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: January 9, 2026

1 Executive Summary

1.1 Key Findings and Implications

- **Performance ceiling:** No general-purpose LLM exceeded 73% 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 basic accounting tasks.
- **Reasoning models:** Reasoning models with higher reasoning effort delivered only marginal accuracy gains, while incurring increased costs and inference latency.

1.2 Methodology Overview

- We evaluated 19 state-of-the-art frontier models for transaction classification under a standardized prompting protocol, using a dataset of 17,792 financial transactions collected from more than 100 randomly selected small businesses.
- This revision used the same benchmark dataset as in our 2025 study for comparability.
- We established a human baseline from 12 compensated professional accountants, and compared the baseline performance metrics against general-purpose LLMs and Digits' specialized ML system.
- This study refreshed our benchmark model list of frontier models to include GPT-5.2, Gemini 3 Flash, Claude Opus 4.5, DeepSeek V3.2, and Kimi K2, while exploring various reasoning levels to assess their performance impact.

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 accounting 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 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.

We were interested in how Digits' proprietary machine-learning system, designed specifically for financial transaction classification, compares 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 solely relying on SOTA models such as Google's Gemini 3 Flash, OpenAI's GPT-5.2, and Anthropic's Claude Opus 4.5.

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 15 to 281. 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.

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

4.2 Establishing a Human Baseline

The human baseline quantifies the effort required for outsourced accountants to perform transaction classification, a core accounting task. It provides a grounded benchmark of human expertise in real-world financial operations and contextualizes the remaining performance gap between state-of-the-art LLMs and human accountants.

To establish this baseline, **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 17,792 transactions dataset utilized in the broader scope of this study. 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.

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⟩}
````

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 o3 which require a temperature of 1.0 per API specifications. [1] [2]
- **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.
- **Domain Expertise Prompting:** All models received the standardized system prompt: `"You are an expert bookkeeper with deep knowledge of accrual-accounting and an eye for detail."`

4.4 Model Providers

We evaluated 19 models with various reasoning configurations from all major providers, including OpenAI, Anthropic, Google, DeepSeek, Moonshot AI, Meta and xAI. The complete model list is shown in Table 1. To standardize deployment for open-source models, we leveraged together.ai and Baseten's infrastructure. For consistency in evaluation, we exclusively used the OpenAI client SDK for all model connections, avoiding provider-specific SDKs.

API rate limits imposed by several providers constrained our ability to evaluate certain models on the full real-world dataset. Although OpenAI's o3-pro model was considered for benchmarking using a smaller subset, daily usage restrictions prevented its evaluation at full scale. Similarly, Google's Gemini 3 Pro was excluded from this study due to comparable access limitations.

4.5 Reasoning Models

For the models with configurable reasoning effort levels, we explored various reasoning effort levels to 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 the purposes of this study, all Charts of Accounts were assumed to be complete and comprehensive. Accordingly, any category proposed by a model that was not present in a client's Chart of Accounts was classified as a true hallucination rather than as evidence of a missing category. This assumption enabled systematic measurement of hallucination rates across models. Furthermore, **all transactions have been reviewed by GAAP accountants for correctness to be confident that the expected ground truth reflects the true category, and that category was already present in the CoA.**

Provider	**Model**
OpenAI	GPT-5.2 (none, low, medium, high) GPT-5 o3-pro (medium) o3 (low, medium, high) o3-mini (medium) GPT-4.1 gpt-oss-120b
Google	Gemini 3 Flash (minimal, high)
Anthropic	Claude Sonnet 4.5 (none, low, high) Claude Haiku 4.5 (none, low, high) Claude Opus 4.5 (none, low, high)
xAI	Grok 3
Meta	Llama 4 Scout Llama 4 Maverick
DeepSeek	DeepSeek-V3.2 DeepSeek-V3 DeepSeek-R1
Alibaba Cloud	Qwen3-235B-A21B (thinking, no thinking)
Moonshot AI	Kimi-K2-Instruct-0905

Table 1: List of evaluated models by provider

4.7 Digits' Proprietary System

While we cannot disclose details about Digits' proprietary ML system, we can share several key characteristics relevant to this comparative study. The Digits ML system 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 workflows.

4.8 Performance Analysis

Our comprehensive evaluation of latest frontier models revealed improved performance across model providers. Top-ranking models generally have higher accuracy and lower latency compared to our 2025 study. The average model accuracy in this study was improved from 61.3% to 65.1% compared to our previous benchmark. This 3.8% improvement indicates overall progress in general-purpose LLMs across major providers.

Notably, Google's Gemini 3 Flash achieved an accuracy rate of 72.2% accuracy at an average of 5.03 sec/inference, compared to the 69.8% accuracy at 6.83 sec/inference from OpenAI's o3 in our previous study. Gemini 3 Flash with low thinking level also emerged as a compelling alternative, achieving 71.6% accuracy (50 basis points below the high thinking level configuration) at a significantly lower latency of 2.2 sec/inference, 128.6% faster than Gemini 3 Flash with high thinking level.

We also documented the very first case of zero hallucination with OpenAI's GPT-5.2 (with low and high reasoning effort) on the 17,792 benchmark dataset, marking a significant milestone for general-purpose LLMs in accounting automation.

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 73% 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 business-specific context. **This missing contextual layer includes business operations patterns, industry-specific accounting practice, and historical bookkeeping preferences for similar transactions.**

Notably, this performance ceiling was overcome by Digits' proprietary ML system which scored 93.5% accuracy on the benchmark dataset of 17,792 transactions. Digits' ML system has been specifically designed to incorporate these additional contextual elements, such as business-specific operational data and historical bookkeeping preferences. Digits system's superior performance

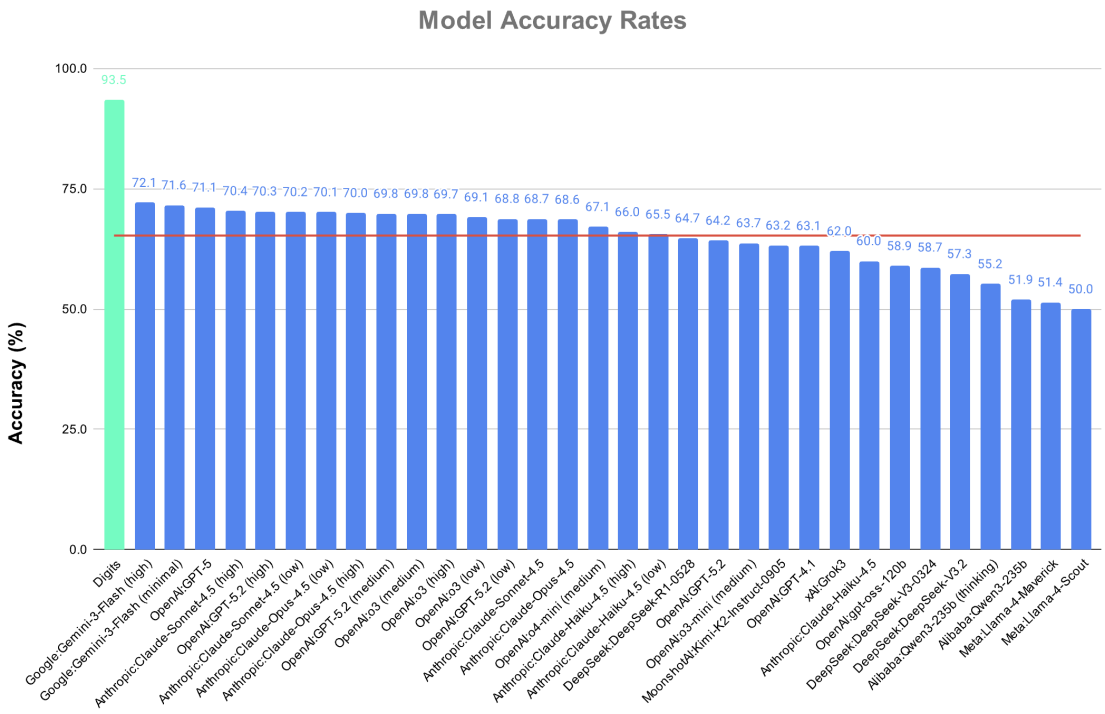

Figure 1: Comparison of Model Accuracy (higher is better)

underscores a crucial finding: **the inherent subjectivity of accounting classification cannot be adequately addressed by even the most advanced general-purpose LLMs in isolation.**

4.8.2 Inference Latency

The latency findings in Figure 2 highlight a critical consideration for organizations implementing AI in accounting workflows: **the trade-off between processing speed and accuracy must be carefully evaluated in the context of business requirements and transaction volumes.** Despite the leading model (Gemini 3 Flash) in this study achieving better accuracy and hallucination rate compared to the forerunner from the previous revision (OpenAI o3), the average latency increased from 5.1 sec/inference to 5.63 sec/inference. The latency increase is primarily driven by LLMs with enhanced reasoning effort levels, which generally yields better accuracy at the expense of inference latency.

The increase in inference latency as reasoning effort dials up stems from three major factors: internal model architecture, the overhead from more reasoning tokens to generate the final output, and more provider-side and client-side timeouts or retries which further extend processing time. In our evaluation code, we instrumented an exponential backoff and retry mechanism for results completeness, which is partially responsible for the high latency. We do acknowledge that in real-world accounting applications, a shorter timeout would be more pragmatic and cost efficient.

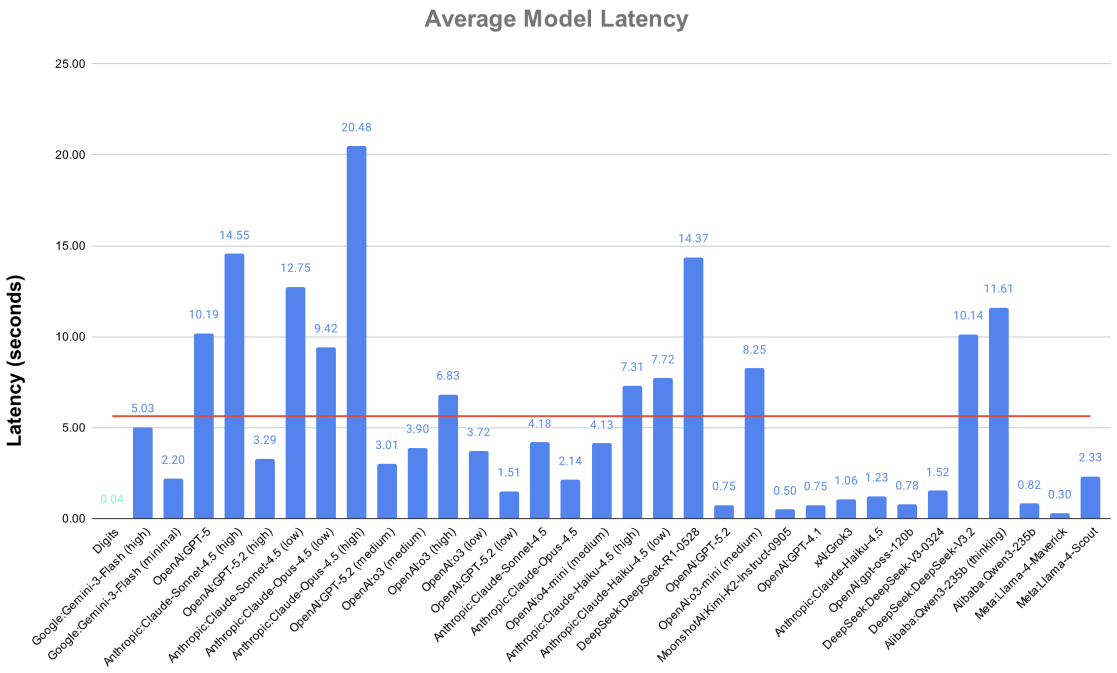

Figure 2: Comparison of Model Latency (smaller is better)

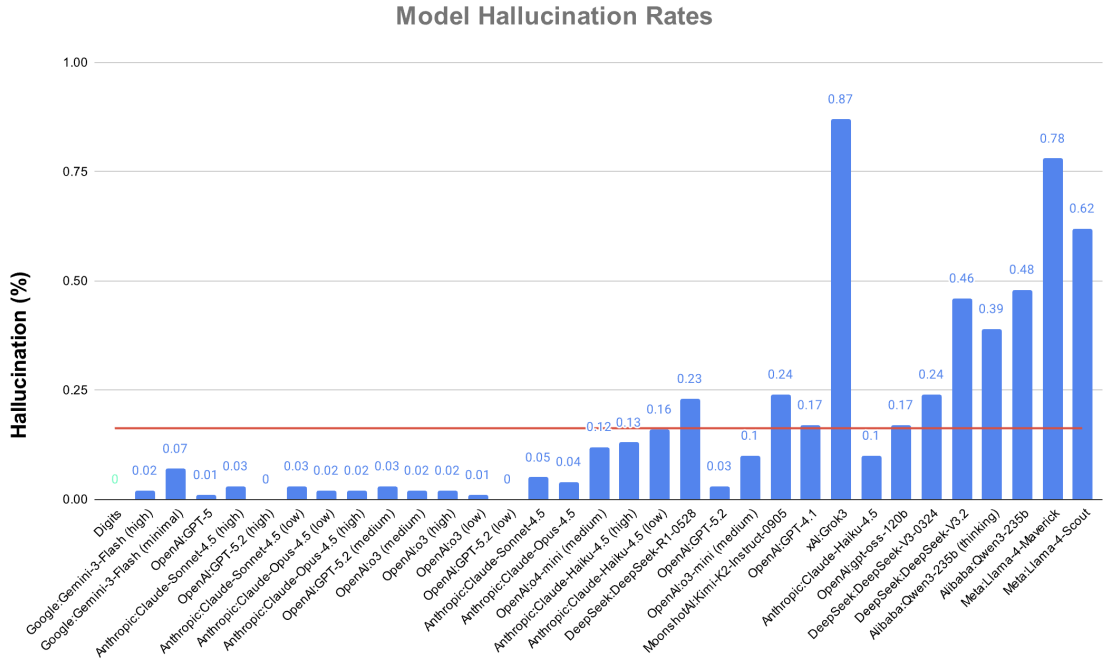

Figure 3: Comparison of Model Hallucination Rates (smaller is better)

4.8.3 Model Hallucination Rates

Our study revealed a strong correlation between hallucination rate and latency, where **we observed a significant reduction in hallucination incidents when no client-side timeout was imposed**. However, this improvement introduces a critical trade-off: the exceedingly long processing time substantially impacts inference throughput, presenting significant operational challenges for accounting systems and other high-volume applications where processing speed is essential for business continuity.

We also found that **while reducing hallucinations is crucial for reliability and user experience, it does not automatically translate to higher overall accuracy**, as shown in Figure 3. Compare GPT-5.2 (high reasoning effort) at 70.3% accuracy to Gemini 3 Flash (minimal thinking level) with 0.07% hallucination and 71.6% accuracy, or GPT-5.2 (none reasoning) at 68.8% accuracy to Claude Opus 4.5 (low, high reasoning effort) with 0.02% hallucination and 70.0%, 70.1% accuracy. Forcing the model to choose in a confined set doesn't guarantee correctness, thus accuracy and hallucination rate should be treated as independent metrics to optimize for in real-world applications.

5 Comparison with Human Baseline

In a controlled study involving 12 compensated professional accountants, the participants were organized into four groups of three accountants each. 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. Each group was responsible for classifying 500 transactions, resulting in a total dataset of 2,000 classified transactions.

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

The results shown in Figure 4 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 accounting tasks. On average, **the accountants took 4 hours and 43 minutes to complete their 500-transaction allocation,** equating to approximately 34 seconds per transaction per accountant.

Note the performance metrics across the human benchmark and full datasets are directionally similar but not exactly the same, though the former was randomly sampled from the larger dataset of 17,792. This characterizes the inherent variability in transaction classification tasks, where the specific composition of the dataset can significantly influence observed performance outcomes. For consistency and comparability, we refer to the 78.8% accuracy of Gemini 3 Flash (high thinking level) as approaching human parity (79.1% accuracy), and otherwise refer to its 72.1% accuracy

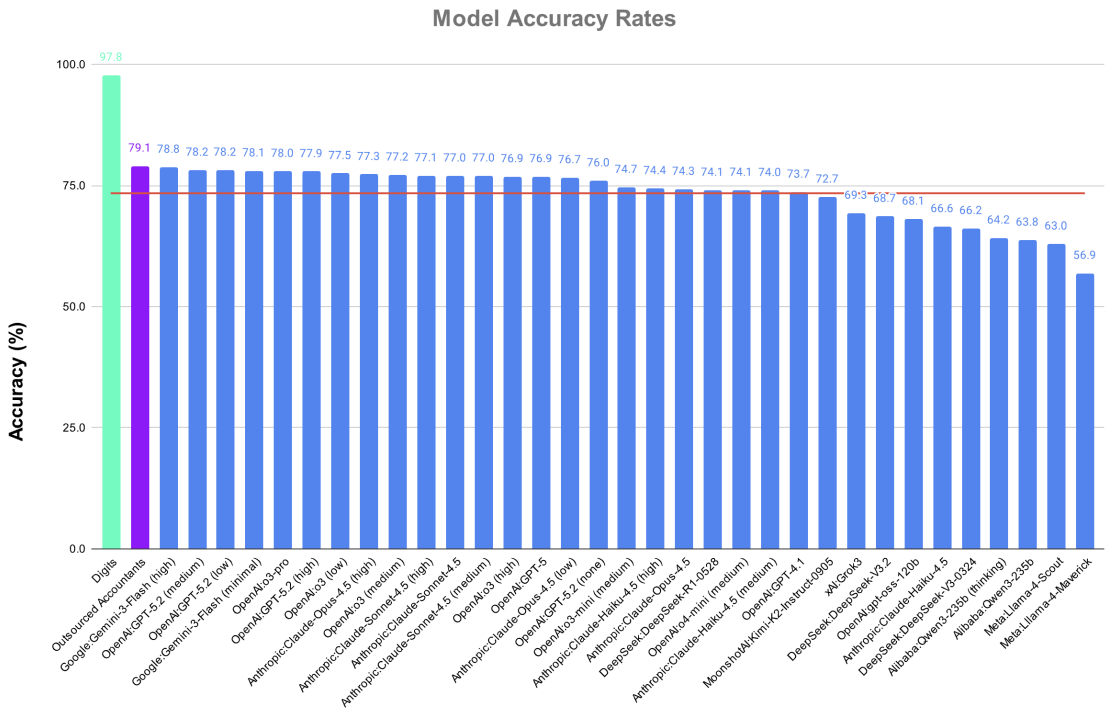

Figure 4: Comparison of Human Accuracy Rates (higher is better)

from the full dataset.

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 interesting performance characteristics when reasoning capabilities were activated. Specifically, our evaluation of OpenAI's GPT-5.2 model demonstrated a near-doubling of inference latency with 1% improvement in accuracy when reasoning effort level increased from low to medium, shown in Table 2; and a more than doubled latency with 0.5% accuracy improvement for Google's Gemini 3 Flash, when the model's thinking depth was updated from minimal to high, see Table 3. Our observations suggest that the computational overhead associated with enhanced reasoning in closed-source architectures may not translate to sustained performance gains in accounting-specific tasks.

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 4. Our results indicated that activating thinking capabilities resulted in a greater than 10-fold increase in latency while yielding

Model Configuration	Accuracy (%)	Latency (s)
OpenAI:GPT-5.2 (high)	70.3	3.29
OpenAI:GPT-5.2 (medium)	69.8	3.01
OpenAI:GPT-5.2 (low)	68.8	1.51
OpenAI:GPT-5.2 (none)	64.2	0.75

Table 2: Performance comparison of OpenAI GPT-5.2 with different reasoning efforts

Model Configuration	Accuracy (%)	Latency (s)
Google:Gemini 3 Flash (high)	72.1	5.03
Google:Gemini 3 Flash (minimal)	71.6	2.20

Table 3: Performance comparison of Google Gemini 3 Flash with different thinking levels

only a modest 3.3% 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-A21B (thinking)	55.2	11.61
Alibaba:Qwen3-235B-A21B (thinking disabled)	51.9	0.82

Table 4: Performance comparison of Qwen3 model with and without reasoning capabilities

Given the high-throughput requirements of modern accounting systems—where processing speed influences operational efficiency and user experience—our findings suggest that reasoning-enhanced models may not consistently provide sufficient benefits to offset their costs. The latency increases observed across both closed- and open-source architectures, combined with relatively modest accuracy gains, suggest that selecting an LLM with an appropriate level of reasoning effort requires careful consideration of the trade-off between accuracy and latency.

7 Challenges and Limitations

Our study encountered several technical challenges that warrant consideration when assessing the feasibility of integrating LLMs into accounting workflows. These issues influenced the benchmarking methodology and raised important considerations regarding the practical deployment of such systems in production environments, where reliability and performance are critical.

API reliability emerged as a consistent challenge across all model providers evaluated in this study. Intermittent failures merits the implementation of robust retry mechanisms to ensure uninterrupted service availability. The reliability and user experience degradation concerns highlight potential risks for accounting applications during peak financial periods, such as month-end or year-end close, when transaction volumes and ledger activity increase substantially.

The substantial gap between the fastest and slowest models in our evaluation indicates that

processing speed is a critical consideration when selecting AI technologies for accounting applications, and in many cases may carry greater practical importance than marginal accuracy improvements.

These technical limitations underscore the importance of evaluating not only the classification accuracy of LLMs for accounting tasks, but also their operational characteristics under realistic deployment conditions. Organizations considering such technologies should assess whether infrastructure requirements, reliability patterns, and response times align with their accounting workflows and user experience expectations.

8 Future Directions

This paper presents the third revision of our industry-recognized study examining general-purpose LLMs for accounting automation. We acknowledge that emerging innovations may effectively address several challenges identified in our current evaluation framework—particularly processing speed, hallucination rates, and domain-specific understanding of accounting principles.

Now that our benchmark work established that the state-of-the-art models are approaching human accuracy, future research will focus on evaluating models' ability to solve more complex accounting tasks when they are provided the tools to access business context and historical patterns.

We are committed to helping accounting professionals make informed decisions as AI technologies mature. We will continue to update our benchmark as new frontier models are released, ensuring that our findings remain relevant and actionable for practitioners in the field.

9 Conclusion

We set out to examine the impact of frontier large language models (LLMs) on accounting automation by conducting a comprehensive study of 19 state-of-the-art general-purpose LLMs on a transaction classification task. Our findings reveal a consistent performance ceiling across these models, suggesting that general world knowledge alone is insufficient to fully automate accounting tasks.

Performance Ceiling and Contextual Limitations The observed accuracy ceiling of 73% suggests that general-purpose LLMs lack the necessary contextual understanding of business-specific patterns, industry practices, and historical bookkeeping preferences that are essential for accurate accounting classification. This limitation underscores the importance of domain-specific context, which general-purpose models are not designed to capture effectively.

Operational Viability Concerns Our analysis reveals substantial operational challenges confronting LLMs's effectiveness in high-volume accounting applications. The average request latency of 5.63 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.

Limited Value of Reasoning Models Increasing reasoning effort in models did not produce

sustained performance gains for accounting tasks, despite common intuition. Across both closed- and open-source architectures, high reasoning capabilities in most cases delivered marginal accuracy gains at the cost of increased latency and computational expense. These results suggest that LLMs' performance ceilings are driven less by reasoning depth than the availability of domain-specific context.

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. Additionally, the 10.4% disagreement rate among professional accountants on identical transactions highlights the inherent subjectivity of accounting classification, underscoring the value of systems that capture and apply business-specific patterns rather than relying solely on generalized knowledge.

References

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

[2] 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.