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

Hannes Hapke, Cole Howard, Jo Pu, Chris Hassell

ml@digits.com

Digits Financial, Inc.

Last revision: March 4, 2025

1 Executive Summary

1.1 Key Findings and Implications

- The best-performing general-purpose LLM (GPT-4.5-preview) reached an accuracy ceiling of approximately 66% for transaction classification, while Digits' purpose-built system achieved over 93% accuracy, highlighting the limitations of general AI for specialized accounting tasks.
- Newer model iterations demonstrated unexpectedly higher hallucination rates, with models like Claude 3.7 and OpenAI o3 returning unparsable JSON or suggesting non-existent accounting categories more frequently than their predecessors.
- Processing latency varied significantly across models (3.7 seconds on average), with reasoning models showing the slowest response times without proportional accuracy improvements, creating scalability challenges for high-volume accounting workflows.
- API reliability issues and throughput limitations were observed across all model providers, raising concerns about the operational viability of LLM integrations in time-sensitive accounting applications.

1.2 Methodology Overview

- We evaluated 17,792 financial transactions from over 100 randomly selected small businesses, providing consistent prompts across 13 different models from major providers, including OpenAI, Anthropic, Google, and xAI, comparing their performance against a specialized accounting ML system.

2 Introduction

In today's rapidly evolving financial technology landscape, applying artificial intelligence to accounting tasks presents promising opportunities and significant challenges. 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.

This white paper examines the performance of leading AI models in handling accounting transaction classification tasks, comparing Digits' proprietary, specialized ML system against general-purpose language models across multiple dimensions, including accuracy, latency, and reliability.

While recent advances in language model capabilities have demonstrated impressive results across many domains, accounting's inherent subjectivity and domain-specific requirements raise questions about whether general-purpose models can effectively address the nuanced needs of financial classification tasks. Our comprehensive evaluation provides actionable insights into these questions.

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-4 and other similar 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 Digits clients.** 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 client 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 clients, with the number of categories in their Charts of Accounts ranging from as few as 15 to as many as 281 categories. The average client maintained 92 categories, while the median was 66, indicating a right-skewed distribution where some clients 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 Prompt Construction

The primary objective of our evaluation was to predict the appropriate category from each client'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 accounting client 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 client-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 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.

4.2.1 Prompt Example

Here is an example of the prompt we used for the LLM comparison is shown in Prompt 1.

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⟩}
```

4.2.2 Experimental Configuration 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 128 tokens for standard models, with an extended limit of 512 tokens for reasoning-focused models to accommodate their explanatory capabilities.
- **Domain Expertise Prompting:** All models except OpenAI's o1/o3 and DeepSeek's R1/V3 received the standardized system prompt: `"You are an expert bookkeeper with deep knowledge of accrual-accounting and an eye for detail."`

4.3 Model Providers

Our evaluation encompassed models from all major providers, including OpenAI, Anthropic, xAI, and Google, with model access occurring between February 17 and February 28, 2025. For open-source models, we utilized together.ai's infrastructure and SDK together.ai [2025] to standardize deployment conditions.

To ensure consistency in our evaluation methodology, we employed Andrew Ng's `aisuite` library Ng [2025] to manage transitions between different model providers, which helped maintain procedural uniformity throughout the testing process.

For this comprehensive study, we included the following models:

Provider	Model
OpenAI	o1 o1-mini o3-mini gpt-4o gpt-4o-mini gpt-4.5o-preview
Google	gemini-2.0-flash
Anthropic	claude-3-5-sonnet-20240620 claude-3-7-sonnet-20250219
xAI	grok-2-1212[1]
Meta	Llama-3.3-70B-Instruct-Turbo
Deepseek	DeepSeek-V3 DeepSeek-R1

Table 1: List of evaluated models by provider

4.4 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

[1]At the time of writing, xAI hasn't made the Grok-3 API generally available yet.

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.**

4.5 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.

4.6 Performance Analysis

Our comprehensive evaluation revealed several significant patterns in model performance across the transaction classification task. The most recently released GPT-4.5-preview emerged as the top-performing general-purpose model, demonstrating a superior ability to interpret transaction descriptions and match them to appropriate accounting categories compared to other tested models.

4.6.1 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 human 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.**

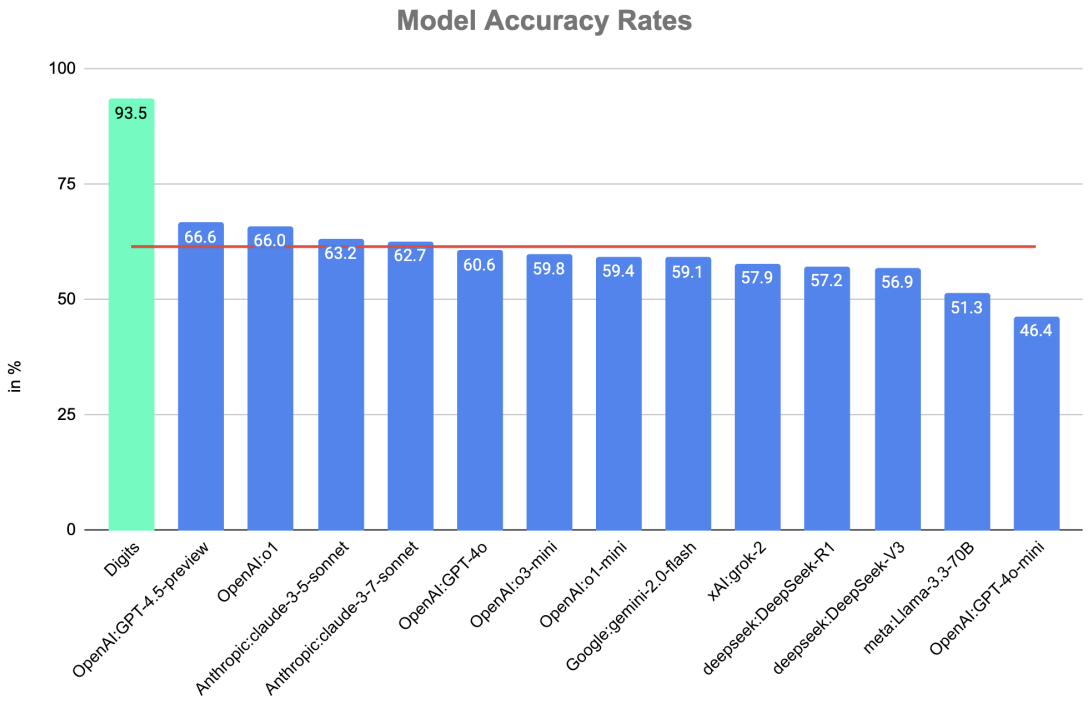

Figure 1: Comparison of Model Accuracy (higher is better)

In addition to the subjectivity, large language models lack the ability to 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 accountants.

4.6.2 Model Request Latency

Our analysis revealed substantial variation in response times across the evaluated models, with reasoning-focused models such as OpenAI's o1 and o3 consistently demonstrating the highest latencies.

These reasoning models exhibited significantly slower processing times than their non-reasoning counterparts (e.g., GPT-4.5), a performance characteristic that aligns with their architectural design priorities. These models are fundamentally optimized for deep reasoning on complex individual requests rather than high-throughput processing of numerous similar tasks.

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 was 3.67 seconds per request, 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 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.

4.6.3 Model Hallucination Rates

Our analysis uncovered a concerning trend in the hallucination rates of newer model iterations shown in Figure 3. Contrary to the general expectation that more recent models would demonstrate improved reliability, **we observed significantly higher hallucination rates in the latest models, such as Claude 3.7 and OpenAI o3, than their predecessors.** These newer models more frequently suggested accounting categories that did not exist within the client's Chart of Accounts, requiring additional validation and correction steps that would diminish the efficiency gains sought through automation.

Reviewing the hallucinated results revealed that recent models slightly altered the predicted category name. For example, an expense category, "Advertising" would turn into `"Advertising (expensed)"` or the bank category, `"Mercury Checking - 1"` would turn into `"Mercury Checking"`.

The practical implications of these hallucination rates are substantial when considered at scale. Even a modest 1% hallucination rate translates to 10,000 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

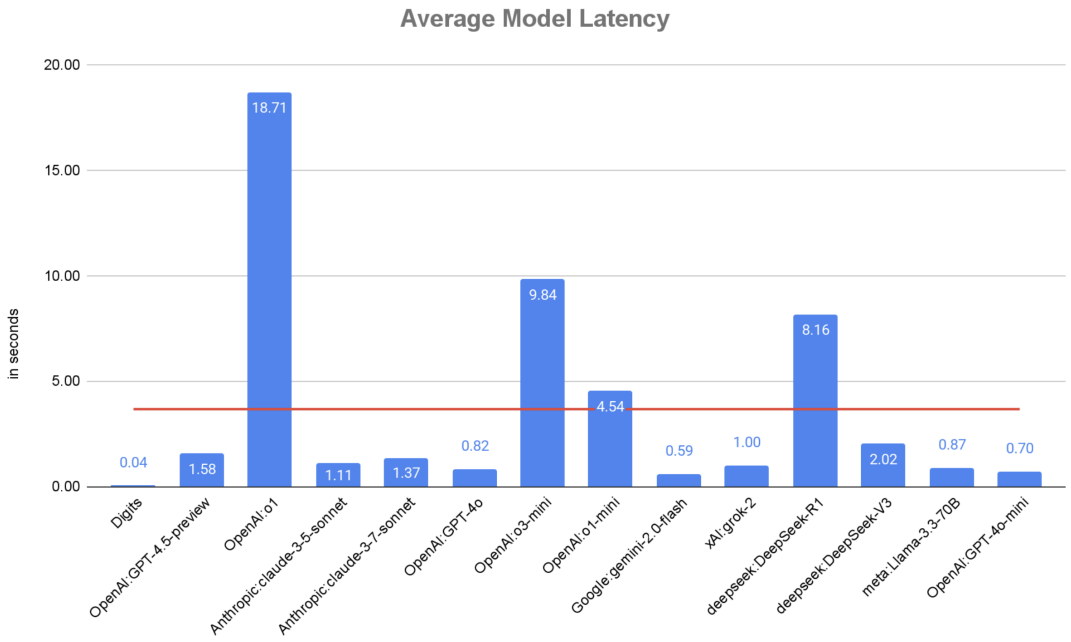

Figure 2: Comparison of Model Latency (smaller is better)

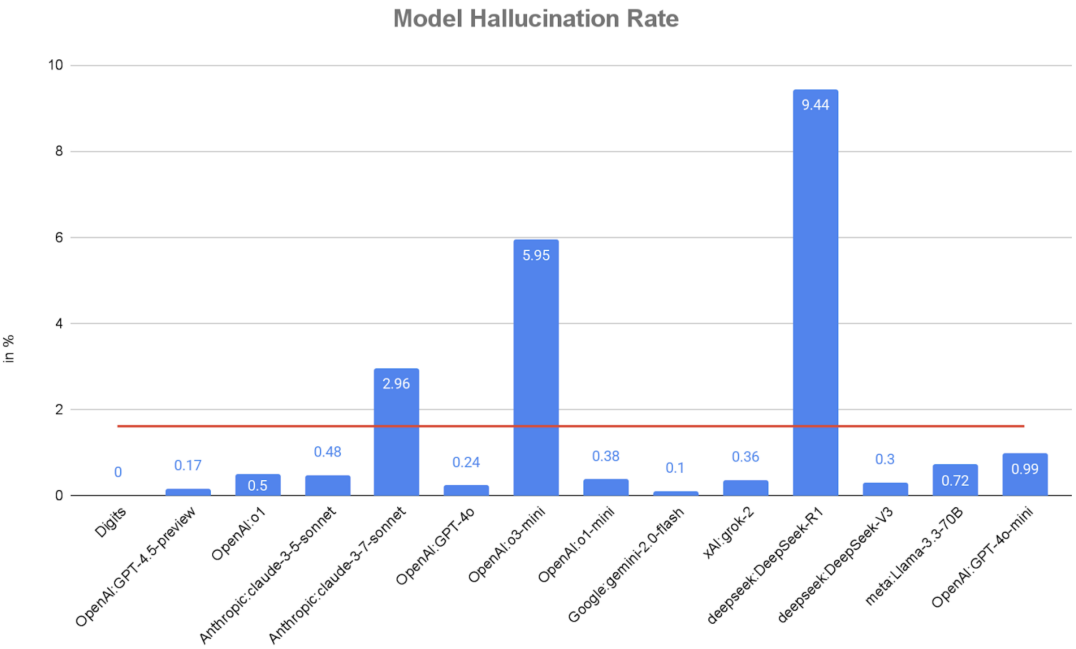

Figure 3: Comparison of Model Hallucination Rates (smaller is better)

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 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 quarter-end closings when transaction volumes spike significantly.

We observed substantial variations in throughput capabilities among different providers. Most notably, xAI's infrastructure imposed strict limitations on request parallelization, forcing us to process transactions sequentially rather than in parallel batches. This constraint significantly extended testing timelines and would present serious scalability challenges in production environments handling large transaction volumes. Similarly, several other providers required aggressive request throttling to prevent rate limit exceptions, extending processing times beyond acceptable thresholds in time-sensitive accounting workflows.

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.

6 Future Directions

Our exploration of LLM performance in accounting tasks represents an initial step in understanding the potential and limitations of these technologies in financial classification. As the artificial intelligence landscape continues to evolve rapidly, we are committed to maintaining an ongoing evaluation process that tracks developments across the industry.

We recognize that model capabilities are advancing quickly, with new architectures, training methodologies, and specialized fine-tuning approaches emerging regularly. These innovations may address some of the challenges we've identified in our current evaluation, particularly regarding processing speed, hallucination rates, and domain-specific understanding of accounting principles.

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.

7 Conclusion

Our comprehensive evaluation of LLMs for accounting transaction classification reveals critical insights for organizations seeking to implement AI-powered financial automation. The consistent performance ceiling observed across even the most advanced general-purpose language models highlights the fundamental challenges accounting's inherent subjectivity presents to AI systems.

Despite recent advances in model capabilities, the contextual understanding and domain-specific judgment required for accurate accounting classification remain beyond the reach of models designed for general applications.

These findings suggest that the future of accounting automation lies not in applying increasingly powerful general-purpose models but in developing specialized systems specifically for financial classification tasks that can adapt to the subjective classification patterns unique to each business context.

The significant performance gap between general-purpose LLMs and purpose-built accounting systems like Digits' proprietary solution demonstrates the value of domain-specific optimization. Organizations seeking to implement accounting automation should carefully evaluate the accuracy metrics of potential solutions and their operational characteristics, including processing latency, hallucination rates, and reliability patterns.

As AI technology continues to evolve, the most successful accounting automation approaches will likely combine domain-specific models with carefully designed workflows that accommodate the unique classification needs of individual businesses while maintaining the rigorous standards of accuracy and reliability that financial processes demand.

References

Andrew Ng. Andrew ng's ai suite, 2025. URL https://github.com/andrewyng/aisuite.

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

together.ai. together.ai sdk, 2025. URL https://docs.together.ai/docs/introduction.

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.