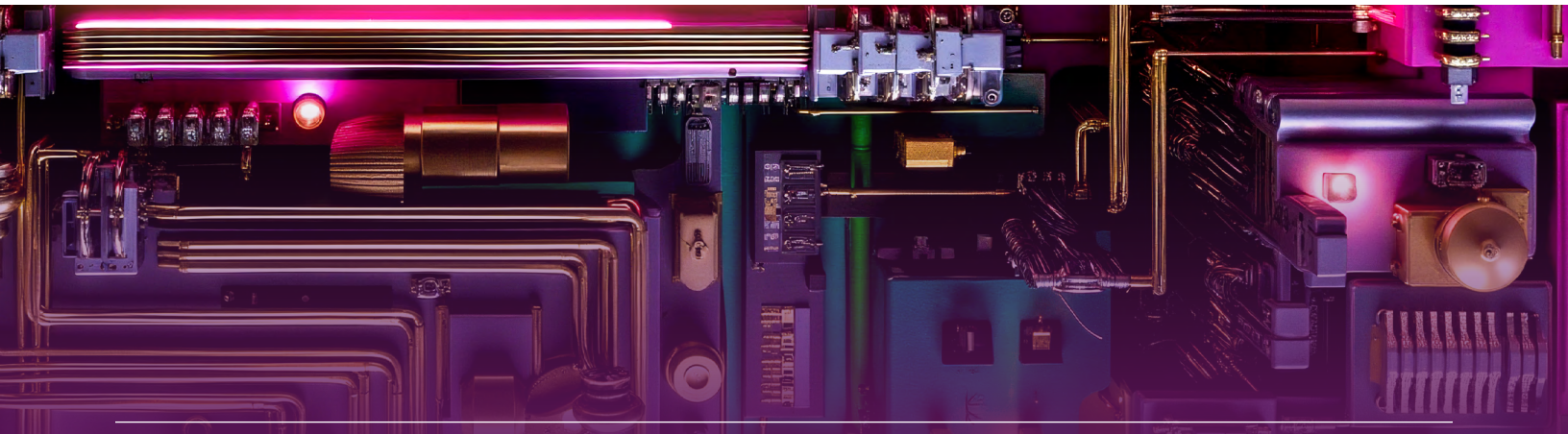
[3] James Ryseff, Brandon De Bruhl, and Sydne J. Newberry, *The Root Causes of Failure for Artificial Intelligence Projects and How They Can Succeed: Avoiding the Anti-Patterns of AI*, Research Report RR-A2680-1 (RAND Corporation, 2024). rand.org/content/dam/rand/pubs/research_reports/RRA2600/RRA2680-1/RAND_RRA2680-1.pdf; Alex Singla, Alexander Sukharevsky, Lareina Yee, and Michael Chui, with Bryce Hall, *The State of AI: How Organizations Are Rewiring to Capture Value* (McKinsey, March 12, 2025), mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-how-organizations-are-rewiring-to-capture-value.

What You'll Accomplish

By completing this phase, you'll have crystal-clear alignment among stakeholders on what success looks like, preventing costly scope creep and ensuring every technical decision ties back to business value.

Requirement Evaluation Action Plan

Step	Action Item	Key Questions	Deliverable
1	**Identify use case**	What specific business problem are you solving? What tasks need automation?	Clear statement of the problem
2	**Define success metrics**	How will you measure success? What are the key performance indicators?	Success criteria document
3	**Assess current state**	What existing systems/processes are in place? What data is available?	Current-state analysis
4	**Set budget and timeline**	What's your budget range? When do you need results?	Budget and timeline constraints



Phase 2: Technical Requirements Analysis

Why This Phase Matters

Many organizations discover too late that their infrastructure can't support their chosen AI solution, or that they lack the data quality needed for accurate results. This phase prevents expensive surprises by mapping technical realities before vendor selection.

What You'll Accomplish

You'll identify potential technical roadblocks early, understand true implementation costs (including hidden infrastructure needs), and ensure your organization has the technical foundation for AI success.

Critical: Data Readiness Assessment

Data quality is the single most critical factor in the success of AI projects. According to Informatica's CDO Insights 2025 survey, the top obstacles to AI success are data quality and readiness (43%), lack of technical maturity (43%), and shortage of skills and data literacy (35%). Gartner predicts that organizations will, through 2026, abandon 60% of AI projects unsupported by AI-ready data. A 2024 Gartner survey found that 63% of organizations either do not have or are unsure if they have data management practices appropriate for AI. Before proceeding, conduct a thorough data quality audit covering completeness, accuracy, consistency, timeliness, and accessibility of your data assets.[4]

Assessment Action and Criteria

Step	Action Item	Key Questions	Evaluation Criteria
1	**Data requirements**	Volume, quality, format, privacy requirements	Data availability score
2	**Performance needs**	Accuracy, speed, throughput, latency requirements	Performance benchmarks
3	**Integration requirements**	Application programming interfaces (APIs), existing systems, deployment environment	Technical compatibility
4	**Scalability needs**	Expected growth, peak usage, geographic distribution	Scalability requirements

[4] Informatica, *CDO Insights 2025 Survey Report*, 2025, informatica.com/campaigns/cdo-insights-2025/index.html; Gartner, Inc., "Lack of AI-Ready Data Puts AI Projects at Risk," press release, February 26, 2025, gartner.com/en/newsroom/press-releases/2025-02-26-lack-of-ai-ready-data-puts-ai-projects-at-risk.

Phase 3: Model Type Selection

Why This Phase Matters

Choosing the wrong type of AI model is like using a hammer when you need a screwdriver. Each AI type excels at specific tasks. Understanding these distinctions prevents costly mistakes — like investing heavily in an LLM when a simpler predictive model would work better.

What You'll Accomplish

You'll narrow the field from hundreds of options to the specific type that matches your needs, saving weeks of evaluation time and ensuring you're comparing apples to apples, not apples to oranges.

Model Types and Considerations

Model Type	Best for	Pros	Cons	Example Tools/ Providers	Example Use Cases
LLMs	Text generation, analysis, conversation	Versatile, powerful, pretrained	High cost, hallucinations	GPT-5, GPT-4 (OpenAI); Claude (Anthropic); Gemini (Google); Perplexity; Cohere; Chipmunk Robotics (Inkwise.ai)	Customer service, content creation, professional technical writing
Computer vision	Image/video analysis	Accurate visual recognition	Requires large datasets	Google Vision API, Amazon Web Services (AWS) Rekognition, Azure Computer Vision, Clarifai, Roboflow	Quality control, security
Predictive analytics	Forecasting, classification	Proven ROI, interpretable	Needs historical data	AWS SageMaker, Google Vertex AI, DataRobot, H2O.ai, Azure ML Studio	Sales forecasting, risk assessment
Recommendation systems	Personalization	Drives engagement	Cold-start issues (performance is limited when new users or items have little to no data for the model to learn from)	AWS Personalize, Google Recommendations AI, Azure Personalizer, Recombee, Dynamic Yield	E-commerce, content platforms
Natural language processing	Text analysis and sentiment detection (e.g., customer or user feedback)	Domain-specific accuracy	Language limitations	AWS Comprehend, Google Natural Language API, Azure Text Analytics, spaCy, Hugging Face	Document processing, social monitoring

Note: Many platforms offer automated machine learning (AutoML) capabilities that simplify model selection and training. AutoML features are available across providers, including Google Vertex AI, AWS SageMaker, Autopilot, Azure AutoML, DataRobot, and H2O.ai. These tools can automatically select algorithms, tune hyperparameters, and optimize models based on your data.

Phase 4: Model Evaluation Framework

Why This Phase Matters

Without objective evaluation criteria, selecting a vendor becomes a political exercise rather than a data-driven decision. This framework transforms subjective preferences into quantifiable scores, defending your choice to stakeholders and ensuring the best technical fit.

What You'll Accomplish

You'll create an audit trail for your decision-making process, enable fair comparison across diverse solutions, and build consensus among technical and business stakeholders using objective data.

Technical Evaluation Matrix

Criteria	Weight	Scoring Method	Benchmark Tests
Accuracy	25%	Validation against test dataset	F1-score, precision, recall
Performance	20%	Speed and throughput testing	Latency, requests per second
Reliability	20%	Stress testing, error handling	Uptime, error rates
Scalability	15%	Load testing	Concurrent users, data volume
Integration	10%	API testing, compatibility	Ease of implementation
Cost efficiency	10%	Total cost of ownership	Cost per transaction/query

Business Evaluation Criteria

Factor	Assessment Method	Risk Level
Vendor stability	Financial health, market position	High/medium/low
Support quality	SLA, documentation, community	Rate 1–5
Compliance[5]	GDPR, HIPAA, SOC 2® requirements	Compliant/noncompliant
Security	Data encryption, access controls	Security audit results

Note: GDPR stands for General Data Protection Regulation; HIPAA stands for Health Insurance Portability and Accountability Act.

[5] AI governance and responsible AI: Beyond regulatory compliance (e.g., GDPR, HIPAA, SOC 2), organizations should evaluate AI -specific governance during selection. This includes human-in-the-loop oversight protocols, bias detection and fairness testing, etc. Emerging regulatory frameworks (including the EU Artificial Intelligence Act) are establishing new expectations for AI transparency and accountability that may affect procurement and attestation requirements.

Phase 5: Vendor and Model Comparison[6]

Why This Phase Matters

The AI vendor landscape changes continuously, with new players emerging and established ones evolving their offerings. This phase helps you navigate vendor marketing hype to understand real capabilities, pricing structures, and long-term viability.

What You'll Accomplish

You'll cut through marketing noise to understand true costs (including hidden fees), identify which vendors align with your technical architecture, and assess vendor stability for long-term partnership potential.

Leading AI Model Providers

Provider	Strengths	Best Use Cases	Pricing Model
OpenAI	Advanced LLMs, broad capabilities	Conversational AI, content generation	Pay-per-token
Anthropic	Safety-focused, reliable	Enterprise applications, analysis	Pay-per-token
Google Cloud AI	Integrated ecosystem	Predictive analytics, computer vision	Various pricing tiers
AWS AI/ML	Comprehensive services, scalability	Enterprise-scale deployments	Pay-as-you-go
Microsoft Azure AI	Enterprise integration, hybrid cloud, graphical user interface	Business process automation, prebuilt templates	Subscription + usage

Model Comparison Template

Model Name	Accuracy Score	Latency (ms)	Cost ($/1K queries)	Support Rating	Overall Score
Model A	94%	150	$2.50	4/5	8.5/10
Model B	91%	80	$1.75	3/5	7.8/10
Model C	96%	200	$3.20	5/5	8.9/10

Note: The preceding template is a sample model comparison. Entries and values will depend on the user's chosen models.

[6] Model providers vs. cloud platforms: AI model providers and cloud infrastructure platforms represent distinct layers of the technology stack, though they are often conflated. Model providers (e.g.,OpenAI, Anthropic) develop and maintain AI models; cloud platforms (e.g., AWS, Google Cloud, Azure) provide the infrastructure to deploy and scale. In practice, many AI models are available through multiple platforms. Therefore, organizations may separately decide which model fits their use case and which platform fits their infrastructure.

Phase 6: Pilot Testing Strategy

Why This Phase Matters

Lab performance rarely matches real-world results. Pilots reveal hidden issues like edge cases, user resistance, and integration challenges that could derail full deployment. Gartner predicted that at least 30% of gen AI projects will have been abandoned after proof of concept by the end of 2025 due to poor data quality, inadequate risk controls, escalating costs, or unclear business value. MIT's 2025 research found that only about 5% of AI pilot programs achieve rapid revenue acceleration, with the vast majority stalling. Thorough pilot testing is essential to validate assumptions before full investment.[7]

What You'll Accomplish

You'll validate vendor claims with actual data, identify and resolve integration issues before they become critical, build user confidence through gradual rollout, and create realistic implementation timelines based on actual experience

Pilot Evaluation Process

Step	Duration	Activities	Success Metrics
Proof of concept	2–4 weeks	Basic functionality testing	Technical feasibility confirmed
Limited pilot	4–8 weeks	Small-scale deployment	Performance targets met
Expanded testing	8–12 weeks	Larger user group, edge cases	Business value demonstrated
Final validation	2–4 weeks	Full integration testing	Ready for production

[7] Gartner, Inc., "Gartner Predicts 30% of Generative AI Projects Will Be Abandoned After Proof of Concept By End of 2025," press release, July 29, 2024, gartner.com/en/newsroom/press-releases/2024-07-29-gartner-predicts-30-percent-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025; Aditya Challapally, Chris Pease, Ramesh Raskar, Pradyumna Chari, *The GenAI Divide: State of AI in Business* 2025 (MIT NANDA, July 2025).

Phase 7: Decision Framework

Why This Phase Matters

After weeks of evaluation, decision paralysis can set in. This framework transforms complex multifactor decisions into clear recommendations, helping you defend your choice to executives, board members, or other stakeholders who weren't part of the evaluation process.

What You'll Accomplish

You'll produce a defensible, data-driven recommendation, align all stakeholders on the final decision, and create documentation for future AI selections, thus establishing a repeatable process for your organization.

Scoring Matrix (100-Point Scale)

Category	Weight	Scoring Criteria
Technical performance	40 points	Accuracy, speed, reliability scores
Business value	25 points	ROI potential, strategic alignment
Implementation feasibility	20 points	Integration complexity, timeline
Risk assessment	15 points	Vendor stability, security, compliance

Decision Tree

Does the model meet minimum accuracy requirements? (>85%)

- **No** → Eliminate from consideration
- **Yes** → Continue evaluation

Is the total cost within budget?

- **No** → Negotiate or consider alternatives
- **Yes** → Proceed to pilot testing

Does it integrate with existing systems?

- **No** → Assess integration effort and costs
- **Yes** → Strong candidate for selection

Is the risk assessment acceptable?

- **No** → Develop mitigation strategies
- **Yes** → Recommend for implementation

Phase 8: Implementation Planning

Why This Phase Matters

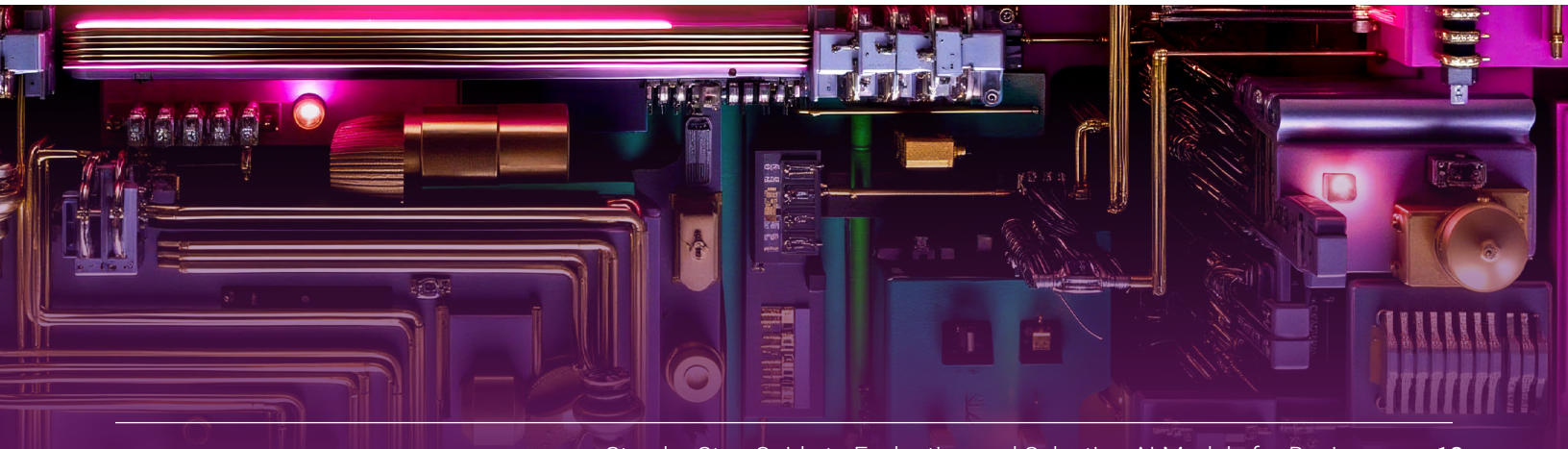
The gap between selection and deployment is where most AI projects fail. Poor implementation planning leads to significant budget overruns, substantial delays, and user rejection. This phase bridges that gap with precision.

What You'll Accomplish

You'll create a realistic roadmap that accounts for technical and human factors, establishes clear accountability and success metrics, makes contingency plans for common problems, and ensures smooth transition from pilot to production scale.

Implementation Process

Milestone	Timeline	Key Activities	Risk Mitigation
Model selection	Week 1	Final decision, contract negotiation	Have backup options ready
Integration setup	Weeks 2–4	API integration, data pipeline setup	Parallel development tracks
User training	Weeks 4–6	Staff training, documentation	Change-management program
Go live	Weeks 7–8	Production deployment, monitoring	Rollback plan prepared
Optimization	Ongoing	Performance tuning, user feedback	Continuous improvement process



Key Success Factors

Critical Do's:

- Start with a clear business case and success metrics.
- Conduct thorough pilot testing before full deployment.
- Plan for ongoing model maintenance and updates.
- Invest in user training and a change-management program.
- Establish monitoring and governance frameworks.

Critical Don'ts:

- Don't select a model based on hype rather than business needs.
- Don't underestimate integration complexity and costs.
- Don't ignore data quality and preparation requirements.
- Don't fail to plan for model drift and performance degradation.
- Don't fail to consider long-term vendor relationship and support.

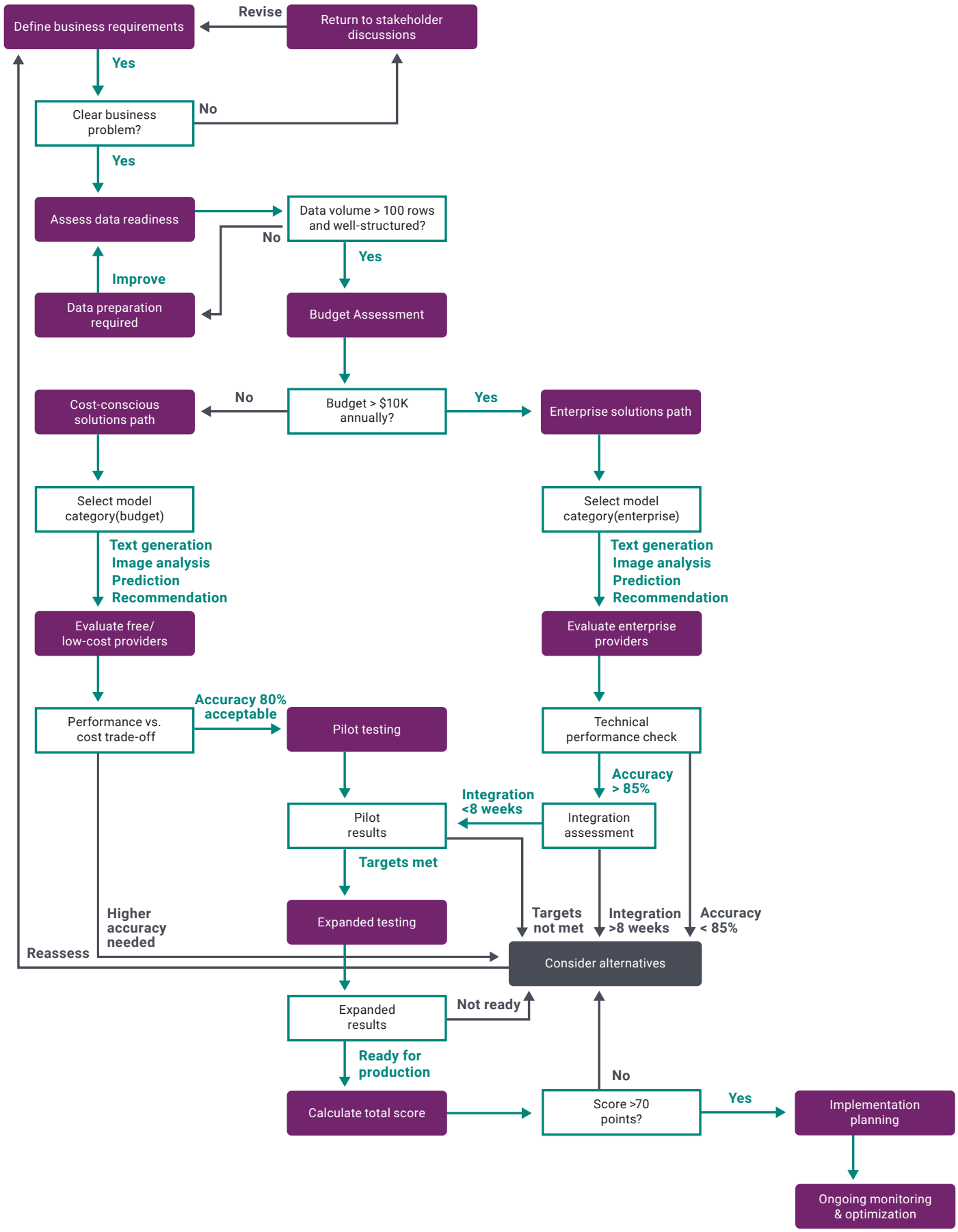
Recommended Timeline

Total Process Duration: 12–16 weeks

- Requirements definition: 2–3 weeks
- Technical analysis: 2–3 weeks
- Vendor evaluation: 3–4 weeks
- Pilot testing: 4–6 weeks
- Final selection and planning: 1 week

The end-to-end evaluation process outlined in this guide is summarized in the flowchart on the following page.

Figure: End-to-End AI Model Evaluation and Selection Framework



Founded by AICPA and CIMA, the Association of International Certified Professional Accountants powers leaders in accounting and finance around the globe.