Beyond the AI Hype
Evaluating LLMs vs. Digits AGL for Accounting Tasks

Hannes Hapke, Jo Pu, Siva Manivannan, Cole Howard, Chris Hassell

ml@digits.com

Digits Financial, Inc.

Last revision: June 18, 2025

1 Executive Summary

1.1 Key Findings and Implications

- **Performance ceiling:** No general-purpose LLM exceeded 70% accuracy.
- **Specialized advantage:** Digits' purpose-built system significantly outperformed all general-purpose models.
- **Human parity:** Specialized systems still outperform both outsourced accountants and LLMs, though general LLMs are approaching human performance in accounting.
- **Reasoning models:** Enhanced reasoning provided no additional benefit, but did increase costs and model request latency.

1.2 Methodology Overview

- We evaluated 17,792 financial transactions from over 100 randomly selected small businesses, providing consistent prompts across 19 different models from major providers, including OpenAI, Anthropic, Meta and xAI.
- To facilitate a clearer comparison of results, this revised study maintained a consistent underlying dataset, as used in our March 2025 study.
- Compared against a baseline created by 12 hired professional accountants and Digits' specialized ML system.
- Since the prior release (March 4th, 2025), we increased the number of tested models from 13 to 19 models and introduced various reasoning configurations for a more holistic study.

2 Introduction

In today's rapidly evolving financial technology landscape, the promise of artificial intelligence for accounting automation has generated significant interest, challenges, and considerable hype. As businesses increasingly seek automation solutions for transaction classification, understanding the capabilities and limitations of large language models (LLMs) in this domain becomes crucial for technology decision-makers. We evaluate Digits' proprietary, specialized ML model against 19 general-purpose LLMs across multiple performance dimensions including accuracy, latency, and reliability.

While recent advances in language model capabilities have demonstrated impressive results across many domains, can general-purpose models effectively address the nuanced, subjective needs of financial classification tasks?

Our comprehensive evaluation provides answers backed by real-world data from over 17,000+ transactions.

3 Background

Accounting practices are inherently subjective, with significant variations in how individual accountants approach classification and decision-making processes. This subjectivity presents a fundamental challenge for large language models (LLMs) attempting to generalize accounting knowledge globally. **While these models can learn general accounting principles, they struggle to replicate the nuanced judgment experienced accountants develop through years of practice in specific industry contexts.** This limitation becomes particularly evident when models trained on generalized data attempt to mimic the decision-making patterns of individual accountants working within specialized domains.

A concerning trend in the accounting technology landscape involves new entrants who rely solely on 3rd party, closed-source models for their accounting automation solutions. These companies frequently share sensitive financial transaction data with large AI providers such as OpenAI, raising significant questions about data privacy and security. This practice creates potential vulnerabilities for businesses that may not fully understand the extent to which their financial information is being processed and stored by third-party AI systems outside their direct control.

We were interested in comparing how Digits' proprietary machine-learning system, designed specifically for financial transaction classification, performs against state-of-the-art LLMs. Digits has developed a specialized ML system tailored for accounting tasks, and we wanted to determine if there is a significant competitive advantage compared to solutions like GPT-o3 or Anthropic's Claude-4 models.

4 Methodology

4.1 Data and Ground Truth

Our study utilized a comprehensive dataset comprising **17,792 financial transactions from over 100 randomly selected businesses that use Digits**. These transactions occurred between November 1, 2024, and February 1, 2025, providing a recent and representative sample of accounting data. The random selection process for clients ensured that our findings would broadly apply across various business types and accounting practices.

The dataset reflected typical transaction patterns with an 88% to 12% split between debits and credits, mirroring the natural distribution commonly observed in accounting systems. The split is skewed towards debit transactions because most transactions' credit side can be defined by their source (e.g., a given bank account).

The complexity of the accounting structures varied significantly across businesses, with the number of categories in their Charts of Accounts ranging from as few as 15 to as many as 281 categories. The average business maintained 92 categories, while the median was 66, indicating a right-skewed distribution where some businesses maintained substantially more complex accounting structures than others.

GAAP-trained US-based accountants have reviewed all transactions and expected categories, providing a solid basis for this comprehensive comparison.

4.2 Establishing a Human Baseline

In this revised study, we introduced a crucial human baseline component to understand the real-world implications of scaling financial operations today. Our objective was to observe and quantify the effort required for outsourced accountants to perform a fundamental task: transaction classification.

To achieve this, **we hired 12 experienced accountants and graduating senior accounting students to participate**. These 12 individuals were divided into four groups, with three participants per group. Each group was tasked with independently classifying 500 financial transactions, resulting in a total of 2000 transactions classified across the entire cohort. To allow the accountants to maximize their time by focusing on a single business's chart of accounts, we selected 500 transactions from the same business.

It is important to note that the 2000 transactions used in this human baseline study constitute a subset of the much larger dataset utilized in the broader scope of this research. This approach ensured that the human classification task was representative of the types of transactions encountered in real-world business scenarios, while also allowing for a manageable and focused evaluation of human performance and scalability challenges.

4.3 Prompt Construction

The primary objective of our evaluation was to predict the appropriate category from each business's Chart of Accounts (CoA) for a given transaction rather than mapping to a generic, standardized CoA. Predicting categories in a standard CoA is a straightforward but highly unrealistic scenario. In practice, each business wants their financial activity reflected in a customized CoA to emphasize their unique business aspects (e.g., travel expenses broken down by department or specific CoGS categories).

Predicting categories of a business-specific CoA better reflects real-world accounting practices where businesses maintain individualized category structures. To provide context for the classification task, we supplied the models with the "other side" of each transaction, as this information would typically be available from their bank feed, the credit card provider, payroll partner, or other integration source and provides valuable contextual clues.

Each model and accountant received the transaction description, amount, and the list of available categories for the relevant client. We excluded the category from the opposing entry side (debit vs. credit) since it's rarely appropriate for classifying the transaction in question.

In double-entry accounting, each transaction has two sides. When classifying a specific transaction, the category that applies to one side (e.g., a debit) is typically not suitable for the other side (the credit), so we intentionally removed it from consideration.

We did not change the prompts for individual models; instead, we used a consistent prompt structure across all evaluations to ensure a fair comparison. While we requested all LLMs to generate JSON output for standardized processing, we chose not to encode categories as structured enums. Although we initially explored this approach, we ultimately decided against it due to the prohibitively slow processing times caused by the ample token search space created by numerous categories.

4.3.1 Prompt Setup

Prompt 1 shows an example of the prompt we used for the model evaluation.

4.3.2 Prompt Parameters

To ensure consistent and comparable results across models, we standardized the following parameters:

- **Temperature Setting:** Set to approximately zero for most models to minimize non-deterministic outputs, with the exception of OpenAI's o1 and o3 variants which require a temperature of 1.0 per API specifications. OpenAI [2025] Wang et al. [2023]
- **Output Length:** Limited to 1,024 tokens for standard models, with an extended limit of 10,240 tokens for reasoning-focused models to accommodate their explanatory capabilities.

Prompt 1: Example Prompt Template for Model Evaluation

```
Given the following transaction description from a liability or asset account:

'''
Description: UBER *TRIP. Merchant name: Uber
Amount: $44.80
'''
Given that this will be recorded as a credit to Mercury Credit (0000) - 1
Which category should receive the debit side?

'''
"Software & Apps"
"Travel"
"Meals"
⟨other categories removed for privacy reasons⟩
'''
* I want you to think of the most likely category from the list above for the described transaction and amount
* Think of a single sentence description of why
* Double check that the category is actually in the list
NOTE: Do not provide explanations. Only provide the most relevant category.
Return them as JSON with the schema:
{"category": ⟨string⟩}
```

- **Domain Expertise Prompting:** All models except OpenAI's o1-mini received the standardized system prompt: [1] `"You are an expert bookkeeper with deep knowledge of accrual-accounting and an eye for detail."`

4.4 Model Providers

We evaluated models from all major providers, including OpenAI, Anthropic, and Meta, between June 10 and June 18, 2025. To standardize deployment for open-source models, we leveraged together.ai's infrastructure.

For consistent evaluation, we exclusively used the OpenAI client SDK for all model connections, avoiding provider-specific SDKs.

For this comprehensive study, we increased the number of evaluated models from 16 to 19, and included various reasoning configurations. We included the following models in this study:

Provider	**Model**
OpenAI	o1 (medium) o1-mini o3 (low, medium, high) o3-mini (medium) o3-pro GPT-4o GPT-4o-mini GPT-4.1
Google	Gemini-2.5-pro
Anthropic	Claude-3-7-sonnet (medium) Claude-4-0-sonnet (medium) Claude-4-0-opus (medium)
xAI	grok-3
Meta	Llama-3.3-70B-Instruct-Turbo Llama-4-Scout Llama-4-Maverick
Deepseek	DeepSeek-V3 DeepSeek-R1
Alibaba Cloud	Qwen3-235b (thinking, no thinking)

Table 1: List of evaluated models by provider

Access to some model APIs was heavily restricted by providers, preventing testing with our real-world dataset. This included OpenAI's recently released o3-pro, and Google's Gemini-2.5-pro. While these models were considered for benchmarking with our smaller dataset, their use with real-

[1] At the time of writing this paper, o1-mini didn't allow system prompts as part of chat completion requests.

world accounting transaction volumes presents significant drawbacks due to the usage restrictions per day.

4.5 Reasoning Models

This study expands upon our March 2025 research by incorporating a greater number of reasoning models. We also increased the allowance for generated tokens, providing these models more ”room” for their thought processes. This differs from our March 2025 study, which prioritized high throughput and capped models at 512 tokens. Given the needs of more recent models, higher token allowances are now necessary. Furthermore, we explore various reasoning settings and assess their impact on domain-specific capabilities in classifying accounting transactions.

4.6 Hallucination Assessment

We defined hallucinations as instances where models generated categories that did not exist in the client’s Chart of Accounts. For this study, we treated all Charts of Accounts as complete and comprehensive, meaning that any category suggested by a model not present in the client’s CoA was considered a true hallucination rather than an indication of a missing category that should have been included. This assumption allowed us to quantify hallucination rates systematically across different models. Furthermore, **all transactions have been reviewed by GAAP-certified accountants for correctness to be confident that the expected ground truth reflects the correct and true category, and that such category was already present in the CoA.**

4.7 Digits’ Proprietary System

While we cannot disclose proprietary details about Digits’ ML systems, we can share several key characteristics relevant to this comparative study. The Digits platform utilizes a compilation of multiple proprietary machine learning models, all of which are hosted and trained in-house to ensure data security and system integrity. One notable strength of the Digits ML system is its approach to hallucination prevention through a proprietary workflow designed for accounting applications.

Digits’ ML system consistently outperforms even the fastest general-purpose LLMs evaluated in this study. This performance advantage is particularly significant given the time-sensitive nature of many accounting processes.

4.8 Performance Analysis

Our comprehensive evaluation of current AI models revealed several significant performance patterns and operational considerations. OpenAI’s o3 model demonstrated superior performance among all tested systems, achieving a marginal but notable improvement over GPT-4.5, which had previously established itself as the leading model in March 2025 with a 66.6% accuracy rate.

Despite these advances, the evaluation highlighted a critical performance ceiling that remains unbroken across the industry. No model in our assessment successfully crossed the 70% accuracy threshold, indicating persistent challenges in achieving higher-level performance benchmarks. The observed lower average performance across our model cohort can be attributed to our expanded testing methodology, which incorporated a broader range of open-source models alongside proprietary systems, providing a more comprehensive view of the current landscape.

A particularly noteworthy finding emerged regarding hallucination rates in reasoning models. We documented a significant reduction in hallucination incidents when response generation processes were allowed to complete their full cycles without early termination. However, this improvement introduces a critical trade-off for practical implementation: the extended processing time required for complete response generation substantially impacts transaction throughput, presenting significant operational challenges for accounting systems and other high-volume applications where processing speed is essential for business continuity.

4.8.1 Model Accuracy

Despite the impressive capabilities of leading models, we observed a consistent performance ceiling across all general-purpose LLMs. None of these models achieved accuracy rates exceeding 70% on the transaction classification task, regardless of size or recency, as shown in Figure 1. This limitation stems from a fundamental constraint: the general-purpose models lack critical contextual information beyond the transaction description that outsourced accountants naturally incorporate into their classification decisions. **This missing contextual layer includes business operations patterns, industry-specific accounting practices, and historical classification precedents for similar transactions.**

Notably, this performance ceiling was overcome by Digits' proprietary ML system, which has been specifically designed to incorporate these additional contextual elements. This system's superior performance underscores a crucial insight from our research: **the inherent subjectivity of accounting classification cannot be adequately addressed by even the most advanced general-purpose language models operating in isolation.**

In addition to the subjectivity, large language models cannot differentiate transactions based on their source accounts. Clients frequently use multiple banks or credit cards for distinct purposes—such as office expenses versus cost of goods sold (CoGS) — yet these different accounts often process identical transaction types. **While a generic LLM typically categorizes transactions based solely on overall probability patterns from its training data, Digits' purpose-designed machine learning system excels by recognizing the source context.** Our system can effectively "blank out" similar transactions from different sources and route them to appropriate categories based on their origin — a capability that remains challenging for generic LLMs.

The most significant takeaway from our analysis is that effective accounting automation requires systems carefully tuned to the unique characteristics of financial classification tasks. **The subjective nature of accounting decisions, which often vary significantly between businesses even within the same industry, creates a challenge that cannot be solved through general language understanding alone.** Instead, effective solutions require specialized systems that can capture and apply the implicit classification patterns specific to individual businesses and their

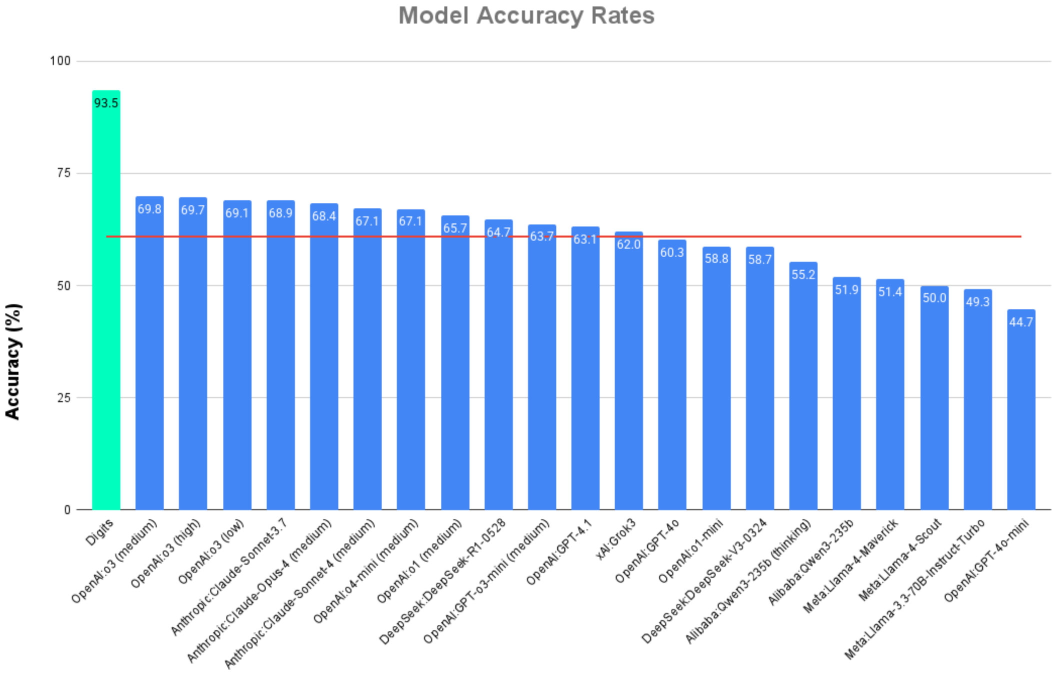

Figure 1: Comparison of Model Accuracy (higher is better)

accountants.

4.8.2 Model Request Latency

Recent advancements in large language models have demonstrated significantly enhanced reasoning capabilities compared to their predecessors. However, these improvements in cognitive performance have come at the cost of substantially increased model latency during inference. This latency increase can be attributed to two primary factors: first, the architectural design decisions implemented by model providers in their latest iterations, and second, the computational overhead associated with generating additional reasoning tokens required for the models to arrive at conclusive categorizations or decisions.

While the extended processing time of reasoning models might be justified in complex decision-making scenarios, **our findings indicate that this additional computational investment did not translate to proportionally better classification accuracy for accounting tasks.** The correlation between increased latency and improved performance was notably weak, suggesting that the factors limiting classification accuracy extend beyond the reasoning depth afforded by longer processing times.

The average latency across all models increased to 5.04 seconds per request (vs. 3.67 seconds per request in our March 2025 study), significantly varying between the fastest and slowest performers. **This timing constraint becomes particularly consequential when considering real-world accounting applications that must process substantial transaction volumes.** Even with

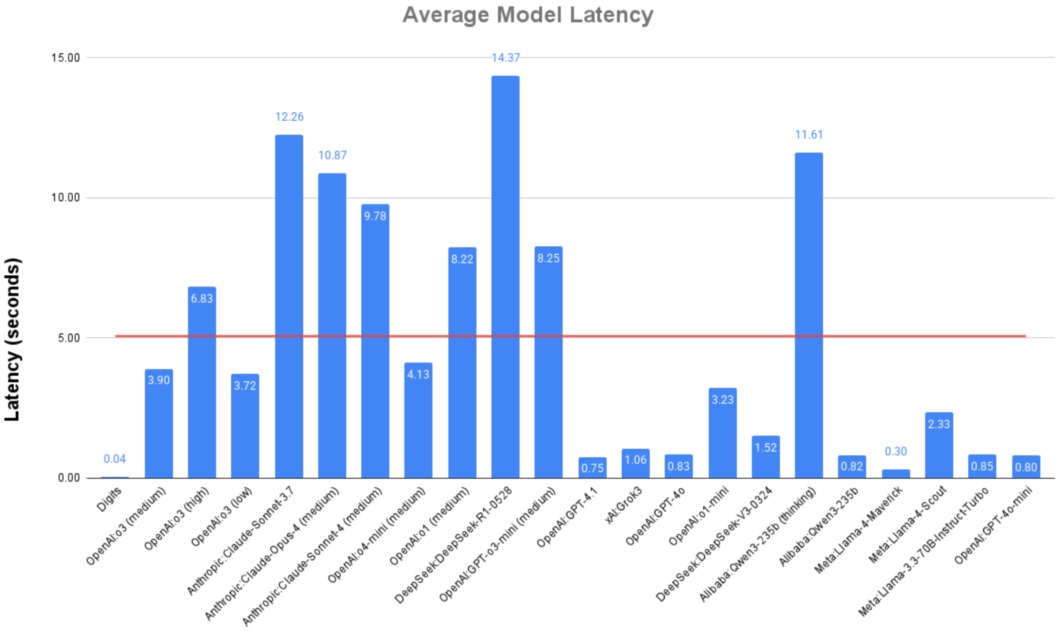

Figure 2: Comparison of Model Latency (smaller is better)

relatively modest 2-second latencies, scaling to millions of transactions creates significant operational challenges for classification systems. This reality necessitates sophisticated parallelization strategies and robust infrastructure to maintain reasonable processing timelines for large-scale accounting operations.

These latency findings in Figure 2 highlight a critical consideration for organizations implementing AI classification in accounting workflows: **the trade-off between processing speed and marginal accuracy improvements must be carefully evaluated within specific business requirements and transaction volumes.** In many practical scenarios, a slightly less accurate model with substantially faster response times may provide a better overall value than a marginally more accurate but significantly slower alternative. This also poses a significant risk for accounting solutions that select model providers before they reach meaningful transaction volumes.

4.8.3 Model Hallucination Rates

Hallucination rate emerged as the most discriminative performance metric. When models were allowed extended reasoning time, hallucination rates decreased significantly. However, this required substantial token generation (often thousands of tokens per classification), resulting in high latency costs.

Unlike our March 2025 baseline, **current evaluations show a strong positive correlation between accuracy and low hallucination rates across all models**, suggesting improvements in model training or architecture shown in Figure 3.

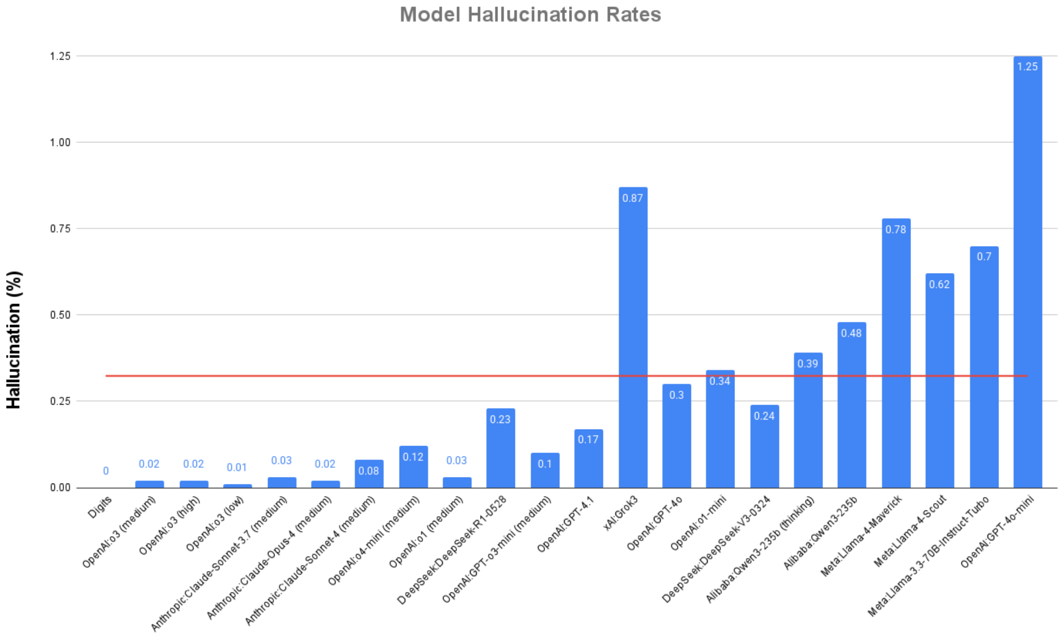

Figure 3: Comparison of Model Hallucination Rates (smaller is better)

The practical implications of these hallucination rates are substantial when considered at scale. Even a modest 0.1% hallucination rate translates to 1,000 misclassified transactions for every million processed. In enterprise accounting environments where transaction volumes regularly reach millions, this error rate would impose a significant operational burden on financial teams needing to identify and correct these misclassifications. Such rework requirements would substantially diminish the expected efficiency benefits of implementing AI-assisted classification systems.

These findings emphasize the critical importance of hallucination prevention mechanisms in accounting automation systems. Models deployed in financial contexts must prioritize accuracy and reliability over generative flexibility, with specialized constraints that prevent the suggestion of non-existent categories. This requirement further supports the case for purpose-built financial classification systems rather than the adaptation of general-purpose LLMs for accounting applications.

5 Comparison with Human Baseline

To establish a comprehensive baseline for transaction classification performance, we conducted a controlled study involving 12 professional accountants who were compensated for their participation. The participants were organized into four groups of three accountants each, with every group responsible for classifying 500 transactions, resulting in a total dataset of 2,000 classified transactions. To achieve a strong performance signal while minimizing human classification effort, we reduced the dataset to 2,000 transactions, selected as a subset of our original dataset.

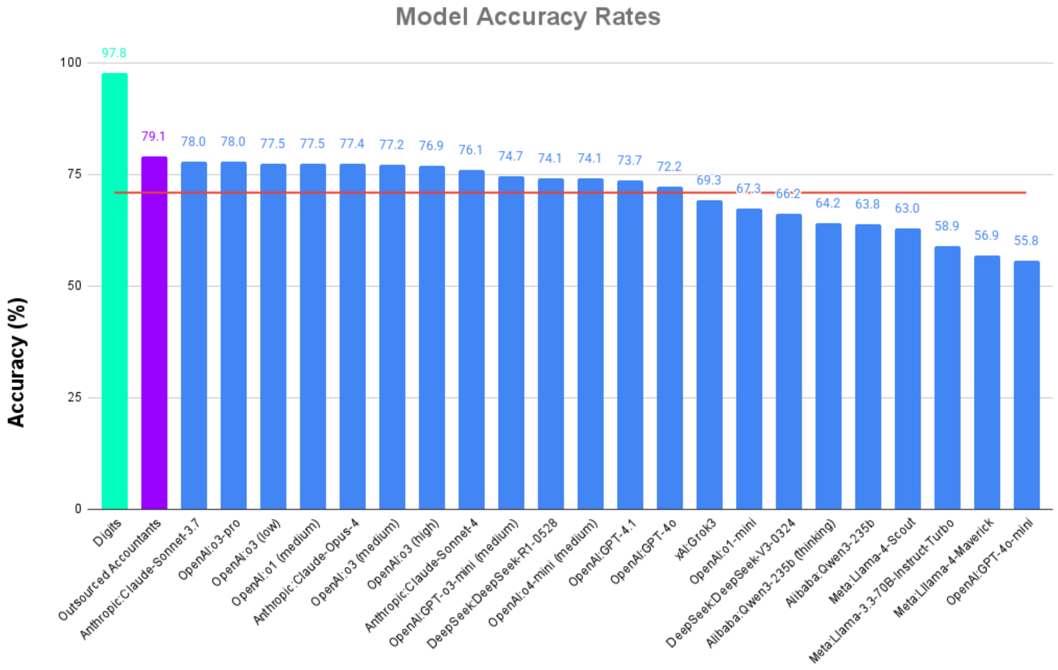

Figure 4: Comparison of Human Accuracy Rates (higher is better)

Critically, all outsourced accountants received identical information sets to those provided to the large language models (LLMs) being evaluated, ensuring consistency in available data across both human and artificial intelligence classification approaches.

The study revealed significant insights into human classification consistency and efficiency. **The accountants showed an accuracy of 79.1% when compared to our GAAP accountant-reviewed dataset. Among the 2,000 total transactions processed, 10.4% exhibited classification disagreement** within the three-person accountant groups, defined as cases where no two accountants converged on the same classification. This disagreement rate provides a crucial benchmark for understanding the inherent complexity and ambiguity present in real-world transaction classification tasks. From a time efficiency perspective, **the accountants took an average of 4 hours and 43 minutes to complete their 500-transaction allocation**, translating to approximately 34 seconds per transaction per accountant.

Analysis of the transaction subset revealed important characteristics that influenced overall performance metrics. The selected transactions contained relatively fewer edge cases and incorporated a higher proportion of repeat transaction types compared to typical operational datasets. This composition contributed to an observable increase in accuracy rates across all evaluated systems, including both the Digits platform and the LLMs under assessment. The reduced complexity of the transaction set suggests that performance metrics derived from this study may represent an upper bound for classification accuracy under more challenging, real-world conditions.

The study's primary finding demonstrates that LLM classification performance aligns closely with the accuracy rates by outsourced accountants, indicating that current artificial intelligence systems have achieved near-human parity in transaction classification tasks. However, both outsourced accountants and LLMs significantly underperformed compared to Digits' GAAP-verified classification system, highlighting the continued value of specialized, compliance-focused platforms. This performance gap suggests that while LLMs represent a substantial advancement in automated accounting processes, purpose-built systems with embedded GAAP expertise continue to provide superior accuracy for regulatory compliance and financial reporting requirements.

6 Do Reasoning Models Actually Provide a Benefit?

Our comprehensive evaluation of reasoning-enhanced AI models revealed significant performance trade-offs that challenge their viability in high-throughput accounting applications. We conducted systematic testing across both closed-source and open-source model architectures to assess the relationship between enhanced thinking capabilities and practical performance metrics.

In our analysis of closed-source models, we observed concerning performance characteristics when reasoning capabilities were activated. Specifically, our evaluation of OpenAI's o3 model demonstrated a near-doubling of latency rates without corresponding improvements in accuracy, shown in Table 2. This finding suggests that the computational overhead associated with enhanced reasoning processes in closed-source architectures may not translate to meaningful performance gains in accounting-specific tasks. The lack of accuracy improvement, combined with substantial latency increases, raises questions about the cost-effectiveness of deploying such models in production accounting environments.

Model Configuration	**Accuracy (%)**	**Latency (s)**
OpenAI:o3 (high)	69.7	6.83
OpenAI:o3 (medium)	69.8	3.90
OpenAI:o3 (low)	69.1	3.72

Table 2: Performance comparison of OpenAI o3 model with different reasoning configurations

To validate these findings across different model architectures, we conducted controlled experiments using Qwen3, an open-source model that allows for granular control over reasoning capabilities. By toggling the thinking capabilities on and off, we were able to isolate the specific impact of reasoning enhancements on system performance as shown in Table 3. Our results indicated that activating thinking capabilities resulted in a greater than 10-fold increase in latency while yielding only a modest 4% improvement in accuracy. This dramatic latency penalty for minimal accuracy gains further reinforces concerns about the practical applicability of reasoning-enhanced models in time-sensitive accounting operations.

Model Configuration	**Accuracy (%)**	**Latency (s)**
Alibaba:Qwen3-235b (thinking)	55.2	11.61
Alibaba:Qwen3-235b (thinking disabled)	51.9	0.82

Table 3: Performance comparison of Qwen3 model with and without reasoning capabilities

Given the high throughput requirements inherent to modern accounting systems, where processing

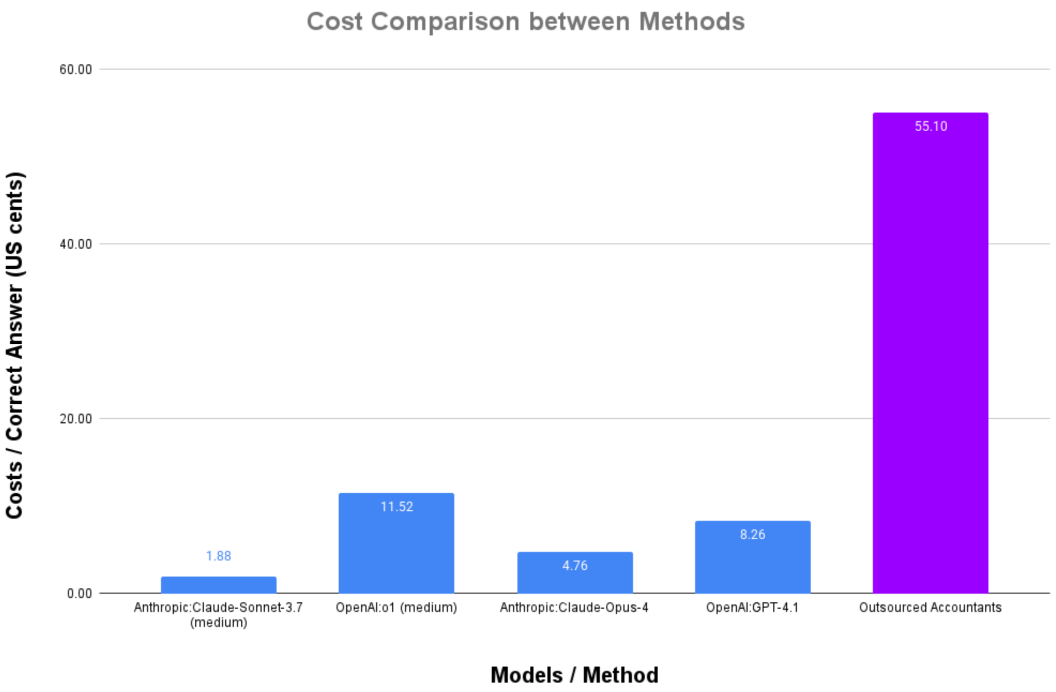

Figure 5: Cost Comparison between Classification Methods (lower is better)

speed directly impacts operational efficiency and user experience, our findings suggest that reasoning-enhanced models do not currently provide sufficient value to justify their implementation costs. The substantial latency penalties observed across both closed-source and open-source architectures, coupled with marginal accuracy improvements, indicate that traditional AI models may remain more suitable for accounting applications where rapid processing of large transaction volumes is prioritized over incremental accuracy gains. Organizations considering the adoption of reasoning models should carefully weigh these performance trade-offs against their specific operational requirements and tolerance for increased processing times.

7 Cost Analysis

Our LLM testing revealed significant cost differences among model types and providers. This led us to investigate the cost per correct answer for each method, calculated as the ratio of total classification generation costs to the number of correct answers produced.

Figure 5 illustrates the cost comparison for four closed-source LLMs and the expenses associated with accountants manually reviewing transactions. This highlights substantial cost discrepancies between manual classification and LLMs, as well as notable variations in cost among the different LLM model types.

8 Challenges and Limitations

Our research encountered several technical obstacles that merit careful consideration when evaluating the feasibility of integrating LLMs into accounting workflows. These challenges affected our benchmark methodology and raised important questions about the practical application of these technologies in production environments where reliability and performance are paramount.

API reliability emerged as a consistent challenge across all model providers tested in our study. Each provider required thoughtful implementation of retry handling mechanisms to manage intermittent failures and ensure complete data collection. These reliability issues highlight potential concerns for accounting applications where consistent availability is essential, particularly during peak financial periods such as month-end or year-end closings when transaction volumes spike significantly.

We observed substantial variations in throughput capabilities among different providers. Most notably, OpenAI's latest models o3-pro and Google's Gemini-2.5-pro imposed strict limitations on request parallelization, forcing us to process transactions sequentially rather than in parallel batches or abandon the tests for the models entirely. This constraint significantly extended testing timelines and would present serious scalability challenges in production environments handling large transaction volumes.

Perhaps most concerning from an implementation perspective were the extended processing times observed across multiple models. While some providers delivered reasonably quick responses, others exhibited latencies that would make real-time transaction classification impractical. This performance variability raises questions about the viability of incorporating certain LLMs into accounting systems where users expect immediate feedback during transaction entry or reconciliation processes. The significant performance gap between our evaluation's fastest and slowest models suggests that processing speed should be a critical consideration when selecting AI technologies for accounting applications, potentially outweighing marginal accuracy improvements in many practical scenarios.

These technical limitations underscore the importance of evaluating the classification accuracy of LLMs for accounting tasks and their operational characteristics in realistic deployment scenarios. Organizations considering these technologies should carefully assess whether the infrastructure requirements, reliability patterns, and response times align with their specific accounting workflow needs and user experience expectations.

9 Future Directions

This paper presents the second revision of our industry-recognized study examining the application of large language models (LLMs) for accounting tasks. As part of our commitment to maintaining a comprehensive and ongoing evaluation process, we continue to track developments across the accounting technology landscape, recognizing that model capabilities are advancing at an unprecedented pace. We acknowledge that emerging innovations in the field may effectively address several of the challenges identified in our current evaluation framework, particularly those concerning processing speed limitations, hallucination rates, and the models' domain-specific understanding of fundamental accounting principles. This iterative approach ensures our research remains relevant and provides practitioners with current insights into the evolving capabilities and limitations of LLM implementations in professional accounting environments.

The accounting domain presents uniquely complex challenges for AI systems due to its combination of structured rules and subjective professional judgment. Future evaluations will expand our focus to assess how emerging models handle increasingly nuanced accounting scenarios, including complex multi-line transactions, industry-specific classification patterns, and adaptation to changing accounting standards.

Our commitment to continued evaluation will help accounting professionals and technology leaders make informed decisions about incorporating artificial intelligence into their financial workflows as these technologies mature. By maintaining rigorous benchmarking standards across new models as they emerge, we aim to provide the accounting community with reliable insights into which approaches truly advance the state of the art for financial classification tasks.

10 Conclusion

Our comprehensive evaluation of leading AI models for accounting transaction classification reveals critical insights that should inform technology adoption decisions in the financial sector. While the rapid advancement of large language models has generated considerable enthusiasm for their potential applications in accounting automation, our findings demonstrate that significant challenges remain for practical implementation in production environments.

While general-purpose LLMs may serve as useful tools for certain accounting tasks, mission-critical transaction classification workflows are best served by specialized systems designed specifically for the unique challenges of financial data processing.

10.1 Performance Limitations of General-Purpose Models

The most significant finding of this study is the persistent performance ceiling observed across all general-purpose LLMs tested. Despite evaluating 19 models from leading providers, including OpenAI's latest o3 variants, Anthropic's Claude 4 family, and other state-of-the-art systems, none achieved accuracy rates exceeding 70% on real-world transaction classification tasks. This ceiling persisted regardless of model size, architectural sophistication, or recency, indicating fundamental limitations in how general-purpose models approach the nuanced requirements of accounting classification.

The performance gap between general-purpose LLMs and Digits' specialized ML system underscores a crucial insight: effective accounting automation requires purpose-built solutions that can incorporate the contextual elements, business-specific patterns, and domain expertise that accountants naturally apply. The inherent subjectivity of accounting decisions, combined with the need to differentiate transactions based on source accounts and historical classification patterns, creates requirements that extend beyond the capabilities of even the most advanced general-purpose language models.

Operational Viability Concerns Our analysis reveals substantial operational challenges that would impede the practical deployment of current LLMs in high-volume accounting environments. The average request latency of 5.04 seconds, combined with API reliability issues and throughput limitations, presents significant scalability concerns for organizations processing large transaction volumes. These performance characteristics would create unacceptable delays in critical accounting workflows, particularly during peak financial periods when rapid processing is essential. The hallucination rate analysis provides perhaps the most compelling argument against current LLM deployment in accounting contexts. Even modest hallucination rates of 0.1% translate to thousands of misclassified transactions when scaled to enterprise volumes, creating substantial operational burdens for financial teams who must identify and correct these errors. This finding highlights the critical importance of accuracy and reliability in financial applications, where errors can have significant compliance and reporting implications.

Limited Value of Reasoning Models Contrary to expectations, our evaluation of reasoning-enhanced models revealed that increased computational sophistication does not translate to meaningful performance improvements in accounting tasks. Across both closed-source and open-source architectures, reasoning capabilities produced dramatic latency increases—often exceeding 10-fold penalties—while delivering only marginal accuracy improvements. This finding suggests that the factors limiting classification accuracy in accounting contexts are not primarily related to reasoning depth, but rather to the availability of domain-specific context and business intelligence.

Human-AI Performance Parity with Important Caveats The establishment of a human baseline revealed that current LLMs have achieved near-parity with the performance of outsourced accountants on transaction classification tasks. However, this parity comes with important qualifications: both outsourced accountants and LLMs significantly underperformed compared to purpose-built systems designed specifically for accounting applications. The 10.4% disagreement rate among professional accountants on identical transactions also highlights the inherent subjectivity in accounting classification, reinforcing the need for systems that can capture and apply business-specific classification patterns rather than relying on generalized knowledge.

References

OpenAI. OpenAI API: Create chat completion, 2025. URL https://platform.openai.com/docs/api-reference/chat/create#chat-create-temperature.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171.