Artificial Intelligence and Law
https://doi.org/10.1007/s10506-025-09465-7

ORIGINAL RESEARCH

Performance of LLMs on VITA test: potential for AI-assisted tax returns for low income taxpayers

Sina Gogani-Khiabani[1] · Ashutosh Trivedi[2] · ShinPing Chyi[3] · Saeid Tizpaz-Niari[1]

Accepted: 9 May 2025

Abstract

This paper investigates the performance of a diverse set of large language models (LLMs) including leading closed-source (GPT-4, GPT-4o mini, Claude 3.5 Haiku) and open-source (Llama 3.1 70B, Llama 3.1 8B) models, alongside the earlier GPT-3.5 within the context of U.S. tax resolutions. AI-driven solutions like these have made substantial inroads into legal-critical systems with significant socio-economic implications. However, their accuracy and reliability have not been assessed in some legal domains, such as tax. Using the Volunteer Income Tax Assistance (VITA) certification tests—endorsed by the US Internal Revenue Service (IRS) for tax volunteering—this study compares these LLMs to evaluate their potential utility in assisting both tax volunteers as well as taxpayers, particularly those with low and moderate income. Since the answers to these questions are not publicly available, we first analyze 130 questions with the tax domain experts and develop the ground truths for each question. We then benchmarked these diverse LLMs against the ground truths using both the original VITA questions and syntactically perturbed versions (a total of 390 questions) to assess genuine understanding versus memorization/hallucinations. Our comparative analysis reveals distinct performance differences: closed-source models (GPT-4, Claude 3.5 Haiku, GPT-4o mini) generally demonstrated higher accuracy and robustness compared to GPT-3.5 and the open-source Llama models. For instance, on basic multiple-choice questions, top models like GPT-4 and Claude 3.5 Haiku achieved 83.33% accuracy, surpassing GPT-3.5 (54.17%) and the open-source Llama 3.1 8B (50.00%). These findings generally hold across both original and perturbed questions. However, the paper acknowledges that these developments are initial indicators, and further research is necessary to fully understand the implications of deploying LLMs in this domain. A critical limitation observed across all evaluated models was significant difficulty with open-ended questions, which require accurate numerical calculation and application of tax rules. We hope that this paper provides a means and a standard to evaluate the efficacy of current and future LLMs in the tax domain.

Extended author information available on the last page of the article

Published online: 08 July 2025

Keywords Large language models · Tax law · VITA certification tests

1 Introduction

Advancements in AI continue to push boundaries of AI-human collaborative problem-solving in a wide range of applications and tasks ranging from finance and manufacturing to healthcare. The capabilities in solving challenging socio-technological problems in high-dimensions regardless of input data type (e.g. code, images, text, audio) has enabled widespread adoption of applications in medical domain (Nori et al. 2023), planning (Webb et al. 2024), tax preparation (Yu et al. 2020; April Technologies 2023; Gogani-Khiabani et al. 2025; Tizpaz-Niari et al. 2023; Srinivas et al. 2023), and finance/banking (Drenik 2023).

While large language models (LLMs) have made substantial impacts across different domains, the applicability of this technology to answer tax-related questions requires further investigation (Tizpaz-Niari et al. 2024). Specifically, the complexity of tax law combined with the requirements for sophisticated numerical reasoning present challenges for LLMs in this domain. Simultaneously, advancements in AI models' capabilities, continues to make exceptional progress. However, the performance and accuracy of LLMs for tax-relevant scenarios and questions have not been studied systematically.

This paper aims to address this gap by assessing the performance of several LLMs, including both closed-source and open-source models, on a set of tax-related questions. We evaluate these LLMs using Volunteer Income Tax Assistance (VITA) certification tests (Internal Revenue Service 2022), employed by the U.S. Internal Revenue Service (IRS) to certify tax volunteers who assist low/moderate income taxpayers. These tests present various tax scenarios at both basic and advanced levels. We initially compiled a dataset of 130 questions from the IRS Form 6744 (Internal Revenue Service 2022) and subsequently generated two distinct sets of syntactic perturbations for these scenarios and questions, resulting in 130 original and 260 perturbed instances (a total of 390). In performing the syntactic perturbations, we ensure that questions remain the same semantically, so the answers are the same.

While the answers to these questions are not publicly available, we recognize the possibility that some LLMs might have encountered similar information during training. Therefore, the inclusion of perturbed questions is crucial for evaluating true understanding versus memorization. We employed prompting as the primary method of interaction with the LLMs, presenting scenarios and related questions without additional context. Ground truth answers were obtained through a consensus process involving domain experts.

The primary method of interacting with LLMs is prompting, where we present the scenarios (separately starting from an empty chat for each scenario), followed by the questions related to the scenario, with no other contexts. To obtain the ground truth answers, we distribute questions to two tax experts and schedule a meeting to discuss their answers and resolve any disagreements. The experts reached agreements on all questions except two. In those cases, we asked a third expert and resolved the disagreements by taking the majority vote

This study evaluates the efficacy of this diverse set of LLMs-including GPT-4, GPT-3.5, GPT-4o mini, Claude 3.5 Haiku, Llama 3.1 70B, and Llama 3.1 8B-across the 390 original and syntactically perturbed questions. The evaluation covers two primary VITA scenario types (basic and advanced) and three distinct question formats (true/false, multiple-choice, and open-ended) to provide a comprehensive performance benchmark. Our analysis reveals significant variations in capability. Leading closed-source models generally demonstrated superior accuracy; for instance, on advanced true/false questions, GPT-4 and Claude 3.5 Haiku both achieved high accuracy (82.76%), significantly outpacing GPT-3.5 (55.17%) and Llama 3.1 8B (55.17%). Model scale also proved influential within the open-source models, with Llama 3.1 70B (e.g., 72.41% on advanced true/false) consistently performing better than Llama 3.1 8B (55.17% on the same task). Robustness to syntactic perturbations varied, with some models showing performance gains (e.g., GPT-4 on perturbed basic true/false: 90.63% vs 65.63% original) while others decreased, suggesting complex reasoning patterns rather than simple memorization. A critical finding across all models, however, was a pronounced difficulty with open-ended questions requiring precise numerical calculation and tax rule application; even the top-performing model, GPT-4, achieved only 48.89% accuracy overall on these questions (22 out of 45), while many others, like Llama 3.1 8B, scored below 10%. We pose the challenge of improving performance on numerically-focused, open-ended questions as crucial for future research aiming to enhance LLMs' applicability in the tax domain.

The rest of this paper is structured as follows: Section 2 describes the LLMs evaluated, data collection and perturbation methods, prompting strategies, and the ground truth generation process. Section 3 presents the results of our experiments. Section 4 discusses the findings, including statistical significance, the RAG experiment, and the effects of perturbations. Section 5 overviews related work on evaluating LLMs, particularly in specialized domains. Section 6 concludes the paper with a summary, discussion of limitations, and directions for future research.

2 Methodology

2.1 Large language models

This study evaluates a range of LLMs, including both closed-source and open-source models, to provide a broader assessment of their capabilities within the tax domain.

GPT-3.5 (OpenAI): Launched prior to GPT-4, GPT-3.5 is based on the GPT-3 architecture (Brown et al. 2020). It features a sophisticated design capable of generating human-like text by predicting subsequent tokens. GPT-3.5 has demonstrated proficiency in a wide range of tasks, from conversational agents to content creation, yet it encounters challenges with complex or domain-specific queries (Floridi and Chiriatti 2020).
GPT-4 (OpenAI): GPT-4, the successor to GPT-3.5, introduces improvements in understanding, reasoning, and contextual awareness due to an expanded training corpus and refinements in architecture and training approaches (OpenAI et al.

2024; Wei et al. 2022). Consequently, GPT-4 demonstrates superior performance, particularly in interpreting intricate scenarios and generating contextually relevant and coherent text (Bubeck et al. 2023).
Llama 3.1 8B (Meta): This is an open-source LLM developed by Meta. It offers a smaller parameter size (8 billion) compared to larger models like GPT-4, making it potentially more accessible for research and deployment in resource-constrained environments. Its performance on complex tasks like tax-related question answering is a key aspect of our evaluation.
Llama 3.1 70B (Meta): Also developed by Meta, Llama 3.1 70B offers a larger parameter size (70 billion) compared to the 8B version. This increased scale often translates to improved performance on complex tasks. Our study investigates whether this holds true for tax-related questions and how its performance compares to both closed-source and smaller open-source models.
GPT-4o mini: This smaller closed-source model allows us to explore the trade-off between model size and performance in the tax domain. Its reduced computational requirements make it potentially suitable for deployment on devices with limited resources.
Claude 3.5 Haiku (Anthropic): Developed by Anthropic, Claude 3.5 Haiku is another smaller closed-source model included in our evaluation to further investigate the impact of model size on performance in the context of tax-related questions.
Summary: Including this diverse set of LLMs allows us to explore a broader spectrum of model capabilities, comparing performance across different architectures, sizes, and access levels (open-source vs. closed-source). Specifically, our study examines their effectiveness in navigating the complexities of tax regulations, evaluating their potential to automate and augment tasks traditionally performed by human tax experts. This focus is pertinent given the intricate nature of tax law and the potential benefits of leveraging LLMs to assist both taxpayers and tax professionals.

2.2 Input sources

Our analysis is derived from the Internal Revenue Service (IRS) Volunteer Income Tax Assistance (VITA) certification tests, as outlined in the IRS Form 6744 (Internal Revenue Service 2022). Specifically, we utilize a dataset that consists of 60 basic and 70 advanced questions included in this publication. These questions are designed to assess the comprehension and application of tax law for volunteers who wish to assist taxpayers, particularly those with low to moderate income.

It is important to note that the answers to these questions are not publicly available, adding a layer of complexity to the task of evaluating the performance of large language models in this domain. The chosen scenarios are deliberately focused on taxpayers under specific circumstances, emphasizing scenarios tailored towards low-income taxpayers. This approach not only aligns with the practical applications of these models in assisting vulnerable communities, but also presents a unique challenge in assessing the models' ability to navigate the intricacies of tax law.

2.3 Syntactic perturbations

To evaluate the resilience and adaptability of large language models to variations in input, we introduced syntactic perturbations to the original question scenarios. This process involved the systematic alteration of question structures without changing their inherent meaning or the information they conveyed.

2.3.1 Perturbation methodology with examples

We employed a range of syntactic manipulations, such as reordering of clauses, and the substitution of synonyms. Additionally, we included more complex syntactic changes, such as rephrasing questions as statements and vice versa. These modifications aim to reflect natural language variations, offering a comprehensive assessment of the models' comprehension abilities. The perturbed scenarios were then used to test the models, allowing us to gauge their performance consistency across syntactically varied yet semantically equivalent inputs. The perturbations included:

- *Changing personal names*: Replacing names in the scenarios while maintaining the other details.
- *Switching genders (if applicable)*: Altering the gender of individuals in the scenarios where it does not impact tax implications.
- *Modifying ages slightly within 2 years*: Adjusting ages within a small range, ensuring these changes do not have tax implication (e.g., changing 47 to 48 is allowed, but modifying 16 to 18 is not allowed).
- *Paraphrasing sentences*: Rephrasing sentences while preserving their original meaning and tax implications.
- *Maintaining monetary values*: Crucially, all monetary values, including wages, interest, tax credits, etc., were kept *exactly* the same to ensure the perturbations did not alter the correct tax calculations.
- *Preserving critical information*: No critical tax-relevant information was added or removed during the perturbation process.

These syntactic modifications create semantically equivalent variations of the original questions to assess LLM's capabilities in resolving the tax returns. The perturbed and original scenarios are used to assess the LLMs' ability to interpret tax-related information despite variations in phrasing or presentation consistently. For each of the 130 original questions/scenarios, we generated two distinct sets of perturbations following these rules, resulting in 260 perturbed instances (a total of 390). By adhering to these specific perturbation rules, we ensured that the difficulty and underlying tax logic of each question remained unchanged. This approach is aligned with the goal of evaluating true comprehension and reasoning capabilities rather than simply assessing the models' ability to recall specific memorized answers.

Illustrative Example Below, we present an original scenario and its perturbed version, along with the associated question to illustrate our approach. This scenario is

adapted from the VITA/TCE certification tests as outlined in Form 6744 (Internal Revenue Service 2022, p. 71).

Original Scenario Jenny Smith, age 57, is single. Jenny earned wages of $52,000 and was enrolled the entire year in a high deductible health plan (HDHP) with self-only coverage. During the year, Jenny contributed $2,000 to her Health Savings Account (HSA) and her mother also contributed $1,000 to Jenny's HSA account.

Original Question 6. Form 8889, Part 1 is used to report HSA contributions made by _____.

a. Jenny
b. Jenny's employer
c. Jenny's mother
d. All of the above

Perturbed Scenario At 57 years of age and single, Laura Johnson earned a salary of $52,000. For the full year, she was covered by a high deductible health plan (HDHP) with just herself. Laura added $2,000 to her Health Savings Account (HSA) over the year, and her father contributed an additional $1,000.

Perturbed Question 6. Form 8889, Part 1 is used to report HSA contributions made by _____.

a. Laura
c. Laura's father
b. Laura's employer
d. All of the above

2.3.2 Rationale and observed effects of perturbation

The introduction of syntactic perturbations is a well-established method in natural language processing research to test whether LLM responses are based on rote memorization or genuine reasoning. Prior studies, such as Ribeiro et al. (2020) and Jiang et al. (2020), have demonstrated the utility of perturbations to evaluate model robustness and understanding across semantically equivalent but syntactically varied inputs. Similarly, Brown et al. (2020) emphasize that performance on rephrased prompts indicates reasoning ability rather than reliance on memorized patterns. By preserving semantic content while altering syntactic structures, we simulate real-world linguistic variability and challenge the models to apply knowledge flexibly. The success of GPT-4 on the perturbed dataset in this study further aligns with findings from Laban et al. (2023).

2.3.2.1 Observations Our analysis across all evaluated models revealed that syntactic perturbations did not yield a uniform effect on performance. Contrary to what

might be expected if models relied heavily on memorization (where performance would likely decrease consistently), we observed varied responses. For some model-scenario-question type combinations, accuracy increased with perturbation (e.g., GPT-4's accuracy on basic true/false questions rose from 65.63% to 90.63%), while for others, it decreased (e.g., Llama 3.1 70B's accuracy on basic multiple-choice questions fell from 67.00% to 60.04%). This inconsistency across the different LLMs may suggest that the models generally engage in reasoning rather than simply recalling memorized answers when faced with linguistic variations. However, the variability may also indicate that their reasoning and comprehension mechanisms handle these variations with differing degrees of success. We left further research to understand the root cause of these inconsistencies for future work.

2.4 Prompting strategy

Interacting with large language models (LLMs) like the LLMs evaluated in this study involves a method known as "prompting". This technique, foundational in natural language processing (NLP), entails presenting the LLM with a scenario followed by a series of prompts (questions), to which the model generates responses based on its pre-existing knowledge and the provided context. Our methodology for prompting was systematic and designed to ensure fair evaluation and mimic realistic usage patterns. We employed a zero-shot prompting approach, meaning no examples of correctly answered questions were included within the prompt itself. For each distinct tax scenario taken from the VITA/TCE certification tests (Internal Revenue Service 2022), a new, independent chat session was initiated with the LLM. This crucial step prevents any potential carry-over of context or information from previous interactions, and it ensures that each scenario is processed based only on the information presented in that specific prompt, akin to how a user might pose distinct questions. The structure of a typical prompt consisted of directly presenting the full text of the tax scenario, followed immediately by the specific question associated with that scenario. Furthermore, to enhance accuracy and gain insight into the model's reasoning process, we generally encouraged the models to articulate their reasoning or provide a step-by-step derivation before presenting the final answer. This approach aligns with the principles of Chain-of-Thought (CoT) prompting (Wei et al. 2023), which have been shown to improve performance on complex reasoning tasks.

The complete set of interactions with the large language models, including detailed prompts and responses, is compiled in an Excel files available here.

2.5 Adjudication

Since the answers to the questions within the IRS VITA certification tests are not publicly available, we use a 2-step process to develop the ground truth answers. First, the third and fourth authors, who are a Certified Public Accountant (CPA) and an IRS-certified tax preparer, independently answer all 130 questions from the IRS Form 6744. Then, they meet during one session for 3 h to resolve and agree on all the answers. At the end of the session, they agreed on all the answers except for two

Table 1 GPT-3.5 Performance results

Scenario	Question type	Total questions	Correct answers	Accuracy (%)
Basic	True/False	32	24	75.00
Basic	Multiple Choice	24	13	54.17
Basic	Open-Ended	4	0	0.00
Advanced	True/False	29	16	55.17
Advanced	Multiple Choice	30	12	40.00
Advanced	Open-Ended	11	4	36.36
P-Basic	True/False	64	45	70.31
P-Basic	Multiple Choice	48	24	50.00
P-Basic	Open-Ended	8	0	0.00
P-Advanced	True/False	58	29	50.00
P-Advanced	Multiple Choice	60	34	56.67
P-Advanced	Open-Ended	22	8	36.36

Table 2 GPT-4 Performance Results

Scenario	Question type	Total questions	Correct answers	Accuracy (%)
Basic	True/False	32	21	65.63
Basic	Multiple Choice	24	20	83.33
Basic	Open-Ended	4	1	25.00
Advanced	True/False	29	24	82.76
Advanced	Multiple Choice	30	18	60.00
Advanced	Open-Ended	11	5	45.45
P-Basic	True/False	64	58	90.63
P-Basic	Multiple Choice	48	32	66.67
P-Basic	Open-Ended	8	4	50.00
P-Advanced	True/False	58	42	72.41
P-Advanced	Multiple Choice	60	38	63.33
P-Advanced	Open-Ended	22	12	54.55

questions. To resolve the disagreement on the two remaining questions, we seek a consultant from a non-profit organization who could confirm the answers from the CPA.

3 Results

Our exploration of the capabilities of Large Language Models (LLMs) encompasses a diverse set of models, including those from OpenAI (GPT-3.5, GPT-4, GPT-4o mini), Meta (Llama 3.1 8B, Llama 3.1 70B), and Anthropic (Claude 3.5 Haiku). We evaluated these models across four structured tax scenarios-basic, advanced, and their respective variations with syntactic modifications, called perturbed basic (P-Perturbed) and perturbed advanced (P-Advanced). Each scenario comprises questions categorized as true/false, multiple choice, and open-ended. Detailed results for each model across all scenarios and question types are presented in Tables 1, 2, 3, 4, 5 and 6. Comprehensive visual summaries of these results are provided via heatmaps

Table 3 Llama 3.1 8B performance results

Scenario	Question type	Total questions	Correct answers	Accuracy (%)
Basic	True/False	32	20	62.50
Basic	Multiple Choice	24	12	50.00
Basic	Open-Ended	4	0	0.00
Advanced	True/False	29	16	55.17
Advanced	Multiple Choice	30	13	43.33
Advanced	Open-Ended	11	1	9.09
P-Basic	True/False	64	33	51.56
P-Basic	Multiple Choice	48	26	54.17
P-Basic	Open-Ended	8	0	0.00
P-Advanced	True/False	58	32	55.17
P-Advanced	Multiple Choice	60	24	40.00
P-Advanced	Open-Ended	22	1	4.55

Table 4 Llama 3.1 70B performance results

Scenario	Question type	Total questions	Correct answers	Accuracy (%)
Basic	True/False	32	19	59.37
Basic	Multiple Choice	24	16	67.00
Basic	Open-Ended	4	0	0
Advanced	True/False	29	21	72.41
Advanced	Multiple Choice	30	19	63.33
Advanced	Open-Ended	11	2	18.18
P-Basic	True/False	64	43	67.18
P-Basic	Multiple Choice	48	29	60.04
P-Basic	Open-Ended	8	0	0.00
P-Advanced	True/False	58	37	63.79
P-Advanced	Multiple Choice	60	36	60.00
P-Advanced	Open-Ended	22	5	22.73

Table 5 GPT-4o mini performance results

Scenario	Question type	Total questions	Correct answers	Accuracy (%)
Basic	True/False	32	26	81.25
Basic	Multiple Choice	24	18	75.00
Basic	Open-Ended	4	0	0.00
Advanced	True/False	29	23	79.31
Advanced	Multiple Choice	30	18	60.00
Advanced	Open-Ended	11	2	18.18
P-Basic	True/False	64	49	76.56
P-Basic	Multiple Choice	48	32	66.67
P-Basic	Open-Ended	8	0	0.00
P-Advanced	True/False	58	46	79.31
P-Advanced	Multiple Choice	60	37	61.67
P-Advanced	Open-Ended	22	5	22.73

Table 6 Claude 3.5 Haiku performance results

Scenario	Question type	Total questions	Correct answers	Accuracy(%)
Basic	True/False	32	22	68.75
Basic	Multiple Choice	24	20	83.33
Basic	Open-Ended	4	0	0.00
Advanced	True/False	29	24	82.76
Advanced	Multiple Choice	30	20	66.67
Advanced	Open-Ended	11	4	36.36
P-Basic	True/False	64	53	82.81
P-Basic	Multiple Choice	48	30	62.50
P-Basic	Open-Ended	8	0	0.00
P-Advanced	True/False	58	46	79.31
P-Advanced	Multiple Choice	60	38	63.33
P-Advanced	Open-Ended	22	3	13.64

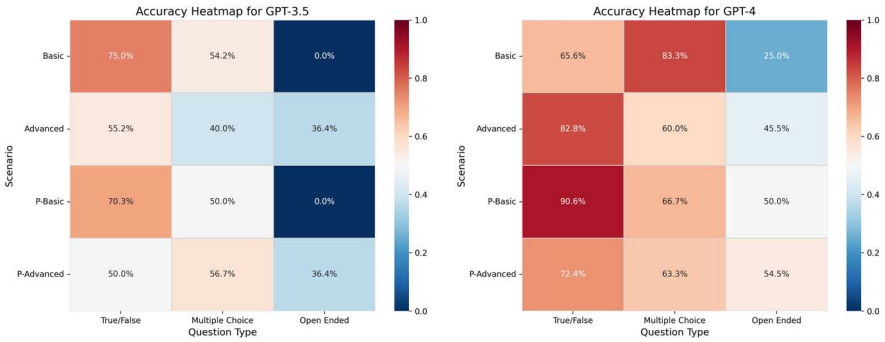
for each individual model (Fig. 1) and comparative bar charts detailing accuracy by question type within each scenario for all models (Fig. 2).

This overview highlights key performance trends across different model architectures, sizes, and access types (closed vs. open-source). It sets the stage for a more detailed examination of performance within each specific scenario in the following subsections, aiming to provide a nuanced understanding of current LLM capabilities and limitations in the tax domain and to identify avenues for future advancements.

3.1 Basic scenario

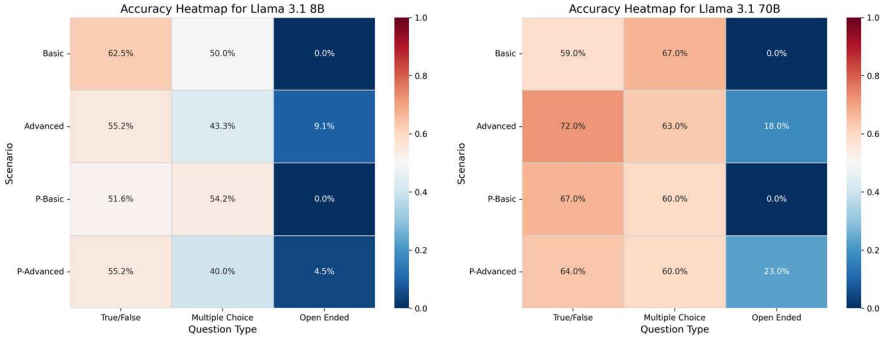
In the basic scenario testing foundational tax knowledge, performance varied significantly. For true/false (T/F) questions, accuracy ranged from Llama 3.1 70B's 59.37% to GPT-4o mini's 81.25%, with GPT-4o mini, GPT-3.5, and Haiku 3.5 generally outperforming GPT-4 and the Llama models. On multiple-choice (MC) questions, GPT-4 and Claude 3.5 Haiku led with 83.33% accuracy, followed by GPT-4o mini (75.00%) and Llama 3.1 70B (67.00%), while GPT-3.5 (54.17%) and Llama 3.1 8B (50.00%) performed lower.

Open-ended questions requiring numerical answers proved extremely challenging for all models. Accuracy was 0.00% across the board (GPT-3.5, Llamas, GPT-4o mini, Haiku 3.5), with the sole exception of GPT-4 achieving 25.00% (1 out of 4 correct). This highlights a critical limitation in the precise numerical application of tax rules at this level. For instance, regarding the maximum EITC investment income (ground truth $10,300), the widespread 0.00% open-ended accuracy means nearly all models failed this question; GPT-4 had only a minimal chance of success by returning $10,000. Also, while GPT-3.5 incorrectly calculated the Pickens's withholding ($10,000 vs. correct $8530), GPT-4 reportedly computed it accurately. This specific success for GPT-4, despite its low 25% open-ended score, suggests a potential edge in numerical processing over the others, which universally failed these basic open-ended tasks. Overall, the basic scenario shows competitive performance among top closed-source models on T/F and MC questions, but reveals a near-universal failure on open-ended numerical questions, with only GPT-4 demonstrating any capability, albeit limited.



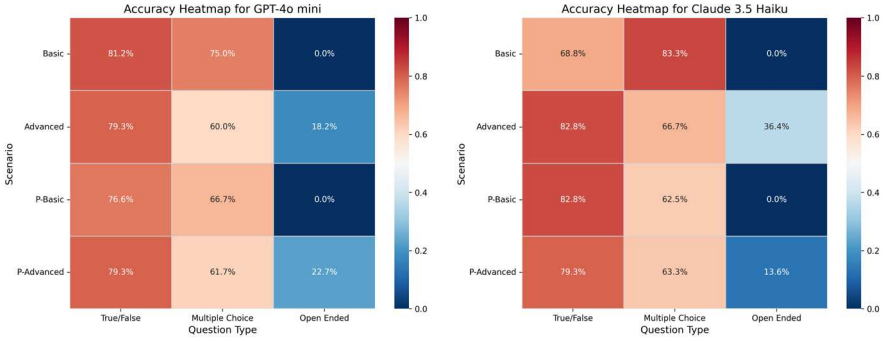
(a) GPT-3.5 Accuracy Heatmap

(b) GPT-4 Accuracy Heatmap



(c) Llama 3.1 8B Accuracy Heatmap

(d) Llama 3.1 70B Accuracy Heatmap

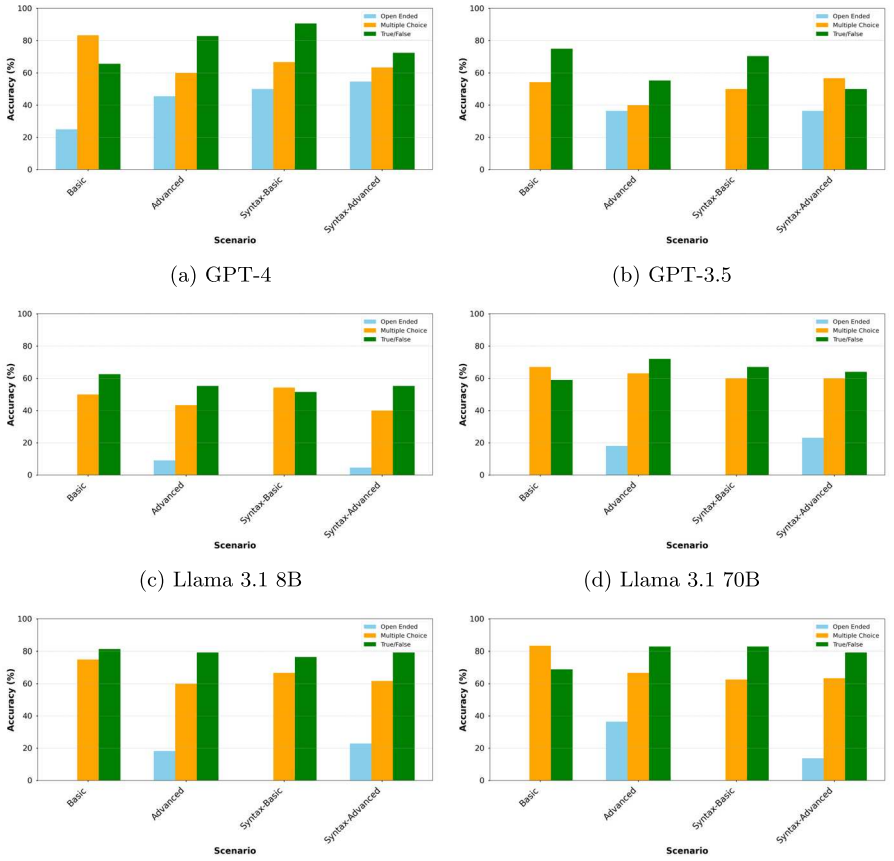


(e) GPT-4o mini Accuracy Heatmap

(f) Claude 3.5 Haiku Accuracy Heatmap

Fig. 1 Accuracy heatmaps for all models across basic and advanced scenarios as well as the corresponding perturbed (P) ones

Figure 1 with Basic scenarios presents accuracy heatmaps for all evaluated LLMs. These visualizations starkly contrast the generally higher accuracy on True/False and Multiple Choice questions with the consistently lower scores for open-ended questions across most models. Comparative performance differences, such as between GPT-4 (b) and Llama 3.1 8B (c), are also readily apparent. While indicating LLM



(a) GPT-4 (b) GPT-3.5

(c) Llama 3.1 8B (d) Llama 3.1 70B

(e) GPT-4o mini (f) Claude 3.5 Haiku

Fig. 2 Accuracy by question type for all models

advancements, the persistent weakness in handling open-ended numerical inquiries emphasizes the critical need for continued improvement.

3.2 Advanced scenario

The advanced scenarios tested nuanced tax comprehension. On true/false questions, GPT-4 and Claude 3.5 Haiku excelled (both 82.76%), significantly outperforming GPT-3.5 and Llama 3.1 8B (both 55.17%). Multiple-choice performance was more varied, led by Haiku 3.5 (66.67%) and Llama 3.1 70B (63.33%), with GPT-3.5 lowest (40.00%).

Open-ended questions saw slightly improved but still low accuracy compared to the basic scenario. GPT-4 led (45.45%), followed by GPT-3.5 and Haiku 3.5 (tied at 36.36%), down to Llama 3.1 8B (9.09%).

Specific examples highlight these differences. In one scenario, GPT-4 correctly identified $1800 taxable unemployment, while GPT-3.5 failed. In another scenario

However, both GPT-4 and GPT-3.5 miscalculate taxpayer's $10 additional IRA tax as $250. This shows that even the best models struggle with precise numerical application of complex rules.

In essence, advanced scenarios confirmed the strengths of models like GPT-4 and Haiku 3.5 in inferential reasoning for T/F questions, but emphasized the universal challenge of accurate numerical calculation in open-ended tax questions, even when overall accuracy slightly improves.

3.3 Syntax-basic scenario

Evaluating robustness to syntactic variations in basic scenarios, GPT-4 excelled on perturbed true/false questions (90.63%), followed by Claude 3.5 Haiku (82.81%) and GPT-4o mini (76.56%). On perturbed multiple-choice, GPT-4 and GPT-4o mini led (both 66.67%), with GPT-3.5 lowest (50.00%). Open-ended questions remained a significant challenge, with all models scoring 0.00% accuracy except for GPT-4, which uniquely improved to 50.00% under perturbation.

Overall, this scenario showcased GPT-4's strong robustness, especially its unique ability to handle some perturbed open-ended numerical questions. Other models showed varying robustness on T/F and MC, but the open-ended barrier persisted for all except GPT-4.

3.4 Syntax-advanced scenario

The syntax-advanced scenarios tested performance on complex tax situations combined with linguistic variations, probing the upper limits of syntactic comprehension and reasoning.

On perturbed advanced true/false questions, Claude 3.5 Haiku and GPT-4o mini demonstrated the highest robustness, both scoring 79.31%. GPT-4 followed closely at 72.41%, while Llama 3.1 70B achieved 63.79%. Llama 3.1 8B (55.17%) and GPT-3.5 (50.00%) found these questions the most challenging.

Multiple-choice questions in this scenario showed relatively tight performance among the top models. GPT-4 and Claude 3.5 Haiku led with 63.33% accuracy. GPT-4o mini (61.67%), Llama 3.1 70B (60.00%), and GPT-3.5 (56.67%) were competitive, while Llama 3.1 8B performed significantly lower at 40.00%.

Open-ended questions remained difficult, though GPT-4 achieved its highest open-ended score across all scenarios here, reaching 54.55%. GPT-3.5 was second with 36.36%. Llama 3.1 70B and GPT-4o mini both scored 22.73%. Claude 3.5 Haiku (13.64%) and Llama 3.1 8B (4.55%) had the lowest accuracy, indicating extreme difficulty with complex, perturbed numerical tasks.

Exploring a specific open-ended question further highlights these differences. For instance, regarding advanced payment of premium tax credit under syntactic modification, GPT-4's accurate response ($4,656, matching ground truth) matches with its leading 54.55% accuracy in this category. Conversely, GPT-3.5's showed discrepancy ($5,258) is consistent with its lower 36.36% score. It is highly probable that models with significantly lower scores in this category (Llama 70B, Mini, Haiku, Llama 8B) also failed this complex calculation.

This comparison demonstrates GPT-4's enhanced ability to maintain semantic understanding and numerical accuracy amidst complex linguistic and regulatory challenges. While other models like Haiku and Mini show strong robustness in certain areas (e.g., True/False), GPT-4's advantage, particularly in the difficult open-ended category, underscores the importance of continued LLM development for handling intricate, real-world tax scenarios.

3.5 Summary of results

While GPT-4 often demonstrates high accuracy, particularly in multiple-choice (e.g., 83.33% in Basic) and advanced true/false questions (82.76%), other closed-source models like Claude 3.5 Haiku achieve comparable or even identical results in specific categories (e.g., 83.33% on Basic Multiple Choice vs. 82.76% on Advanced True/False). GPT-4o mini also shows strong performance, frequently approaching or exceeding GPT-3.5's accuracy (e.g., 75.00% vs. 54.17% on Basic Multiple Choice). In contrast, GPT-3.5 generally lags behind these leading models.

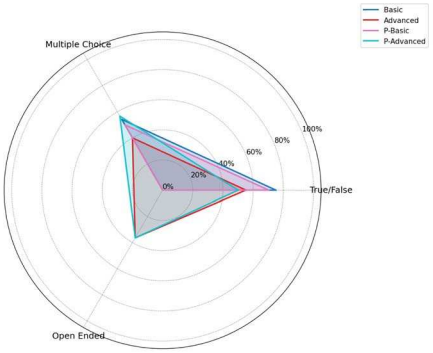
A consistent finding across *all* evaluated models is the significant challenge posed by open-ended questions requiring numerical calculation or specific value extraction. Accuracy on these questions is markedly lower than on true/false or multiple-choice types. Performance often starts at 0% for basic scenarios (observed for GPT-3.5, Llama 8B, Llama 70B, GPT-4o mini, and Haiku 3.5) and remains relatively low even for the best-performing models in more complex scenarios (e.g., GPT-4 achieved 25.00% on Basic open-ended and peaked at 54.55% on Perturbed Advanced open-ended).

Overall, our evaluation reveals a varied landscape of LLM capabilities in the tax domain. Among the leaders, GPT-4 frequently exhibits enhanced performance and adaptability, particularly when handling complex syntactic constructs or challenging open-ended questions. However, models like Haiku and GPT-4o mini show competitive, and sometimes superior, robustness on true/false and multiple-choice tasks. With open-ended questions, while GPT-4 showed the most improvement and highest scores, its performance still remained limited even for it. The distinct performance profile of each model across the four scenario types (basic, advanced, perturbed basic, perturbed advanced) is illustrated in the radar charts in Fig. 3.

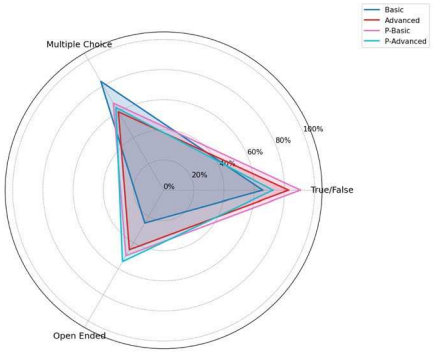
4 Discussion

4.1 Statistical significance

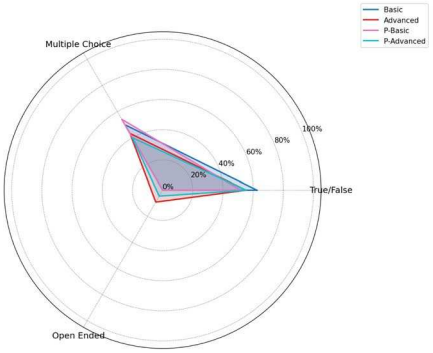
To rigorously assess performance differences across all evaluated Large Language Models (LLMs), pairwise Chi-Square tests of independence were conducted. We compared the proportion of correct versus incorrect answers between each pair of models, first based on their overall performance across all questions, and subsequently broken down by question type (True/False, Multiple Choice, and open-ended). A significance level of $p < 0.05$ was used. Figure 4 provides a visual comparison of overall



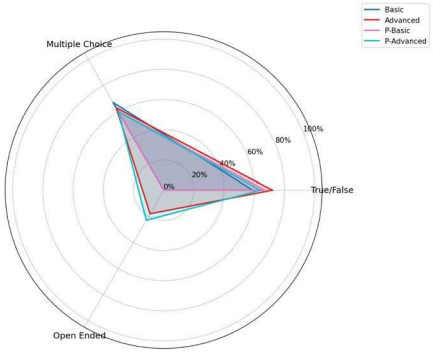
(a) GPT-3.5 Scenario Comparison



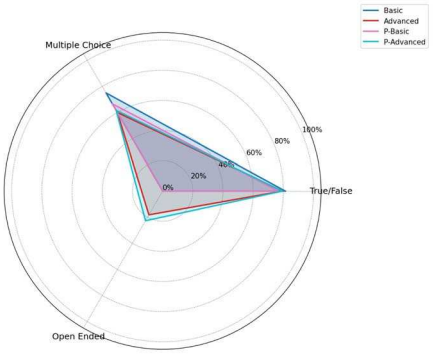
(b) GPT-4 Scenario Comparison



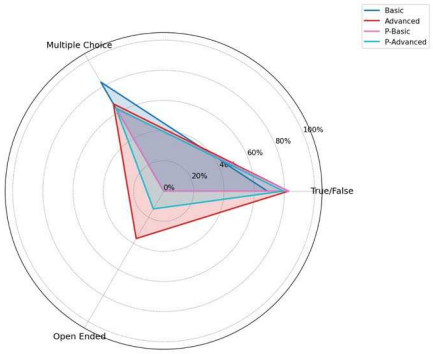
(c) Llama 3.1 8B Scenario Comparison



(d) Llama 3.1 70B Scenario Comparison

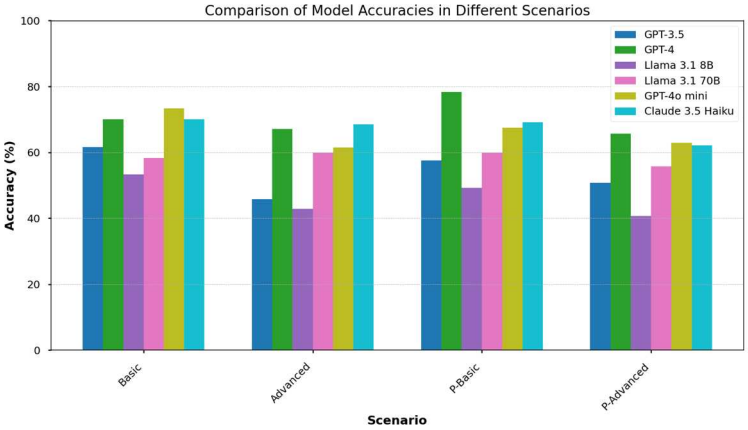


(e) GPT-4o mini Scenario Comparison

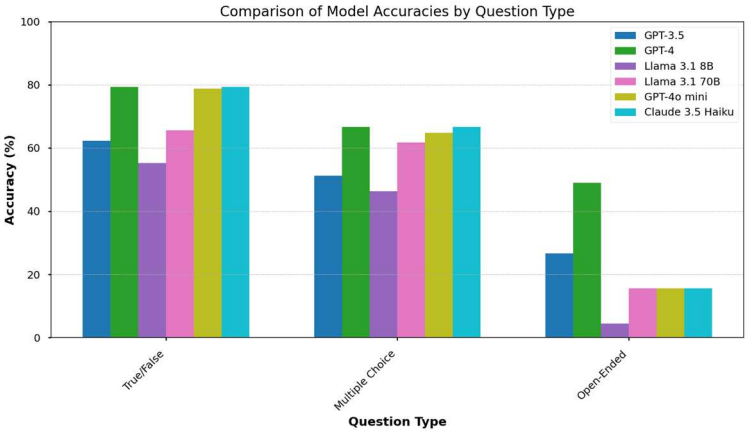


(f) Claude 3.5 Haiku Scenario Comparison

Fig. 3 Scenario comparison radar charts for all models



(a) Comparison of Model Accuracies in Different Scenarios



(b) Comparison of Model Accuracies by Question Type

Fig. 4 Comparison of model accuracies

model accuracies, illustrating these performance tiers across different scenarios (a) and question types (b).

4.1.1 Overall performance

The analysis revealed statistically significant differences in overall accuracy between many model pairs. The top-performing closed-source models demonstrated significant advantages over others. For instance, GPT-4 significantly outperformed GPT-3.5 ($\chi^2(1) = 23.72, p < 0.001$), degree of freedom is (1), and Llama 3.1 8B ($\chi^2(1) = 49.54, p < 0.001$). Similarly, Claude 3.5 Haiku significantly outperformed GPT-3.5 ($\chi^2(1) = 13.91, p < 0.001$) and Llama 3.1 8B ($\chi^2(1) = 35.01, p < 0.001$). Within the open-source models, Llama 3.1 70B significantly outperformed Llama 3.1

8B ($\chi^2(1) = 12.33, p < 0.001$). Interestingly, no statistically significant differences in overall performance were found among the highest-performing group: GPT-4, Claude 3.5 Haiku, and GPT-4o mini (e.g., GPT-4 vs. Haiku: $p = 0.247$; GPT-4o mini vs. Haiku: $p = 0.762$).

4.1.2 Performance by question type

Analyzing performance within each question category provided further granularity:

4.1.3 True/False questions

Significant differences were observed, largely mirroring the overall trend where top closed-source models outperformed others. For example, GPT-4 significantly outperformed GPT-3.5 ($\chi^2(1) = 12.69, p < 0.001$) and Llama 3.1 8B ($\chi^2(1) = 24.00, p < 0.001$). However, within the top performance tier for this question type (GPT-4, Haiku, GPT-4o mini), there were no statistically significant differences (e.g., GPT-4 vs. Haiku: $p = 1.000$; GPT-4o mini vs. Haiku: $p = 0.898$). Llama 3.1 70B showed significantly better performance than Llama 3.1 8B ($\chi^2(1) = 4.12, p = 0.042$.

4.1.4 Multiple choice questions

A similar pattern emerged. GPT-4 and Haiku performed statistically identically (p=1.000) and significantly better than GPT-3.5 ($\chi^2(1) = 7.97, p = 0.005$ for both) and Llama 3.1 8B ($\chi^2(1) = 13.67, p < 0.001$ for both). Llama 3.1 70B performed significantly better than Llama 3.1 8B ($\chi^2(1) = 7.77, p = 0.005$) and showed no significant difference compared to the top group (e.g., vs GPT-4: $p = 0.354$).

4.1.5 Open-ended questions

The statistical landscape changed considerably for this challenging question type. GPT-4 demonstrated a statistically significant advantage over *all* other models evaluated, including GPT-3.5 ($\chi^2(1) = 4.73, p = 0.030$), Claude 3.5 Haiku ($\chi^2(1) = 11.45, p = 0.001$), Llama 3.1 70B ($\chi^2(1) = 11.45, p = 0.001$), and Llama 3.1 8B ($\chi^2(1) = 22.73, p < 0.001$). Below GPT-4, fewer significant differences were found; notably, GPT-3.5 significantly outperformed only Llama 3.1 8B ($\chi^2(1) = 8.46, p = 0.004$). Many pairs among the lower-performing models (Llamas, Mini, Haiku excluding GPT-4) showed no significant differences from each other (e.g., Llama 70B vs. GPT-4o mini: $p = 1.000$).

4.1.6 Summary of statistical findings

These results indicate that while overall performance differences establish clear tiers among the models, the statistical significance varies notably by question type. Top closed-source models (GPT-4, Haiku, Mini) often exhibit statistically similar performance on True/False and Multiple Choice questions, significantly outperforming

GPT-3.5 and the open-source models. Llama 3.1 70B in most cases surpasses the 8B version. However, GPT-4 uniquely stands out with a statistically significant performance advantage specifically in the difficult domain of open-ended tax questions compared to all other tested models.

4.2 Preliminary RAG experiments

To address the limitations of relying solely on model outputs, we experimented with a Retrieval-Augmented Generation (RAG) system using IRS official documents 4012 and 4491 as retrieval sources. However, this simple RAG system, which retrieved the top-3 passages based on question or scenario-based text search, failed to yield meaningful improvements in performance. The large size of the documents and the limited relevance of retrieved passages often resulted in redundant or unhelpful information. The initial experiments suggest that more sophisticated RAG strategies are required to effectively utilize such extensive retrieval sources.

4.3 Effect of syntactic perturbations

Contrary to expectations, we observed no consistent relationship between performance changes and syntactic perturbations in the dataset. As shown in results, accuracy sometimes increased for perturbed questions and sometimes decreased, depending on the model and question type. For example, Llama 3.1 70B improved for advanced open-ended perturbed questions (22.73% perturbed vs. 18.18% original) but showed reduced accuracy for basic multiple-choice perturbed questions (60.04% perturbed vs. 67% original).

This inconsistency suggests that the LLMs did not rely on memorization of the questions and answers. If the models had relied on rote memorization, performance would likely decrease consistently when questions were perturbed. Instead, the observed variability indicates that the models leverage reasoning and comprehension capabilities, albeit inconsistently. This highlights the need for further investigation into how perturbations influence model reasoning and context interpretation.

5 Related work

The advent and evolution of Large Language Models (LLMs) like GPT-3.5 and GPT-4 have spurred significant interest across various specialized domains, including medicine, law, and clinical text processing. This section outlines contributions from several key studies that have explored the application, adaptation, and evaluation of LLMs within these domains, providing a context for our investigation into their performance in the tax domain.

Kung et al. (2023) assessed the capability of ChatGPT on the United States Medical Licensing Exam (USMLE), illustrating the potential of LLMs in medical education and clinical decision-making. ChatGPT performed at or near the passing threshold for all three exams without any specialized training or reinforcement, with

accuracies ranging from 45.4% to 75.0% for Step 1, 54.1% to 61.5% for Step 2CK, and 61.5% to 68.8% for Step 3, depending on the encoding format used.

Laban et al. (2023) introduced SUMMEDITS, a benchmark for evaluating LLMs' factual reasoning through summarization tasks across various domains. Their work underscores the importance of accurate information processing by LLMs, a critical aspect for applications within the tax domain.

Chen et al. (2024) demonstrated how domain-specific enhancements, notably through retrieval-augmented generation (RAG) and instructional prompts, could significantly improve the performance of LLMs in medical contexts. Specific improvements included an increase in accuracy by over 5% on clinical question answering tasks compared to baseline LLMs.

Arefeen et al. (2023) proposed LeanContext, a system designed to reduce the operational costs of LLM deployment in domain-specific QA systems. By optimizing context relevance and size, LeanContext reduced the costs by 37-67% while maintaining a performance drop of only 1.41-2.65% in terms of ROUGE-1 score.

Furthermore, Alsentzer et al. (2019). explored the development and application of clinical BERT embeddings, focusing on adapting BERT for clinical texts to align with the broader objective of enhancing domain-specific accuracy in models, including applications in the medical field. They developed BERT models specifically for clinical text, demonstrating notable improvements on clinical natural language processing (NLP) tasks when compared to non-specific embeddings. Their work highlighted the benefits of domain-specific model training through achieving new state-of-the-art performance on the MedNLI task with an accuracy of 82.7%, and showcasing improvements across various clinical NLP tasks including i2b2 2010 and i2b2 2012 tasks.

Lastly, Katz et al. (2023) provided a comprehensive overview of NLP in the legal domain, emphasizing the growing sophistication of methods deployed and the increasing integration of LLMs into legal applications. Their analysis of trends in Legal NLP offers insights into the parallels and potential applications of LLMs in tax law and policy. These studies collectively underscore the versatility and potential of LLMs across specialized domains, setting the stage for our exploration of their applicability and performance within the tax domain.

6 Conclusion and future work

This paper assessed the efficacy of a diverse set of large language models (LLMs)-including GPT-4, GPT-3.5, GPT-4o mini, Claude 3.5 Haiku, Llama 3.1 70B, and Llama 3.1 8B-in answering tax-related questions derived from IRS VITA certification tests. Using an expanded dataset of 130 original and 260 syntactically perturbed questions, we benchmarked performance across true/false, multiple-choice, and open-ended tax scenarios.

Our findings reveal distinct performance tiers, with leading closed-source models (GPT-4, Haiku, Mini) generally outperforming GPT-3.5 and the open-source Llama models (70B, 8B). While these top models demonstrated considerable proficiency on true/false and multiple-choice questions (often achieving 60%-90% accuracy), a crit-

ical limitation emerged across *all* evaluated LLMs: significant difficulty with open-ended questions requiring precise numerical reasoning and application of tax rules. Even the best performer in this category, GPT-4 (peak accuracy 54.55%), showed inconsistencies, while others often scored near zero. Notably, GPT-4 demonstrated a statistically significant advantage over all other models specifically on these challenging open-ended questions.

This study provides a benchmark and methodology for evaluating LLM capabilities in the specialized domain of tax law. Key areas for future research include enhancing numerical reasoning for open-ended questions (via fine-tuning, advanced prompting, or tool use), developing robust RAG systems capable of navigating dynamic tax codes, and conducting thorough reliability and ethical assessments. Continuous evaluation, using standardized tests like VITA, remains crucial to ensure the safe and effective integration of LLMs into tax assistance and education.

Acknowledgements The authors thank Nina E. Olson, the Executive Director of the Center for Taxpayer Right, for the consultant and guidance on this project. We also appreciate anonymous reviewers of Artificial Intelligence and Law as well as Jack G. Conrad (the guest editor) for their feedback and comments to improve this paper. This project has been partially supported by the NSF DASS program under grants CCF-2317206 and CCF-2317207.

References

Alsentzer E, Murphy J, Boag W, Weng W-H, Jindi D, Naumann T, McDermott M (2019) Publicly available clinical BERT embeddings. In Rumshisky A, Roberts K, Bethard S, Naumann T (eds) Proceedings of the 2nd clinical natural language processing workshop. Minneapolis, Minnesota, USA. Association for Computational Linguistics, pp. 72–78

April Technologies (2023) Embed intelligent tax software. https://www.getapril.com

Arefeen MA, Debnath B, Chakradhar S (2023) Leancontext: cost-efficient domain-specific question answering using llms

Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, Lee P, Lee YT, Li Y, Lundberg S, Nori H, Palangi H, Ribeiro MT, Zhang Y (2023) Sparks of artificial general intelligence: early experiments with gpt-4

Chen X, You M, Wang L, Liu W, Fu Y, Xu J, Zhang S, Chen G, Li K, Li J (2024) Evaluating and enhancing large language models performance in domain-specific medicine: Osteoarthritis management with docoa

Drenik G (2023) Ai-powered tax system is creating a new paradigm: Will banks and fintechs adopt the technology to help their customers save on their tax bill?. Forbes, Accessed: 2025-April-18

Floridi L, Chiriatti M (2020) Gpt-3: its nature, scope, limits, and consequences. Minds Mach 30(4):681–694

Gogani-Khiabani S, Dewangan V, Olson N, Trivedi A, Tizpaz-Niari S (2025) Technical challenges in maintaining tax prep software with large language models. arXiv:2504.18693

Internal Revenue Service (2022) VITA/TCE Volunteer Assistor's Test/Retest

Jiang Z, Xu FF, Araki J, Neubig G (2020) How can we know what language models know? Trans Assoc Comput Ling 8:423–438

Katz DM, Hartung D, Gerlach L, Jana A, Bommarito II MJ (2023) Natural language processing in the legal domain

Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggarwal R, Gamurot A, Lirio MMR, Diamante K, Cariño L, Logrono PEC, Tsitos JM (2023) Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2(2):e0000198

Laban P, Kryscinski W, Agarwal D, Fabbri A, Xiong C, Joty S, Wu C-S (2023) SummEdits: measuring LLM ability at factual reasoning through the lens of summarization. In Bouamor H, Pino J, Bali K (eds) Proceedings of the 2023 conference on empirical methods in natural language processing. Singapore. Association for Computational Linguistics, pp 9662–9676

Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) Language models are few-shot learners

Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, King N, Larson J, Li Y, Liu W, Luo R, McKinney SM, Ness RO, Poon H, Qin T, Usuyama N, White C, Horvitz E (2023) Can generalist foundation models outcompete special-purpose tuning? case study in medicine

OpenAI, et al (2024) Gpt-4 technical report

Ribeiro MT, Wu T, Guestrin C, Singh S (2020) Beyond accuracy: behavioral testing of nlp models with checklist

Srinivas D, Das R, Tizpaz-Niari S, Trivedi A, Pacheco ML (2023) On the potential and limitations of few-shot in-context learning to generate metamorphic specifications for tax preparation software. In Preoțiuc-Pietro D, Goanta C, Chalkidis I, Barrett L, Spanakis G, Aletras N (eds) Proceedings of the natural legal language processing workshop 2023. Singapore. Association for Computational Linguistics, pp 230–243

Tizpaz-Niari S, Darian S, Trivedi A (2024) Metamorphic debugging for accountable software

Tizpaz-Niari S, Monjezi V, Wagner M, Darian S, Reed K, Trivedi A (2023) Metamorphic testing and debugging of tax preparation software. In 2023 IEEE/ACM 45th international conference on software engineering: software engineering in society (ICSE-SEIS), pp 138–149

Webb T, Mondal SS, Wang C, Krabach B, Momennejad I (2024) A prefrontal cortex-inspired architecture for planning in large language models

Wei J, Bosma M, Zhao VY, Guu K, Yu AW, Lester B, Du N, Dai AM, Le QV (2022) Finetuned language models are zero-shot learners

Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le Q, Zhou D (2023) Chain-of-thought prompting elicits reasoning in large language models

Yu J, McCluskey K, Mukherjee S (2020) Tax knowledge graph for a smarter and more personalized turbotax

Authors and Affiliations

Sina Gogani-Khiabani[1] · Ashutosh Trivedi[2] · ShinPing Chyi[3] · Saeid Tizpaz-Niari[1]

✉ Sina Gogani-Khiabani
sgoga3@uic.edu

Ashutosh Trivedi
Ashutosh.Trivedi@colorado.edu

ShinPing Chyi
epchinesechyi@gmail.com

Saeid Tizpaz-Niari
saeid@uic.edu

[1] CS Department, University of Illinois Chicago, Chicago, IL, USA

[2] CS Department, University of Colorado Boulder, Boulder, CO, USA

[3] CPA, 6452 Amposta Dr, El Paso 79912, TX, USA