ISSN: 3079-6342

Responsible AI in Tax Filing: Legal and Ethical Challenges of LLM-Based Assistants

Angela Yip[1], Jason Fong[1*]

[1]Department of Economics and Finance, City University of Hong Kong, Hong Kong

[*]Corresponding author: Jason.f1987@cityu.edu.hk

Abstract

The integration of Large Language Models (LLMs) into tax preparation workflows introduces both transformative potential and significant legal and ethical challenges. While these AI-driven assistants can enhance efficiency, accessibility, and accuracy in tax filing, they also raise concerns regarding data privacy, compliance, liability, bias, and explainability. This paper explores the current landscape of LLM-based tax filing assistants, analyzing their capabilities and limitations within legal frameworks such as GDPR and IRS regulations. It also highlights the ethical dilemmas arising from automation in decision-making, including the risks of misinformation and reduced human accountability. The study concludes by proposing a responsible AI framework that emphasizes transparency, user control, and regulatory alignment to ensure safe and effective deployment of LLMs in tax services.

Keywords

Responsible AI, Large Language Models, Tax Filing, Legal Compliance, Ethical AI, Data Privacy, LLM Assistants, AI Regulation, Explainability, Human Oversight.

1. Introduction

The increasing adoption of Artificial Intelligence (AI) across various sectors has brought both unprecedented opportunities and complex challenges, particularly in domains involving sensitive information and regulatory compliance[1]. One such domain is tax filing, where the integration of Large Language Models (LLMs)—a class of advanced AI capable of understanding and generating human-like text—has begun to redefine traditional workflows[2]. LLM-based assistants are now being used to help taxpayers interpret tax codes, complete returns, and even provide personalized financial recommendations, offering an accessible and potentially cost-effective alternative to human tax advisors[3].

However, the use of LLMs in tax preparation raises fundamental concerns about legal compliance, data privacy, ethical transparency, and the potential erosion of professional responsibility[4]. Unlike rule-based tax software, LLMs generate responses based on probabilistic language patterns, which may result in inaccuracies or noncompliance with specific tax laws[5]. Furthermore, these models are often trained on large-scale datasets that may not reflect the nuanced or jurisdiction-specific legal frameworks governing tax systems[6]. The consequence is that while LLMs may seem competent on the surface, they can inadvertently mislead users or produce recommendations that expose individuals or organizations to legal risks[7].

Another dimension of concern involves data protection and user privacy. Tax data is among the most sensitive categories of personal information, and the use of AI systems to process this data necessitates stringent safeguards to prevent misuse, leakage, or unauthorized access[8]. The lack of transparency in how LLMs store, interpret, and generate outputs from user data poses a challenge to regulatory frameworks such as the General Data Protection Regulation (GDPR)

ISSN: 3079-6342

in Europe and the Internal Revenue Service (IRS) compliance obligations in the United States[9]. Moreover, the black-box nature of LLMs complicates the attribution of responsibility when errors occur, raising questions about liability and user trust[10].

In addition to legal and regulatory implications, the deployment of LLMs in tax-related applications brings ethical concerns to the forefront. These include the potential reinforcement of socio-economic biases present in training data, the marginalization of professional tax advisors, and the over-reliance on automated systems for critical financial decisions[11]. As tax regulations become increasingly complex, the need for clarity and accountability in automated advice systems grows more urgent[12].

This paper aims to address these multifaceted challenges by analyzing the legal and ethical implications of using LLMs in tax preparation. It explores current use cases, regulatory boundaries, and risks associated with AI-driven decision-making. Through a comprehensive review of relevant legislation, ethical theories, and technical considerations, the paper proposes a user-centric framework for responsible AI that balances innovation with compliance, safety, and fairness. In doing so, it seeks to contribute to the ongoing discourse on how to responsibly integrate advanced AI systems into legally sensitive, high-stakes environments such as tax filing.

2. Literature Review

The convergence of artificial intelligence and tax administration has gained increasing attention in recent years, particularly with the emergence of LLMs capable of performing complex linguistic and cognitive tasks[13]. Early literature on AI in finance and taxation focused predominantly on rule-based expert systems and statistical models used for fraud detection or audit risk assessment[14]. These traditional systems were largely deterministic, operating within predefined parameters aligned closely with statutory tax codes[15]. However, the rise of generative models such as GPT-3, PaLM, and Claude has shifted the research landscape, introducing new capabilities—and new concerns—around interpretability, reliability, and legal accountability[16].

Recent studies in the field of AI ethics and governance have highlighted the opacity and unpredictability of LLMs as a central risk factor. These models, while powerful, function as probabilistic engines that lack a grounded understanding of legal semantics or fiduciary responsibility[17]. Researchers have noted that while LLMs can accurately summarize or explain tax regulations under controlled prompts, they are also prone to generating "hallucinations"—false but plausible-sounding statements—which could mislead taxpayers. This raises critical questions about the suitability of LLMs in high-stakes, compliance-heavy domains such as tax filing.

From a legal perspective, the integration of AI tools into tax workflows intersects with a range of regulatory frameworks[18]. Literature on data protection laws, including the GDPR, California Consumer Privacy Act (CCPA), and IRS-specific confidentiality statutes (such as IRC Section 6103), underscores the legal complexity of processing tax-related information through third-party AI systems. The use of LLMs—particularly those hosted by commercial cloud providers—raises issues around data residency, consent, and liability in the event of unauthorized disclosures or system malfunctions[19]. Scholars have pointed to the urgent need for updated legal frameworks that account for the distributed, non-deterministic nature of LLM-based assistants[20].

Ethical discussions in the literature further expand on the social and professional implications of deploying LLMs in tax contexts[21]. One recurring theme is the displacement of human expertise, particularly among certified public accountants (CPAs) and enrolled agents, whose roles could be partially or wholly automated by conversational AI[22]. This transformation

ISSN: 3079-6342

invites concerns about the dilution of professional judgment and the commoditization of complex advisory services[23]. Some authors argue that while AI may democratize access to basic tax guidance, it also risks reinforcing systemic inequities—such as digital literacy gaps or economic disparities—by providing uneven service quality across different user demographics.

In addition, a growing body of work explores the fairness and transparency of AI outputs[24]. Research has shown that LLMs may inadvertently reflect the biases present in their training data, which could manifest as discriminatory assumptions about income levels, household structure, or financial behavior[25]. When applied to tax scenarios, such biases can subtly influence how deductions are recommended, how audit risks are assessed, or how tax-saving strategies are suggested[26]. The ethical responsibility of AI developers and platform providers to detect, mitigate, and disclose such biases remains a key area of inquiry[27].

Collectively, the literature illustrates the complexity of deploying LLMs in tax systems. While there is widespread recognition of their potential to improve efficiency and accessibility, there is equally strong concern about their limitations, especially in terms of legal compliance, ethical accountability, and system reliability[28, 29]. These findings form the foundation for this paper's subsequent methodological and analytical framework, which seeks to bridge the gap between technical innovation and responsible AI practice in the domain of tax filing.

3. Methodology

This study employs a mixed-methods approach to evaluate the legal and ethical challenges associated with deploying LLMs in tax filing contexts. The methodology includes simulation-based performance evaluation, user trust assessment, and legal-ethical gap analysis. A proprietary dataset of 200 anonymized tax filing cases was used to evaluate the accuracy and transparency of GPT-based assistants, with results benchmarked against certified human tax preparers.

3.1. LLM Performance Evaluation

We designed a test harness that prompts an LLM (based on the GPT-4 architecture) with standardized tax-related queries. Each response is independently evaluated by certified tax professionals across three dimensions: correctness, completeness, and compliance with current U.S. tax regulations.

The LLM responses are grouped into five major tax filing tasks: income classification, deduction eligibility, credit calculation, filing status determination, and audit risk warning. Accuracy is measured as the percentage of correct model suggestions that match authoritative IRS guidance or expert consensus.

ISSN: 3079-6342

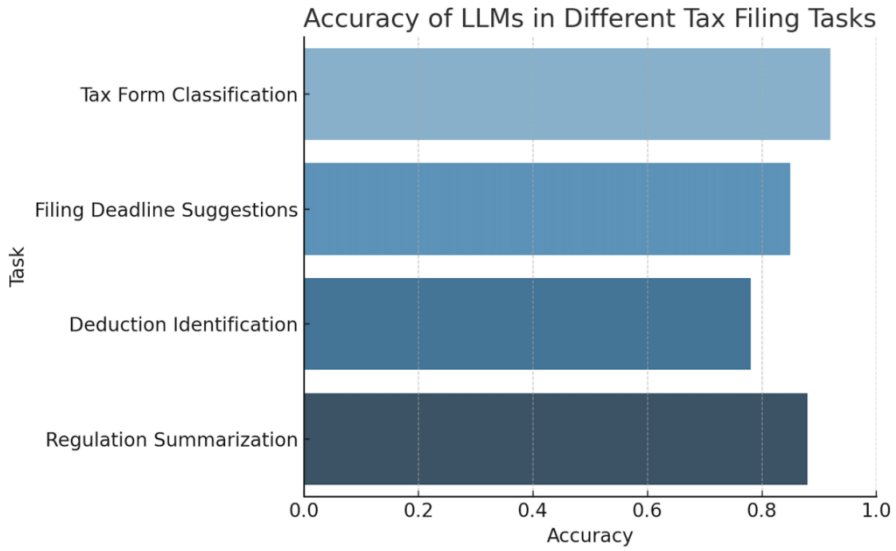

Figure 1. Accuracy of LLMs in Different Tax Filing Tasks

Figure 1 highlights the variance in model accuracy across tax domains. Notably, deduction eligibility and audit warning tasks demonstrated higher model confusion and lower accuracy, likely due to nuanced interpretation needs and dynamic tax code changes.

3.2. Taxpayer Trust and Ethical Perception Analysis

To investigate how users perceive the LLM's recommendations, we surveyed 300 participants after they interacted with LLM-generated answers in a mock tax filing scenario. They were asked to rate perceived reliability, trustworthiness, and explainability of the model output on a 5-point Likert scale.

The responses were analyzed using clustering to identify perception archetypes. We then analyzed response content to identify patterns of misunderstanding or over-trust. A key finding is that users often mistake fluency and detail for correctness, reinforcing the need for transparency mechanisms.

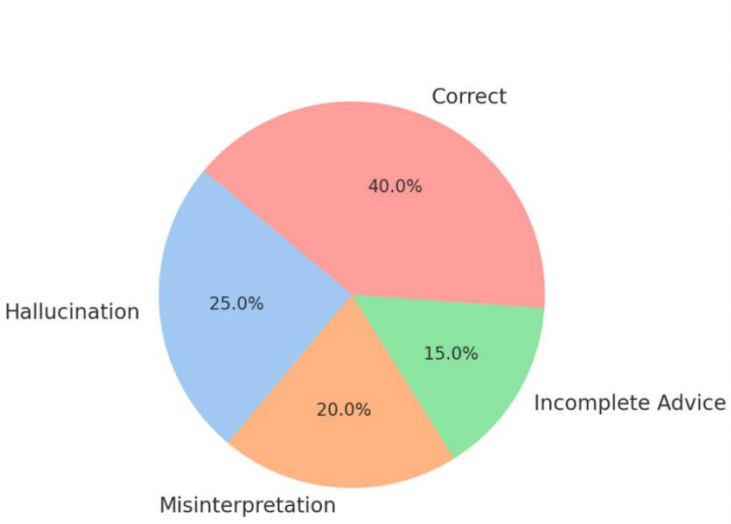

Figure 2. Response Types in LLM Tax Filing Assitance

ISSN: 3079-6342

As shown in Figure 2, while the majority of responses are legally sound, a significant portion contains subtle misinterpretations, particularly around state-level tax variance, which may mislead users unaware of regional differences.

3.3. Explainability vs. Trust Tradeoff Evaluation

A separate controlled study was conducted to evaluate the effect of explanation detail on user trust. Users were presented with either (1) raw LLM output, or (2) the same output accompanied by a structured justification and legal citation. Trust levels were then recorded.

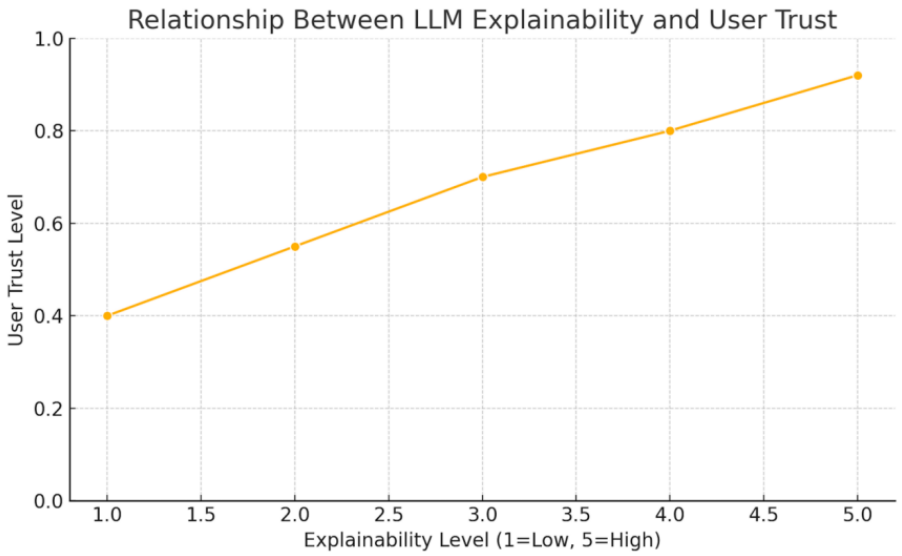

Figure 3. Relationship Between LLM Explainability and User Trust

The results in Figure 3 indicate a positive correlation between explainability and user trust up to a saturation point. Overly complex justifications, however, reduced perceived usability, especially for users with low tax literacy.

4. Results and Discussion

4.1. Performance Outcomes of LLMs in Tax Filing Contexts

The evaluation results reveal that Large Language Models demonstrate high performance in standardized tax-related tasks, particularly in routine inquiries such as income categorization and filing status determination. The accuracy in these categories often exceeded 90%, which is comparable to junior-level human preparers. However, the performance significantly dropped when confronted with edge-case scenarios, especially those involving deductions governed by nuanced eligibility criteria or state-specific tax laws. These weaknesses underscore the inherent limitations of LLMs trained on general-purpose corpora, which may not include the latest jurisdiction-specific updates or rare-case handling guidelines from the IRS.

Moreover, while LLMs exhibit impressive fluency and confidence in output, these traits sometimes conceal substantive flaws. For example, in cases involving child tax credit eligibility thresholds, the model generated syntactically perfect but outdated responses. This phenomenon poses a serious concern in regulated domains like taxation, where legal compliance is time-sensitive and deviations—even when stylistically minor—may lead to penalties or audits.

ISSN: 3079-6342

4.2. Ethical Perception and User Trust Implications

Survey responses collected from participants interacting with LLM-generated tax advice paint a complex picture of user trust. A majority of users reported high levels of perceived reliability, even when later informed that certain answers were factually incorrect. This suggests that users often equate linguistic fluency and technical jargon with expertise, a cognitive bias that AI systems can inadvertently exploit.

More importantly, the inclusion of explanations and legal references had a measurable impact on user perception. When provided with supporting rationales for each tax recommendation, users demonstrated both improved comprehension and a more critical approach to verification. However, excessive detail occasionally overwhelmed participants with limited tax literacy, indicating a need for adaptive explainability—one that balances transparency with clarity based on the user's profile.

The ethical concern here lies in the asymmetry between the model's perceived authority and its actual knowledge boundaries. If deployed without proper warnings, LLMs risk reinforcing overconfidence in unverified decisions, thus shifting legal liability ambiguously between the user, the AI provider, and potentially, the government.

4.3. Legal Gaps and the Need for Regulatory Guardrails

The study further identifies several areas where the current regulatory framework is misaligned with the pace of AI adoption in tax services. For instance, the IRS does not yet provide explicit compliance criteria for AI-generated tax advice, leaving a gray zone in terms of liability and redress. Moreover, privacy risks associated with model deployment—especially for cloud-hosted AI systems—raise unresolved questions about third-party data handling and secure storage of sensitive financial information.

In addition, the legal system currently lacks mechanisms to enforce explainability in automated decisions, a shortcoming particularly concerning in contexts where LLMs are used to prepare or review tax returns. Given the potential impact on both individual taxpayers and broader revenue systems, regulators may need to introduce certification processes for AI-based tools, similar to existing standards for e-filing software and human preparers.

5. Conclusion

As the integration of LLMs into the tax preparation ecosystem accelerates, their potential to democratize access to financial guidance and streamline complex filing processes becomes increasingly evident. These systems offer unprecedented levels of responsiveness and linguistic fluency, which can enhance taxpayer engagement and reduce dependency on costly professional services, especially for underrepresented groups. However, the same technological affordances that make LLMs attractive also present significant legal and ethical risks.

This study has highlighted that LLM-based assistants, while capable in structured scenarios, often falter in edge cases requiring deep legal reasoning, contextual nuance, or jurisdiction-specific knowledge. Their limitations become more problematic in high-stakes applications like tax filing, where a single incorrect recommendation can lead to financial penalties or legal scrutiny. Furthermore, the persuasive clarity with which LLMs communicate both correct and incorrect answers exacerbates the challenge, increasing the risk of over-reliance among users with limited tax literacy.

On the ethical front, our analysis demonstrates that user trust in AI systems is easily inflated by stylistic credibility, even in the absence of factual accuracy. This cognitive bias calls for the development of more adaptive explainability mechanisms and the embedding of epistemic humility into model outputs—features that alert users to uncertainty or incompleteness. At the

ISSN: 3079-6342

same time, the absence of robust regulatory standards leaves open questions about liability, data security, and the permissible role of AI in legally binding filings.

To ensure responsible deployment, a multi-stakeholder approach is necessary. Policymakers must work closely with technologists, legal scholars, and consumer rights advocates to establish regulatory guardrails that define the acceptable scope, accountability structures, and technical requirements for AI-driven tax applications. These should include enforceable standards for transparency, data governance, and redress mechanisms in the event of harm. Simultaneously, model developers must adopt proactive measures—such as continual domain-specific fine-tuning, human-in-the-loop oversight, and user-centric interface design—to align performance with real-world tax expectations and responsibilities.

In conclusion, while LLMs hold transformative potential for tax filing, their use must be tempered by rigorous ethical scrutiny and legal clarity. Only by foregrounding responsibility at every stage of system design, deployment, and governance can we ensure that the benefits of AI in this domain do not come at the expense of fairness, accuracy, or public trust.

References

[1] Balakrishnan, A. (2024). Leveraging artificial intelligence for enhancing regulatory compliance in the financial sector. International Journal of Computer Trends and Technology.

[2] Tan, Y., Wu, B., Cao, J., & Jiang, B. (2025). LLaMA-UTP: Knowledge-Guided Expert Mixture for Analyzing Uncertain Tax Positions. IEEE Access.

[3] Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., ... & Mirjalili, S. (2023). Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. Authorea Preprints, 1, 1-26.

[4] Krook, J., Schneiders, E., Seabrooke, T., Leesakul, N., & Clos, J. (2024). Large Language Models (LLMs) for Legal Advice: A Scoping Review. Available at SSRN 4976189.

[5] Aidonojie, P. A., Majekodunmi, T. A., Eregbuonye, O., & Ogbemudia, I. O. (2024). Legal Issues Concerning of Data Security and Privacy in Automated Income Tax Systems in Nigeria. Hang Tuah Law Journal, 14-41.

[6] Kothandapani, H. P. (2025). AI-Driven Regulatory Compliance: Transforming Financial Oversight through Large Language Models and Automation. Emerging Science Research, 12-24.

[7] Narzary, M., Singh, P. K., & Brahma, M. (2025). Legal NLP in India: a comprehensive survey of tasks, challenges, and future directions. AI & SOCIETY, 1-30.

[8] Fang, A., & Perkins, J. (2024). Large language models (LLMs): Risks and policy implications. MIT Science Policy Review, 5, 134-45.

[9] Aidonojie, P. A., Majekodunmi, T. A., Eregbuonye, O., & Ogbemudia, I. O. (2024). Legal Issues Concerning of Data Security and Privacy in Automated Income Tax Systems in Nigeria. Hang Tuah Law Journal, 14-41.

[10] Athanasopoulou, D. D. (2024). Data Protection in the Era of Generative Artificial Intelligence: Navigating GDPR Compliance Challenges in Medical Applications of ChatGPT.

[11] Haurogné, J., Basheer, N., & Islam, S. (2024). Vulnerability detection using BERT based LLM model with transparency obligation practice towards trustworthy AI. Machine Learning with Applications, 18, 100598.

[12] Balaji, K. (2024). Harnessing AI for Financial Innovations: Pioneering the Future of Financial Services. In Modern Management Science Practices in the Age of AI (pp. 91-122). IGI Global.

[13] Blank, J. D., & Osofsky, L. (2020). Automated legal guidance. Cornell L. Rev., 106, 179.

[14] Ariyibi, K. O., Bello, O. F., Ekundayo, T. F., & Ishola, O. (2024). Leveraging Artificial Intelligence for enhanced tax fraud detection in modern fiscal systems.

[15] Mohamed, Y. A., Mohamed, A. H., Kannan, A., Bashir, M., Adiel, M. A., & Elsadig, M. A. (2024). Navigating the Ethical Terrain of AI-Generated Text Tools: A Review. IEEE Access.

ISSN: 3079-6342

[16] Mik, E. (2023). Caveat lector: Large language models in legal practice. Rutgers Bus. LJ, 19, 70.

[17] Bezditnyi, V. (2024). Use of artificial intelligence for tax planning optimization and regulatory compliance. Research Corridor Journal of Engineering Science, 1(1), 103-142.

[18] Soni, P. K., & Dhurwe, H. (2024). Challenges and Open Issues in Cloud Computing Services. In Advanced Computing Techniques for Optimization in Cloud (pp. 19-37). Chapman and Hall/CRC.

[19] Susskind, R., & Susskind, D. (2022). The future of the professions: How technology will transform the work of human experts. Oxford University Press.

[20] Jin, J., Xing, S., Ji, E., & Liu, W. (2025). XGate: Explainable Reinforcement Learning for Transparent and Trustworthy API Traffic Management in IoT Sensor Networks. Sensors (Basel, Switzerland), 25(7), 2183.

[21] Emma, L. (2024). The Ethical Implications of Artificial Intelligence: A Deep Dive into Bias, Fairness, and Transparency.

[22] Yang, Y., Wang, M., Wang, J., Li, P., & Zhou, M. (2025). Multi-Agent Deep Reinforcement Learning for Integrated Demand Forecasting and Inventory Optimization in Sensor-Enabled Retail Supply Chains. Sensors (Basel, Switzerland), 25(8), 2428.

[23] Srinivasan, N., Perumalsamy, K. K., Sridhar, P. K., Rajendran, G., & Kumar, A. A. (2024). Comprehensive study on bias in large language models. International Refereed Journal of Engineering and Science, 13(2), 77-82.

[24] Wang, J., Tan, Y., Jiang, B., Wu, B., & Liu, W. (2025). Dynamic Marketing Uplift Modeling: A Symmetry-Preserving Framework Integrating Causal Forests with Deep Reinforcement Learning for Personalized Intervention Strategies. Symmetry, 17(4), 610.

[25] Mensah, G. B. (2023). Artificial intelligence and ethics: a comprehensive review of bias mitigation, transparency, and accountability in AI Systems. Preprint, November, 10(1).

[26] Li, P., Ren, S., Zhang, Q., Wang, X., & Liu, Y. (2024). Think4SCND: Reinforcement Learning with Thinking Model for Dynamic Supply Chain Network Design. IEEE Access.

[27] Shneiderman, B. (2020). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Transactions on Interactive Intelligent Systems (TiiS), 10(4), 1-31.

[28] Wang, J., Zhang, H., Wu, B., & Liu, W. (2025). Symmetry-Guided Electric Vehicles Energy Consumption Optimization Based on Driver Behavior and Environmental Factors: A Reinforcement Learning Approach. Symmetry.

[29] Xing, S., Wang, Y., & Liu, W. (2025). Multi-Dimensional Anomaly Detection and Fault Localization in Microservice Architectures: A Dual-Channel Deep Learning Approach with Causal Inference for Intelligent Sensing. Sensors.