accenture

EnterpriseVAL: Accenture's frontier AI enterprise benchmark

Improving model performance with domain expertise, fit-for-purpose data and benchmarks built around real business workflows.

1 Introduction Page 1-2

2 EnterpriseVAL Page 3-4

3 Benchmarking results Page 5-9

4 Pathway to success Page 10-12

5 Conclusion Page 13-14

I.

Introduction

Enterprises are investing billions of dollars in AI, yet many still struggle to connect model performance to measurable business outcomes. The challenge is not a lack of AI capability, but a lack of evidence that AI can reliably perform the complex, judgement-intensive work that drives value in critical business functions. AI developers are working to make models more capable, efficient and reliable in real business environments. Yet, as our research report, The CIO's guide to AI tokenomics shows, four in five dollars of AI token spend lack a quantified link to business outcomes, "whether that is revenue influenced, cost avoided or productivity converted to a dollar figure."[1]

Across the enterprise, AI is reshaping how work is distributed between people and technology, creating new opportunities to augment human expertise and automate selected activities.

The economic surface area is significant. Finance and accounting provides one illustration of this scale: APQC estimates the median cost of the finance function at 0.79% of revenue across industries.[2] Applied illustratively to the $43.1 trillion in revenue generated by the Fortune Global 500,[3] this represents an annual finance-function cost base of approximately $340.5 billion.

Finance is only one part of the enterprise. Similar opportunities exist across functions including procurement, cybersecurity, customer service and HR. The greater potential lies in improving how work is performed using AI to increase capacity, speed and quality while enabling people to focus on innovation, relationships and the expert judgement where human capabilities create the greatest value.

AI investment continues to grow, but sustained business value remains harder to achieve. Accenture's latest Pulse of Change research found that 75% of C-suite leaders plan to increase AI investment over the next 12 months, yet only 25% report widespread and sustained value from AI.[4]

This suggests an AI readiness gap has emerged, characterized by uncertainty over the performance of models and their reliability to complete enterprise tasks to the same standards as an industry or function practitioner.

The question is: how can we bridge that gap and give confidence to end customers and key AI partners?

The next challenge for enterprise AI is not necessarily access to better models; it is making the switch from offering strong capabilities to delivering actual business value. To do so, enterprises must learn which models, or which release of a model, performs to their desired levels at the right price point. That requires technology, operating decisions and investment priorities to converge around the outcomes that matter most. Leaders need a clear view of where AI can perform, where it creates economic advantage and where intervention is still required. Without that discipline, AI investment risks scaling faster than the value it delivers.

Enterprise domain knowledge is becoming one of the most valuable inputs in AI development. Continuously capturing and benchmarking this knowledge against real enterprise work creates a bridge between AI capability and business outcomes.

II.

EnterpriseVAL

Traditional benchmarks measure general intelligence and technical performance, but they rarely reflect the realities of enterprise workflows where accuracy, context, compliance and professional judgement matter most.

EnterpriseVAL was created to address this gap. It evaluates AI models against real-world business tasks, expert-defined standards and function-specific success criteria. This provides a practical way to determine whether AI is truly ready to deliver business value, where it can be trusted to operate and where human oversight remains essential.

EnterpriseVAL is Accenture's new enterprise AI benchmark series, built to bring greater clarity to where AI is ready for enterprise deployment.

It addresses a critical evidence gap: enterprises need to know whether AI models can perform the complex, judgement-intensive work embedded in critical enterprise functions. By evaluating realistic workflows and practitioner-ready deliverables, EnterpriseVAL provides the evidence to show where models can deliver value reliably, safely and at an acceptable cost.

Our new proprietary approach to enterprise benchmarking creates a decision-relevant evidence base for model development and adoption. It captures the most demanding tasks from real enterprise workflows, alongside the source data, expert judgement and rigorous evaluation criteria needed to assess them.

We built 81 finance tasks across record-to-report, order-to-cash, procure-to-pay, financial planning & analysis and treasury. The tasks were written by expert practitioners averaging 18 years' experience. Each comes with real source files, an expert answer and a rubric that marks the errors that matter most. Every task passed an independent five-phase review and the benchmark is sensitive enough to pick up improvement between two releases of the same model.

The gap

Frontier models look ready on paper. Scored on the rubric alone, the models we tested sit close together and score well. Count only the errors that would stop a reviewer signing off, and every score drops sharply. Even the best model fell short of sign-off on a large share of the tasks. Models need data built by the people who do the work, tested against the standard those people apply.

Finance and Accounting (F&A) is the first domain in the EnterpriseVAL series, which will extend across enterprise functions and industries where frontier model performance is consequential and currently unmeasured. F&A provides a strong test bed for this work. Practitioner workflows combine multistep processes that are both deterministic and dependent on professional judgement. Practitioners work across information retrieval, calculation, modelling, validation, visualization and reporting using real-world enterprise artifacts. This allows EnterpriseVAL to test the agentic capabilities of AI models across multiple file formats, interconnected tasks and points requiring professional judgement.

EnterpriseVAL shows where a model is ready, where it falls short and why. Our process reveals:

- **What every model can do:** Tasks all models pass, marking work that is ready for wider automation.
- **What no model can do yet:** Tasks all models fail, showing where models, prompts or workflows still need to improve.
- **Where the model matters:** Tasks where the choice of model changes the outcome.
- **Strengths and weaknesses by function:** How performance varies across the finance and accounting sub-domains.
- **Strengths and weaknesses by type of work:** How models handle the kinds of work that run across every sub-domain, such as calculation, judgement and presentation.

III.

Benchmarking results

Benchmarking approach

Experienced finance and accounting practitioners identified representative, judgement-intensive activities from their own workflows and developed each as a complete task package. They provided a realistic scenario, supporting business files, an expert reference deliverable and a task-specific scoring rubric. The tasks spanned six finance and accounting domains and were designed to require the interpretation, reasoning and decision-making expected in real enterprise workflows.

Accenture and Faculty then tested nine frontier AI models in a controlled environment and assessed their outputs across three dimensions: overall quality, detailed rubric achievement and whether a critical error would prevent professional sign-off. Quality assurance was embedded throughout the benchmark lifecycle. Every task was reviewed independently by a separate finance expert through a five-phase process covering completeness, realism, calculations, assumptions, expected outcomes and rubric clarity. This was followed by peer challenge, remediation and final approval. Tasks underwent an average of six refinement cycles, supported by automated language and accuracy checks and no task entered the final dataset with unresolved QA.

The evaluation method was also validated on a stratified 25-task sample using blinded expert preference, independent expert rubric grading and an automated judge on the same outputs. This allowed the team to measure evaluator agreement, identify bias and determine where automation was sufficiently reliable for scaling the full 81-task study.

We tested the models in two benchmarking studies: a technical benchmark and a human benchmark.

Technical benchmark

The technical benchmark ran nine frontier models, eight proprietary and one open-weight, across all 81 tasks in our finance and accounting dataset.

Every model received the same expert-curated task instruction and the same supporting files and every output was scored against the same rubric that the task expert wrote for that task. The tasks are drawn from six sub-domains within finance and accounting: record-to-report, order-to-cash, procure-to-pay, financial planning & analysis, risk & compliance and treasury.

Headline scores make models look interchangeable, which doesn't tell the full picture; **critical-failure gating is the real differentiator. Tasks that had a critical failure are unusable in the enterprise.**

Performance varies more by workflow dimension than by Finance and Accounting domain. The 3 weakest criterion types are all quantitative: calculation, numerical accuracy and quantitative accounting treatment while models score stronger on qualitative tasks including: tone, source fidelity and qualitative accounting treatment. Some examples of criteria that no model passed included:

- **Quantitative calculation** of formulating all totals and percentages rather than hardcoding them.
- **Numerical accuracy** when reporting figures such as year-to-date current income tax accrued.
- **Qualitative accounting treatment** when deciding what corporate policy prevails for consolidation and justifying it.

Model performance was the highest in the treasury domain where there is already a great deal of automation in the work process. It was lowest in financial planning and analysis (FP&A) where a higher degree of numerical accuracy, quantitative accounting and calculation is required resulting in the greatest critical failures. **Tasks that had a critical failure are unusable in the enterprise.**

Figure 1: Critical checks passed, by process

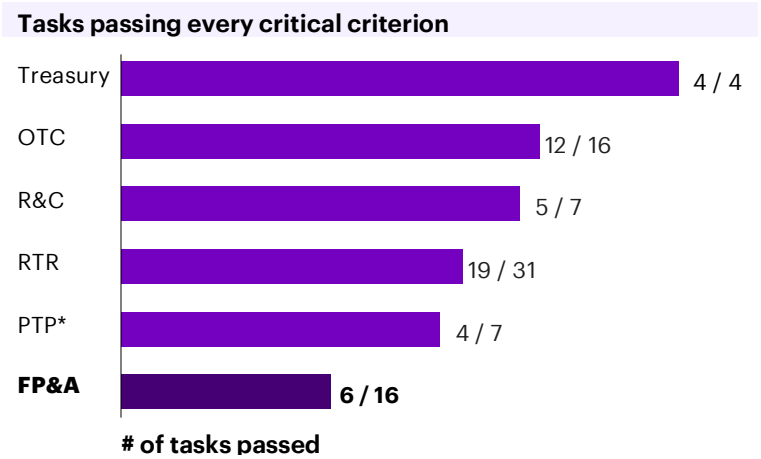
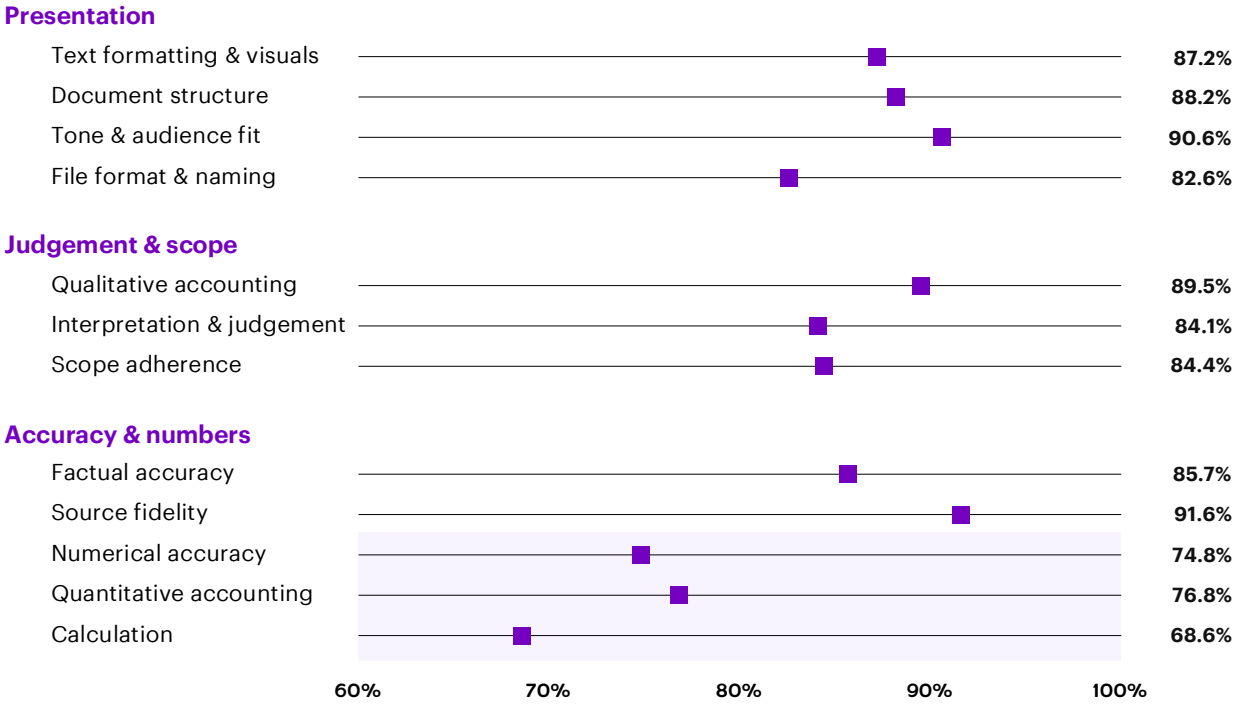


Figure 2: Model performance by workflow dimension



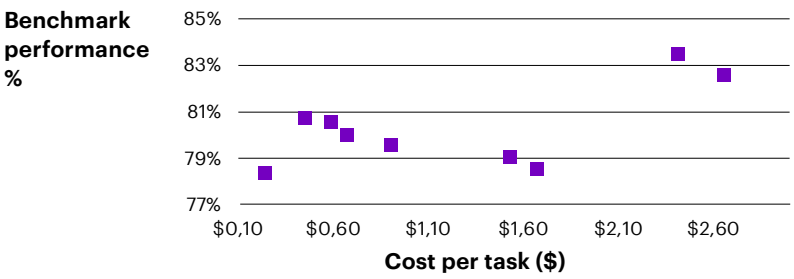
EnterpriseVAL for Finance, 81 use cases. Peer median across the other frontier models

With new model release, performance gains compound.

In the EnterpriseVAL study, each new model release achieved higher competency scores. Weaknesses identified in earlier versions often corrected in subsequent ones. In several cases, models were able to complete tasks with a perfect score of 100%. With one model, the number of tasks passed cleanly increased by 5% in the new model release.

Cost can become the deciding factor once you clear the quality bar.

Figure 3: The trade-off between model performance and cost



10X

Range of cost variance per task. Cost varies from $0.25 to $5 while all nine models score within 5 points of each other.

Human review is still critical where outputs feed into material financial decisions.

For workloads that already include a practitioner, the cost difference is difficult to ignore. Human involvement should be optimized by workflow dimension, weighting review toward critical-fail risk rather than spreading it evenly.

Human benchmark

The human benchmark tested five frontier AI models on the same 25 finance and accounting tasks using three complementary methods:

1. Blinded expert comparison across structure, style, format, accuracy, factual faithfulness and professional judgement.
2. Detailed expert scoring against task-specific criteria, including critical pass-or-fail requirements.
3. An automated judge applying the same rubrics.

No model matched a qualified practitioner consistently. Across 300 blinded comparisons the **qualified practitioners response was preferred in 65% of cases.**

Nearly half of model outputs contained substantive errors, e.g. incorrect figures, postings, rankings or decisions. Models often produced strong answers with low-volume but high-stakes errors (critical failures).

64 of 125

model outputs contained at least one error a practitioner could not sign off.

Frontier models produce finance work that reads as sign-off ready but is wrong about half the time. **Only 27 of 125 outputs were preferred by one expert despite a disqualifying error.**

Expert judgement is the final test of AI readiness: although some models performed well on automated scoring, most did not meet the standard required for expert approval. Overall, the human benchmark helps us measure both the quality of the work and whether it is sufficiently correct and appropriate for a qualified practitioner to approve and use.

IV.

Pathway to success

Whether you are developing AI systems or deploying them, evidence-led insight and a deep understanding of model performance is key. Understanding model performance can help identify which areas of the business where AI can add the most value. We propose a pathway to success designed to maximize the value of enterprise AI ambitions and investments. Building greater confidence in where AI will perform, where it will create value and where further intervention is required.

For model developers, the opportunity goes beyond benchmarking models, by turning enterprise knowledge, process expertise and proprietary data into a repeatable system for improving AI performance and setting up deployments for success.

- **Turn evaluation into an optimization roadmap.** Use benchmark results to identify capability gaps and failure modes, then translate those insights into a prioritized roadmap for model improvement, workflow redesign and targeted data acquisition.

- **Optimize AI deployment around value and performance.** Compare quality, cost and performance to determine where AI can operate independently, where human oversight remains necessary and which combination of human and AI delivers the strongest outcome across the workflow.

- **Build benchmarks around the work that matters.** Create organization-specific benchmarks using representative workflows, input files, expert reference outputs and purpose-built grading rubrics. This provides a repeatable and defensible way to test models, agents and AI solutions against unique operating context before deployment.
- **Establish deployment guardrails and assurance.** Translate benchmark results into clear thresholds for autonomy, human oversight, escalation and control, so models and agents can be deployed with confidence in higher-risk workflows.
- **Create a continuous evaluation capability.** Track quality, failure modes, cost and performance as models and AI solutions evolve. Continuous evaluation creates a learning system that helps both enterprises and AI labs understand how performance is changing and where investment should be directed next.

How Accenture can help

Accenture brings together deep domain expertise, curated enterprise data and the evaluation infrastructure needed to turn feedback into improvement. EnterpriseVAL shows how these capabilities can be connected: real workflows become expert-authored tasks with realistic input files, reference deliverables, task-specific rubrics and critical criteria. This creates evidence at the level at which work is reviewed and approved, rather than relying on a general model score.

01 **Deep domain expertise.** Accenture practitioners identify the decisions, exceptions and tacit standards that determine whether an output is usable. They translate those expectations into measurable criteria and distinguish between work that is ready for greater autonomy and work that still needs human judgement, controls or redesign. This helps clients focus investment while making the required safeguards explicit.

02 **Curated datasets.** Public benchmarks provide a broad market view, but enterprise deployment requires private evaluations built around an organization's own processes, documents and standards. Accenture can create held-out benchmark sets alongside related post-training, fine-tuning, preference and feedback data. Keeping evaluation data separate preserves its integrity, while failure analysis directs new data towards the behaviors that matter. These assets support model selection, targeted improvement and the development of agents that operate within a client's context.

03 **Reinforcement learning and evaluation infrastructure.** Rubrics and critical criteria can operate as regression guardrails and hill-climbing targets. When a model, tool, orchestration layer or policy changes, the full system can be retested against the same business outcomes. Accenture is immersed in the digital core and enterprise platforms, the very environments critical to agentic testing and training, with more specialists who understand the systems behind it from ERP, CRM and data platforms to cloud, controls and operations.

04 **From evidence to deployment.** Accenture can help clients evaluate readiness, diagnose where performance breaks down, improve the model and surrounding system and operate continuous assurance once it is live.

V.

Conclusion

The evidence is the differentiator. Enterprise AI leadership now comes down to proof. The companies that pull ahead are not the ones with access to the newest model, they are the ones who can say, with evidence, which model does which work, to what standard and at what cost. Everyone else is buying capability on trust and hoping value follows.

The headline numbers invite that trust. On rubric scores alone, nine frontier models sit inside a five-point band, close enough that a reader would reasonably call them interchangeable for finance work. **Apply the critical-failure gate, the criteria a practitioner marked as unrecoverable and the picture inverts: scores fall to between 47% and 56%. No model keeps much more than two thirds of its raw score and models that looked identical separate by several points.** Roughly half of outputs contained an error serious enough to block sign-off, and expert reviewers still preferred a quarter of those flawed outputs over a qualified practitioner's own work. The errors are real, material and largely invisible to the naked eye.

What separates strong and weak performance is not the finance domain but the nature of the work. Quantitative criteria, calculation, numerical accuracy and quantitative accounting treatment are consistently the weakest. While qualitative ones, tone, source fidelity and qualitative treatment, hold up well. It tells you where autonomy is safe, where a reviewer belongs, and where the next round of data, tuning and workflow redesign should be aimed.

Once quality is understood, cost becomes the decision. Cost per task varies tenfold across a field separated by five points of quality, and the highest scorers are also the dearest. For any workflow that already has a practitioner in the lead, that difference is difficult to justify on capability alone.

None of this is specific to finance. The method holds wherever work is complex enough to require judgement and consequential enough that errors matter: banking, cybersecurity, supply chain, energy, HR, insurance. Accenture has practitioners in all of them.

The questions arriving from the board, the CEO and the CFO are no longer whether to invest in AI, but which models, in which workflows, at what cost and with what controls.

Build the evidence base now and you can answer the questions that matter.

Citations

1. A. Menon, L. Guan, J. Burrows, and S. Mukherjee. The CIO's guide to AI tokenomics: how to see, control and account for AI token spend at scale. Accenture Research, 2026. https://www.accenture.com/content/dam/accenture/final/accenture-com/document-fy27/q1/ACN-CIO-Guide-AI-Tokenomics-Report.pdf

2. APQC. Total cost to perform the finance function as a percentage of revenue. Open Standards Benchmarking, n.d. https://www.apqc.org/resources/benchmarking/open-standards-benchmarking/measures/total-cost-perform-finance-function-1

3. Fortune. Fortune Global 500 explorer. Interactive data dashboard, 2026. https://fortune.com/ranking/global500/2026/explore/

4. Accenture. "Pulse of Change" Accenture Research, 2026. https://www.accenture.com/en/insights/pulse-of-change

About Accenture

Accenture is a business reinvention and technology company. Our ambition is to be the reinvention partner of choice for our clients and lead in the safe, widespread adoption of AI, and to be the most client-focused, AI-enabled, great place to work in the world. Our approximately 814,000 people bring deep insights across industries and functions, with capabilities spanning strategy, process, operating model, change management, talent, data, technology and AI. We deliver this expertise through consulting, managed services, proprietary assets and platforms, and increasingly, software. We bring innovation and new solutions informed by our experience across industries, enabling us to apply relevant insights, capabilities and leading practices from adjacent sectors to address emerging business challenges and opportunities. We serve more than 9,000 clients and generated revenues of approximately $74 billion in fiscal year 2026.
Visit us at accenture.com.

Disclaimer: The material in this document reflects information available at the point in time at which this document was prepared as indicated by the date provided on the front page, however the global situation is rapidly evolving and the position may change. This content is provided for general information purposes only, does not take into account the reader's specific circumstances, and is not intended to be used in place of consultation with our professional advisors. Accenture disclaims, to the fullest extent permitted by applicable law, any and all liability for the accuracy and completeness of the information in this document and for any acts or omissions made based on such information.

Accenture does not provide legal, regulatory, audit, or tax advice. Readers are responsible for obtaining such advice from their own legal counsel or other licensed professionals. This document refers to marks owned by third parties. All such third-party marks are the property of their respective owners. No sponsorship, endorsement or approval of this content by the owners of such marks is intended, expressed or implied.

Copyright © 2026 Accenture. All rights reserved.
Accenture and its logo are registered trademarks of Accenture.