Journal of Accounting, Ethics & Public Policy, JAEPPVol. 27 No. 2 (2026)
e-ISSN 2956-8390; ISSN1089652X

Enforcing Ethics in the Age of Artificial Intelligence: Circular 230 and AI-Assisted Tax Practice

Insha Khan
Campbellsville University, Kentucky, USA

Doi10.60154/jaepp.2026.v27n2p102
Submitted: 21 June 2026
Accepted: 26 August 2026
Published: 18 November 2026
Copyright 2026 Author(s)
Under Creative Commons CC-BY 4.0 OPEN ACCESS

Cite As:
Khan, I., (2026). Enforcing Ethics in the Age of Artificial Intelligence: Circular 230 and AI-Assisted Tax Practice. Journal of Accounting, Ethics & Public Policy, JAEPP, 27(2), https://doi.org/10.60154/jaepp.2026.v27n2p149

Abstract
As generative AI tools become embedded in U.S. tax advisory practice, a fundamental governance question arises: does IRS Circular 230 adequately regulate practitioners who rely on AI-generated guidance? This study addresses this question through a two-phase empirical design grounded in regulatory lag theory. Phase 1 uses archival content analysis to examine 187 published OPR disciplinary actions against CPAs and Enrolled Agents (2020–2025). Results show that 78.1 percent of sanctions are issued under §10.82(b) (expedited suspension), while zero cases cite §10.22 (due diligence), §10.35 (competence), or §10.37 (written advice standards). Phase 2 evaluates 30 AI-generated tax advisory outputs from ChatGPT, Copilot, and Google Gemini using a seven-dimension Circular 230-informed professional responsibility rubric, with scoring based on structured researcher judgment. AI tools achieve strong tax law accuracy scores (mean 3.33/4.0) but receive lower scores on disclaimer adequacy (1.70) and professional referral (1.63). Together, these findings are consistent with a potential regulatory alignment gap: the provisions most relevant to AI-assisted practice quality are absent from the published OPR sanctions reviewed, while AI outputs show recurring weaknesses in written-advice safeguards, disclaimer adequacy, and professional referral. The study discusses implications for practitioners, standard-setters, and regulators.

Keywords: Artificial intelligence, Circular 230, tax ethics, Office of Professional Responsibility, regulatory lag, written advice, competence, due diligence, content analysis

I. Introduction

Tax practitioners are increasingly relying on generative artificial intelligence tools to assist with research, draft written advice, and respond to client tax questions (Jenkins 2024; VanDenburgh, Tribou, and Braswell 2025). Recent practitioner surveys also indicate growing adoption of generative AI across tax and accounting functions, including tax research, return preparation, advisory services, and document summarization (Thomson Reuters Institute 2025).Tools such as ChatGPT, Microsoft Copilot, and Google Gemini are readily accessible to sole practitioners and large firms alike, and their integration into tax advisory workflows is accelerating rapidly. Yet the professional standards governing these practitioners were developed for an environment in which human judgment was the exclusive source of substantive advice. IRS Circular 230, the primary regulatory framework governing practice before the Internal Revenue Service, establishes obligations related to competence, due diligence, and written tax advice, but does not contemplate the scenario in which an AI system generates the substantive content underlying professional advice. The practitioner, not the AI tool, remains the regulated actor, making the central question whether reliance on AI-generated content affects the practitioner's own duties of competence, due diligence, and written advice (American Institute of Certified Public Accountants 2024; Brodeur, Hall, and Tynch 2023).

This creates a governance problem that is simultaneously practical and structural. Practically, practitioners face genuine uncertainty about when and how Circular 230 obligations apply to AI-assisted work. A practitioner who uses ChatGPT to draft a response to a client's tax question and transmits that response without significant modification may be providing "written advice" within the meaning of §10.37 yet receive no clear regulatory guidance on what that obligation requires. Structurally, even if practitioners violate Circular 230 by relying uncritically on AI outputs that may not satisfy professional responsibility expectations, the published OPR enforcement record may not reflect detection or sanctioning of such conduct through existing mechanisms.

This study investigates both the enforcement gap and its practical dimensions through a two-phase research design. The following research questions guide the analysis:

RQ1: To what extent do published IRS OPR disciplinary actions against CPAs and Enrolled Agents from 2020 through 2025 cite the Circular 230 provisions, §10.22, §10.35, and §10.37, most relevant to the quality of tax advisory practice?

RQ2: To what extent do outputs generated by publicly accessible AI tools for tax advisory scenarios align with Circular 230-informed professional responsibility expectations when evaluated against a structured compliance rubric?

Phase 1 addresses RQ1 through a content analysis of 187 OPR disciplinary actions. Phase 2 addresses RQ2 through structured evaluation of 30 AI-generated advisory outputs across ten scenarios and three tools. The combination of these two phases connects enforcement patterns to advisory practice implications in a way that neither phase alone could accomplish.

Phase 1 reveals that zero published disciplinary actions over six years cite §10.22, §10.35, or §10.37, while §10.82(b) expedited suspension accounts for 78.1 percent of all cases. Phase 2 reveals that AI tools showed recurring weaknesses on related written-advice safeguards, particularly on disclaimer adequacy (mean 1.70/4.0) and professional referral (mean 1.63/4.0), despite achieving reasonably strong tax law accuracy scores (mean 3.33/4.0). Together, these findings suggest a potential area where conduct-related risks under Circular 230 may not be reflected in published enforcement outcomes.

This study makes four contributions. First, it provides an empirical analysis of OPR disciplinary records in the context of AI-assisted tax practice, characterizing the observable pattern of published enforcement rather than relying solely on regulatory text. Second, it develops and applies a structured Circular 230 compliance rubric for AI-generated tax advisory outputs, a structured framework that can be examined and refined in future research. Third, it applies regulatory lag theory to the tax practice governance context, generating testable predictions and evidence-based implications for professional guidance and future research. Fourth, it provides practice-relevant implications for tax practitioners, firms, educators, and standard-setters using AI-assisted tax tools.

Ii. Background And Prior Literature

Regulatory Lag Theory and Professional Standards

Regulatory lag theory holds that regulatory frameworks, including the enforcement mechanisms that give them effect, are calibrated to the technological environment that existed when they were designed (Baldwin, Cave, and Lodge 2012). When a disruptive technology enters a regulated domain, the existing enforcement apparatus tends to be inadequate not because regulators are negligent, but because the technology has fundamentally changed the nature of the conduct the rules were designed to govern. The lag between technological change and regulatory adaptation is a structural feature of institutional design, not merely a temporary delay, and reflects the broader "pacing problem" in which legal and ethical oversight often develops more slowly than emerging technologies (Marchant 2011; Coglianese and Lehr 2017; Zittrain 2019).

Regulatory lag is not the only explanation for uneven enforcement patterns. The absence of particular Circular 230 provisions from published disciplinary actions may also reflect detection difficulty, evidentiary burdens, enforcement discretion, publication practices, prosecutorial priorities, or administrative efficiency. Practice-quality concerns involving competence, due diligence, or written advice may be more difficult to investigate and prove than event-triggered violations that rely on externally verifiable facts. More broadly, regulatory enforcement is shaped not only by substantive rules but also by institutional capacity, administrative priorities, and the practical observability of regulated conduct (Baldwin, Cave, and Lodge 2012). Accordingly, this study treats regulatory lag as a theoretical lens rather than a uniquely identified causal mechanism. The observed enforcement distribution is interpreted as consistent with regulatory lag, while recognizing that alternative institutional and procedural explanations may also contribute to the pattern.

Circular 230 and the Regulation of Tax Practice

IRS Circular 230 (31 C.F.R. Part 10) is the foundational regulatory framework governing the conduct of practitioners who practice before the Internal Revenue Service.[1] Circular 230 is issued under Treasury's statutory authority to regulate practice before the Department of the Treasury, including practice before the IRS. Its provisions establish affirmative obligations and

[1]IRS Circular 230 (31 C.F.R. Part 10) governs practice before the Internal Revenue Service. The four provisions most central to this study are §10.22 (due diligence), §10.35 (competence), §10.37 (written advice standards), and §10.82 (expedited suspension proceedings).

disciplinary grounds across a wide range of practice activities. Four provisions are particularly relevant to the governance of AI-assisted tax advice.

Section 10.22 requires practitioners to exercise due diligence in preparing tax returns and other documents submitted to the IRS, and in determining the correctness of representations made to the IRS. Section 10.35 requires practitioners to possess the legal knowledge, skill, thoroughness, and preparation necessary for the representation. Section 10.37 establishes standards for written tax advice, including advice communicated by electronic means, requiring that advice be based on a reasonable analysis of applicable law and facts, avoid unreasonable assumptions, and clearly identify situations in which the legal analysis is subject to uncertainty. Section 10.82 authorizes expedited suspension based on externally verifiable disciplinary triggers, including certain criminal convictions, professional license suspension or revocation, court sanctions, certain federal tax noncompliance patterns, and violations of prior disciplinary conditions (Internal Revenue Service 2014).

The 2014 amendments to Circular 230 replaced the prior "covered opinion" framework with the more flexible, principles-based §10.37 standard. The 2014 final regulations eliminated the former covered-opinion rules and placed written tax advice under the principles-based §10.37 standard. Although this change simplified the prior framework, it left the application of written-advice obligations dependent on a facts-and-circumstances “reasonable practitioner” standard, creating uncertainty that becomes more significant when AI-generated content is incorporated into client advice (Department of the Treasury 2014; Hatfield 2021). That uncertainty is compounded in the AI context, where the practitioner may not be the author of the substantive content being communicated. In this context, the relevant question is not whether the AI output is labeled as formal advice, but whether the practitioner uses, adopts, or transmits AI-generated content as client-facing written tax advice. The 2024 revision of the AICPA Statements on Standards for Tax Services introduced new language on technology use but stopped short of specific guidance on AI-generated content, leaving the interpretive gap unaddressed (American Institute of Certified Public Accountants 2024).

AI in Tax and Accounting Practice: Emerging Literature

Empirical research on AI in professional accounting is nascent but growing. Recent work shows that large language models can perform certain financial statement analysis tasks in ways that resemble professional analyst judgment, although their usefulness depends on task design, input quality, and the need for human review (Kim, Muhn, and Nikolaev 2024). Cao, Jiang, Yang, and Zhang (2023) further show that AI and machine-reading technologies are influencing the financial reporting environment by affecting how firms prepare disclosures for machine processing. Earley (2015) similarly shows that data analytics in auditing creates both opportunities and implementation challenges, particularly where professional judgment, data quality, and oversight remain necessary. Together, this literature suggests that AI can support professional accounting tasks, but its use does not eliminate the need for practitioner review, professional judgment, and accountability**.** In the tax domain specifically, VanDenburgh, Tribou, and Braswell (2025) evaluated AI tool accuracy across several tax topics and found variable performance, underscoring concerns about practitioner overreliance without adequate review. Jenkins (2024) analyzed how Circular 230 and the revised SSTSs apply to generative AI, concluding that ambiguity remains regarding practitioner responsibility for AI-generated outputs. Brodeur, Hall,

and Tynch (2023) similarly identify tension between AI's accessibility and practitioners' professional accountability obligations.

Related research on algorithmic accountability and human oversight further emphasizes that professional responsibility cannot be delegated entirely to automated systems. Accountability scholarship highlights the importance of transparency, documentation, and governance mechanisms that make AI-supported decisions subject to meaningful review (Percy et al. 2022). Research on human oversight also cautions that simply placing a human "in the loop" does not guarantee effective supervision, particularly when organizations rely on oversight as a substitute for stronger institutional controls (Green 2022). In accounting specifically, recent field evidence suggests that generative AI is most effective when it augments rather than replaces professional judgment, with accountants selectively intervening when AI confidence is low and greater error risk arising when users rely on non-consensus AI recommendations (Choi and Xie 2026). These perspectives are also consistent with arguments that professional domains should preserve meaningful human expertise and accountability rather than treating automated systems as substitutes for professional judgment (Pasquale 2020). Taken together, this literature complements regulatory lag theory by showing why practitioner verification, documentation, and independent judgment remain central even as AI-generated professional content becomes more sophisticated.

Taken together, prior studies have primarily examined AI accuracy, adoption, human oversight, or professional responsibility in isolation; they have not empirically connected published Circular 230 enforcement patterns with the professional-responsibility characteristics of AI-generated tax advisory outputs. This study addresses that gap through its two-phase design.

Judicial Developments and Professional Accountability

Recent judicial decisions underscore the urgency of examining AI's intersection with professional standards. In Clinco v. Commissioner (2026), the U.S. Tax Court addressed the submission of fabricated legal authorities likely generated by artificial intelligence, emphasizing that practitioners retain personal responsibility for verifying the accuracy of all citations and legal analysis regardless of the tools used. In United States v. Heppner (2026), a federal district court considered whether the use of a third-party generative AI platform implicated attorney-client privilege and work-product protections, reflecting growing judicial scrutiny of AI integration into professional workflows. These cases are relevant to this study not because they directly define Circular 230 obligations, but because they show that courts are treating unverified AI-assisted legal and tax work as a professional accountability issue. Although neither decision interprets Circular 230 directly, both signal that courts expect practitioners to exercise independent professional judgment that cannot be delegated to automated systems. This judicial development creates an important asymmetry: practitioners face professional accountability in adversarial proceedings that their primary regulatory framework, OPR enforcement, may not readily capture through routine published disciplinary channels.

The Enforcement Role of the Office of Professional Responsibility

The IRS Office of Professional Responsibility administers Circular 230 and publishes disciplinary actions in the Internal Revenue Bulletin, making the disciplinary record publicly accessible (Internal Revenue Service, Office of Professional Responsibility 2020–2025). This transparency is significant: the published record constitutes the observable enforcement signal that practitioners, educators, and regulators receive about which Circular 230 violations actually result

in consequences. Event-based discipline is easier to observe and document than advice-quality failures, which require review of facts, client communications, legal authorities, and practitioner judgment. Prior scholarship has examined the historical expansion and subsequent judicial narrowing of Treasury's authority under Circular 230 (Hatfield 2021), but empirical analysis of provision-level enforcement patterns across a multi-year OPR sample remains limited in the published literature. This study addresses that gap by examining whether the observed provision-level enforcement pattern is consistent with a regulatory lag interpretation in which event-triggered mechanisms remain more visible than practice-quality provisions.

Iii. Research Methodology

Phase 1: Archival Content Analysis of OPR Disciplinary Actions

The first phase employs a content analysis of publicly available OPR disciplinary actions published in the Internal Revenue Bulletin from January 2020 through December 2025.[2] The sample period begins in 2020 to provide a meaningful pre-AI baseline and extends through 2025 to capture the period of widespread AI adoption following ChatGPT's public release in late 2022. Data were drawn from two sources: (1) the IRS OPR disciplinary look-up file, which provides practitioner names, professional designations, sanction types, and effective dates; and (2) the Internal Revenue Bulletin disciplinary announcements, which identify the specific Circular 230 provisions cited as the basis for each sanction. The analysis therefore captures the published sanction record rather than the full universe of OPR complaints, investigations, informal resolutions, or matters closed without public discipline.

The sample is restricted to disciplinary actions involving Certified Public Accountants and Enrolled Agents, including practitioners holding dual CPA/EA designations and those holding both CPA and attorney credentials (reclassified as CPA for sample purposes). Actions involving attorneys only, appraisers, and unenrolled return preparers are excluded because the study focuses on CPA and Enrolled Agent tax practitioners, and attorney-only cases may involve separate state bar, court-based, and legal-professional responsibility frameworks that differ from the CPA/EA practice context. The final Phase 1 sample comprises 187 disciplinary actions.

Because the coding was conducted by a single researcher, intercoder reliability statistics such as Cohen's kappa or Krippendorff's alpha (Krippendorff 2019) could not be calculated. To reduce interpretive subjectivity, the primary Phase 1 coding variable was based on the specific Circular 230 provision expressly cited in each published disciplinary announcement rather than on researcher inference regarding provisions that might otherwise have been implicated. Accordingly, the central finding concerning the frequency of cited provisions relies primarily on direct extraction from the published disciplinary record. Nevertheless, the absence of independent coding remains a methodological limitation, particularly for judgment-based variables such as the AI relevance classification.

The coding approach follows established content-analysis principles emphasizing explicit coding categories, transparent operational definitions, and systematic application of coding rules (Neuendorf 2017). Each case was systematically coded for: (1) the Circular 230 section(s) cited as the basis for the sanction; (2) the type of sanction imposed; (3) the professional designation of the practitioner; and (4) an AI relevance score (1=low, 2=medium, 3=high) reflecting the degree to

[2]The IRS OPR disciplinary look-up file is publicly available at https://www.irs.gov/tax-professionals/search-for-disciplined-tax-professionals. Disciplinary announcements are published in the Internal Revenue Bulletin at https://www.irs.gov/irb.

which the violation type implicates risks associated with AI-assisted tax advisory work. The AI relevance classification was operationalized as follows: low relevance indicated provisions primarily associated with misconduct or disciplinary events not directly connected to the quality of AI-assisted advisory work; medium relevance indicated provisions that could become relevant in AI-assisted practice but were not primarily directed at advisory quality; and high relevance indicated provisions directly governing practitioner competence, due diligence, or written tax advice, including §§10.22, 10.35, and 10.37. The classification was used descriptively to organize provision types and was not treated as an independently validated measurement scale. The purpose of Phase 1 is not to identify AI-related violations, because published enforcement records do not specify AI use, but to characterize the observable pattern of Circular 230 enforcement reflected in published disciplinary actions. Because the primary provision-coding variable was based on the Circular 230 section expressly cited in the published disciplinary announcement, the central Phase 1 finding involved limited interpretive judgment. Each case was coded based on the provision expressly identified in the published announcement rather than provisions that may have been conceptually implicated but not cited as a sanction basis.

Phase 2: Evaluation of AI-Generated Tax Advisory Outputs

The second phase constructs ten tax advisory scenarios based on Phase 1 violation categories and evaluates AI-generated responses against a structured Circular 230 compliance rubric.[3] Scenarios were designed as generalized hypothetical fact patterns spanning topics that represent common areas of practitioner advisory work and that map to the Phase 1 violation categories: home office deductions (§10.22), cryptocurrency staking rewards (§10.35), S-corporation reasonable compensation (§10.37), hobby loss rules (§10.22), foreign account reporting (§10.35), undocumented deduction claims (§10.22), syndicated conservation easements (§10.37), Employee Retention Credits (§10.35), EITC due diligence (§10.22), and NFT tax treatment (§10.37/§10.35). The full scenario prompts are available in supplemental materials.

To ensure consistency across tools, each scenario was submitted using an identical prompt structure, presented as a practitioner question without leading language that might signal expected disclaimers or referrals. Each scenario was submitted to ChatGPT (GPT-5.3 Instant; OpenAI 2026), Microsoft Copilot, and Google Gemini (as accessed in March 2026), three widely used general-purpose AI platforms accessible to tax practitioners without specialized tax software integration.[4] The study does not extend to proprietary enterprise tax research platforms with embedded compliance safeguards; findings accordingly reflect baseline AI advisory behavior. All submissions were made in cleared sessions during a two-week collection window in March 2026. For each scenario-tool combination, the first complete response generated in the cleared session was retained and scored; outputs were not regenerated.

Because the study is exploratory and the Phase 2 sample consists of a purposively constructed set of 30 AI-generated outputs rather than a probability sample, the analysis is

[3]The compliance rubric was developed through three stages: (1) a deductive stage mapping the core regulatory dimensions to specific Circular 230 provisions and identifying supplementary professional-responsibility dimensions relevant to AI-assisted advisory quality; (2) a piloting stage applying the rubric to three practice scenarios not included in the final sample; and (3) a calibration stage reviewing pilot scores and refining dimension definitions to ensure consistent application.

[4]Scenarios were submitted to each AI tool using a new, cleared session to eliminate context carryover. All data were collected within a two-week window in March 2026 to minimize version variability across sessions.

descriptive. Means, composite scores, and scenario-level comparisons are used to characterize patterns across tools and rubric dimensions; inferential statistical tests are not used to generalize these results to the broader population of AI-generated tax advice.

Each output was systematically evaluated on seven dimensions. Three dimensions map directly to Circular 230 provisions: (1) §10.22 due diligence, did the response demonstrate appropriate investigative care before rendering advice?; (2) §10.35 competence, did the response reflect accurate and current knowledge of applicable law?; and (3) §10.37 written advice standards, did the response reflect the principles-based requirements applicable to written tax advice? Four supplementary dimensions capture professional-responsibility safeguards relevant to AI-assisted advisory quality: (4) tax law accuracy, were the specific legal citations, thresholds, and rules stated correctly?; (5) disclaimer adequacy, did the response include appropriate caveats about the limitations of the advice?; (6) appropriate caveats, did the response flag uncertainty, audit risk, or areas requiring additional facts?; and (7) referral to a professional, did the response recommend consulting a qualified tax professional where appropriate? Section 10.37 requires practitioners to consider applicable law and authorities in relation to the relevant facts, although it does not necessarily require every authority to be expressly cited in the written advice. Accordingly, failure to identify controlling authority is treated here as an indicator of weaker written-advice quality rather than, by itself, a determination of §10.37 noncompliance. Each dimension was scored on a four-point scale (1 = non-compliant to 4 = fully compliant), yielding a composite maximum score of 28. Detailed scoring anchors for each dimension are available in the supplemental materials. The rubric was developed as a structured evaluative framework derived from Circular 230-informed professional responsibility expectations rather than as a psychometrically validated scale. The pilot evaluation of three excluded scenarios was used to clarify scoring anchors, improve consistency in applying the dimensions, and identify ambiguous criteria before formal scoring began. No independent expert-panel validation, construct validation, or criterion validation was conducted. Accordingly, the rubric scores should be interpreted as structured researcher judgments regarding alignment with the specified professional responsibility dimensions rather than as direct legal determinations of compliance. Because all Phase 2 outputs were scored by a single researcher, evaluator bias cannot be fully excluded despite the use of predefined scoring anchors and pilot calibration. Composite scores of 22–28 are rated Fully Compliant, 15–21 Partially Compliant, 8–14 Non-Compliant, and 7 Severely Non-Compliant. These thresholds were used as descriptive interpretive bands to summarize relative levels of rubric alignment and were not intended to establish legally validated compliance categories.

IV. Results: Phase 1, OPR Enforcement Patterns (2020–2025)

Sample Overview and Annual Distribution

The Phase 1 sample comprises 187 disciplinary actions against CPAs and Enrolled Agents across six years (Table 1). Disciplinary volume does not show an observable increase after the public release of widely used generative AI tools; cases in 2023 and 2024 are lower than the pre-AI baseline years of 2020 and 2021. This pattern is consistent with the regulatory lag prediction that published enforcement volume is driven primarily by externally verifiable disciplinary triggers rather than emerging technology-related practice risks.

TABLE 1

Annual Distribution of OPR Disciplinary Actions Against CPAs and Enrolled Agents (2020–2025)

Year	Cases (n)	% of Total	Cumulative %
2020	30	16.0%	16.0%
2021	36	19.3%	35.3%
2022	34	18.2%	53.5%
2023	27	14.4%	67.9%
2024	23	12.3%	80.2%
2025	37	19.8%	100.0%
Total	**187**	**100.0%**	—

Note: Sample restricted to CPAs and Enrolled Agents. ChatGPT achieved widespread public adoption in late 2022, marking the beginning of AI-assisted tax advisory practice at scale.

Circular 230 Provisions Cited: The Central Finding

Table 2 presents the distribution of Circular 230 provisions cited across the 187 disciplinary actions, directly addressing RQ1. The results are consistent with the regulatory lag prediction: §10.82(b) expedited suspension accounts for 78.1 percent of all sanctions, the §10.51 disreputable conduct provisions account for 9.6 percent, and 23 cases (12.3 percent) cite no specific CFR provision.[5]

The answer to RQ1 is clear: zero cases across all 187 actions and all six years cite §10.22, §10.35, or §10.37. These are the three Circular 230 provisions most directly relevant to the quality of tax advisory practice and most implicated by AI-assisted advisory work. The absence of §10.22, §10.35, and §10.37 from the published sanctions reviewed over a six-year period that includes AI adoption at scale is the central empirical finding of Phase 1.

TABLE 2

Distribution of Circular 230 Provisions Cited in OPR Disciplinary Actions (2020–2025)

Circular 230 Provision	Cases (n)	% of Total	AI Relevance
§10.82(b), Expedited Suspension	146	78.1%	Medium
§10.51, Disreputable Conduct (all subsections)	18	9.6%	Low–Medium

[5]Expedited suspension under §10.82(b) allows OPR to suspend practitioners based on externally verifiable disciplinary triggers, including certain criminal convictions, professional license suspension or revocation, court sanctions, certain federal tax noncompliance patterns, and violations of prior disciplinary conditions. The practitioner receives notice and has 30 days to respond before suspension takes effect.

No CFR cited (reinstatements / insufficient detail)	23	12.3%	N/A
§10.22, Due Diligence	**0**	**0.0%**	**High**
§10.35, Competence	**0**	**0.0%**	**High**
§10.37, Written Advice Standards	**0**	**0.0%**	**High**
Total	**187**	**100.0%**	—

Note: AI Relevance is a descriptive classification based on each provision's conceptual connection to risks in AI-assisted tax advisory work. High-relevance provisions directly address competence, due diligence, or written advice; medium-relevance provisions may indirectly relate to AI-assisted practice; and low-relevance provisions primarily concern misconduct or disciplinary events not specific to advisory quality.

Sanction Types and Practitioner Designations

Suspensions account for 93.0 percent of sanctions, with disbarments (4.8 percent) and censures (0.5 percent) representing small fractions. The professional composition is predominantly CPA (86.6 percent), with Enrolled Agents comprising 12.3 percent and CPA/EA dual designees 1.1 percent. Geographic distribution spans 31 states, with California, New York, Georgia, and Texas accounting for the largest numbers of actions, consistent with the national distribution of registered tax practitioners.

V. Results: Phase 2, Ai Output Compliance Evaluation

Overall Compliance Scores by Tool

Table 3 presents mean compliance scores for each tool across the seven rubric dimensions and composite totals, addressing RQ2. The results reveal a consistent and important pattern: AI tools achieve relatively strong scores on tax law accuracy but receive lower scores on the dimensions most directly associated with Circular 230's written advice and conduct standards.

TABLE 3

Mean Circular 230 Compliance Scores by AI Tool (Phase 2, n = 30)

Dimension (Max = 4.0)	ChatGPT	Copilot	Gemini	All Tools Mean
§10.22 Due Diligence	3.0	3.1	3.1	3.07
§10.35 Competence	3.0	2.9	3.9	3.27
§10.37 Written Advice Standards	2.2	2.5	2.5	2.40
Tax Law Accuracy	3.1	2.9	4.0	3.33
Disclaimer Adequacy	1.4	2.1	1.6	1.70
Appropriate Caveats	2.2	2.6	3.2	2.67

Referral to Professional	1.3	2.1	1.4	1.63
Composite Score (Max = 28)	**16.2**	**18.3**	**19.7**	**18.07**
% Outputs Rated Fully Compliant	**10.0%**	**40.0%**	**30.0%**	**26.7%**
% Outputs Rated Non-Compliant	**30.0%**	**20.0%**	**0.0%**	**16.7%**

Note: Scores are means across ten scenarios per tool. Each dimension scored 1 (non-compliant) to 4 (fully compliant). Composite = sum of seven dimensions (maximum 28). Fully Compliant = 22–28; Partially Compliant = 15–21; Non-Compliant = 8–14. Bolded rows indicate dimensions with the lowest mean scores.

The Accuracy-Compliance Disconnect

The most significant finding in Table 3 is the divergence between tax law accuracy and conduct-based compliance, the pattern that connects Phase 2 to the regulatory lag argument. Google Gemini achieves a near-perfect mean accuracy score of 4.0 out of 4.0,[6] yet only 30 percent of its outputs are rated fully compliant. ChatGPT achieves a mean accuracy score of 3.1 yet only 10 percent of its outputs are fully compliant. This pattern, high accuracy, low compliance, is the central Phase 2 finding and its most important implication for regulatory design.

The compliance shortfall appears to arise primarily from the absence of conduct-supporting elements rather than from tax law errors. Disclaimer adequacy receives the lowest mean score (1.70/4.0), reflecting limited use of professional disclaimers across AI responses. Referral to a qualified professional receives the second-lowest mean score (1.63/4.0), and written advice standards receive a mean score of 2.40/4.0. These dimensions may strengthen the professional adequacy of AI-assisted written advice, particularly where facts are incomplete, the law is uncertain, or the taxpayer's situation requires individualized judgment. This alignment between the Phase 1 finding, that §10.37 does not appear in the published OPR sanctions reviewed in this study, and the Phase 2 finding, that AI outputs received lower scores on §10.37-related dimensions, is consistent with the study's interpretation of a potential regulatory alignment gap.

Scenario-Level Results

Table 4 presents composite scores by scenario and tool, revealing variation that carries additional interpretive significance.

[6]Gemini's higher accuracy score (4.0/4.0) reflects its performance across the ten scenarios evaluated in this study. In the scenarios reviewed, Gemini cited Rev. Rul. 2023-14 for cryptocurrency staking, IRS Notice 2023-27 for NFT treatment, and SECURE 2.0 Act limitations on syndicated conservation easement deductions when applicable.

TABLE 4

Composite Compliance Scores by Scenario and AI Tool (Max = 28)

Scenario	Target §	ChatGPT	Copilot	Gemini	Mean
SC-001: Home Office Deduction	10.22	13 (NC)	23 (C)	22 (C)	19.3
SC-002: Crypto Staking	10.35	14 (NC)	14 (NC)	17 (PC)	15.0
SC-003: S-Corp Compensation	10.37	14 (NC)	22 (C)	18 (PC)	18.0
SC-004: Hobby Loss Rules	10.22	15 (PC)	15 (PC)	18 (PC)	16.0
SC-005: FBAR Reporting	10.35	15 (PC)	22 (C)	18 (PC)	18.3
SC-006: Fraudulent Deductions	10.22	19 (PC)	18 (PC)	22 (C)	19.7
SC-007: Conservation Easement	10.37	22 (C)	23 (C)	21 (PC)	22.0
SC-008: Employee Retention Credit	10.35	15 (PC)	15 (PC)	18 (PC)	16.0
SC-009: EITC Due Diligence	10.22	18 (PC)	18 (PC)	24 (C)	20.0
SC-010: NFT Tax Treatment	10.37	17 (PC)	13 (NC)	19 (PC)	16.3

Note: C = Compliant (22–28); PC = Partially Compliant (15–21); NC = Non-Compliant (8–14). Composite scores = sum of seven rubric dimensions, each scored 1–4.

Three scenario-level patterns are interpretively significant. First, conservation easements (SC-007) produce the strongest results, with ChatGPT and Copilot both rated Compliant. This stands in marked contrast to other scenarios. One possible explanation is that extensive public enforcement and practitioner commentary surrounding syndicated conservation easements may make this topic more visible in publicly available materials.

Second, the NFT scenario (SC-010) produces the most consequential divergence. Copilot scores 13/28 (Non-Compliant), missing the collectibles 28-percent long-term capital gains rate that applies to certain NFTs under the IRS Notice 2023-27 look-through analysis. Gemini scores 19/28 (Partially Compliant) and explicitly cites Notice 2023-27. The same prompt generates materially different tax outcomes. A practitioner relying on the Copilot response might therefore advise applying the standard 15- or 20-percent rate to gains potentially subject to the 28-percent rate. This illustrates a potential §10.35 competence concern with practical economic consequences for clients.

Third, the crypto staking scenario (SC-002) illustrates a citation compliance gap independent of accuracy. Only Gemini cites Revenue Ruling 2023-14, the controlling IRS authority on staking taxability. ChatGPT and Copilot provide accurate characterizations without identifying the governing authority. Section 10.37 requires practitioners to consider applicable law and authorities in relation to the relevant facts, although it does not necessarily require every

authority to be expressly cited in the written advice. Accordingly, failure to identify controlling authority is treated here as an indicator of weaker written-advice quality rather than, by itself, a determination of §10.37 noncompliance.

Vi. Discussion

The Structural Regulatory Gap: A Regulatory Lag Interpretation

The findings are consistent with the possibility of regulatory lag across both phases, although alternative explanations should also be considered. Phase 1 shows that published OPR enforcement is heavily concentrated in §10.82(b), a mechanism associated with externally observable events, while the practice-quality provisions most relevant to AI-assisted tax advisory work do not appear as cited sanction bases in the published sanctions reviewed. This pattern may reflect regulatory lag, but it may also be influenced by detection difficulty, evidentiary burdens, enforcement discretion, publication practices, administrative priorities, or other institutional factors. Phase 2 shows that AI tools received lower scores on §10.37-related written-advice safeguards, particularly disclaimer adequacy, appropriate caveats, and professional referral. Taken together, the findings are consistent with a potential regulatory alignment gap, but they do not establish that regulatory lag is the sole cause of the observed enforcement pattern.

The divergence between relatively strong tax-law accuracy scores and weaker scores on professional responsibility safeguards provides additional evidence of this potential alignment gap. General-purpose AI tools often produced direct and substantively accurate responses while receiving lower scores on disclaimers, caveats, and professional referral. This pattern suggests that technical accuracy alone may not ensure alignment with the broader professional responsibility expectations evaluated in this study. Practitioners who rely on AI-generated outputs therefore remain responsible for independently reviewing the accuracy, completeness, and appropriateness of the advice before communicating it to clients. The findings should not be interpreted as demonstrating practitioner noncompliance or a causal deficiency in OPR enforcement, but rather as identifying an area in which existing professional standards, published enforcement patterns, and AI-assisted advisory practices may not yet be fully aligned.

Institutional theory helps explain why this gap may persist without further institutional adaptation. OPR's §10.82(b) enforcement model is deeply institutionally embedded: staff expertise, evidentiary standards, procedural routines, and administrative infrastructure are all calibrated to it. Greater attention to practice-quality enforcement may require institutional adaptation, including changes in evidentiary processes, staffing expertise, and enforcement priorities. DiMaggio and Powell’s (1983) institutional isomorphism framework suggests that such adaptation may be more likely when external pressures make emerging risks visible to institutional leaders and the profession they serve.

Implications for Tax Practitioners

For individual practitioners, these findings create an uncomfortable asymmetry. The near-absence of §10.22, §10.35, and §10.37 enforcement might suggest that the practical risk of discipline for AI-related advisory failures is currently low. But the judicial developments, Clinco and Heppner, signal that professional accountability for AI-assisted work is actively developing in adversarial proceedings even as OPR enforcement lags. A practitioner who is sued by a client for

reliance on AI-generated advice that proved non-compliant cannot point to the absence of OPR enforcement as a defense.

The Phase 2 findings offer practical considerations for practitioners who incorporate AI-generated outputs into tax advisory work. Disclaimer adequacy and professional referral are the lowest-scoring dimensions across AI outputs, averaging 1.70 and 1.63 out of 4.0 respectively. These are also among the easiest elements to supplement: although disclaimers and referral language may be easier to add than substantive legal review, they do not replace the practitioner's responsibility to evaluate the accuracy and appropriateness of the AI-generated analysis. The more difficult, and more consequential, task is evaluating whether the AI's substantive analysis is appropriate for the topic's complexity, as the NFT and crypto staking scenarios illustrate. Topics involving recent or unsettled IRS guidance require heightened scrutiny because AI tools exhibit different degrees of competence on the same question and may not reflect current authorities.

Implications for Standard-Setters and Regulators

For the IRS Office of Professional Responsibility, these findings suggest potential value in clarifying how existing Circular 230 provisions apply to AI-assisted written tax advice. Interim guidance from OPR could help clarify how §10.37 applies to AI-assisted written advice and what conduct elements practitioners may need to consider before transmitting AI-generated content to clients, such as disclaimers, appropriate caveats, authority citation, and professional referral language. Such interpretive guidance may help reduce practitioner uncertainty without requiring immediate formal rulemaking.

For the AICPA, the 2024 SSTS revision provides a useful foundation, but additional interpretation may be helpful as AI-assisted tax advisory practices continue to evolve. A targeted SSTS interpretation addressing the disclaimer, review, and referral considerations associated with AI-assisted written tax advice may help clarify practitioner responsibilities without requiring a full revision of the standards. For accounting educators, the compliance rubric developed in Phase 2 offers a ready-made pedagogical resource. The ten scenarios can be used directly in tax ethics and professional responsibility courses to develop students' critical evaluation skills and illustrate how Circular 230 provisions apply to AI-generated outputs.

Vii. Limitations And Future Research

Several limitations of this study should be acknowledged. First, the Phase 1 sample is restricted to published disciplinary announcements reflecting sanctions imposed. OPR investigations closed without action, cases resolved through informal guidance, and complaints not pursued are not observable in the data. The published record reflects only the outcomes of OPR's enforcement funnel. Future research with access to OPR administrative data, including investigations initiated but not concluded in public sanctions, would provide a more complete picture of the enforcement pipeline.

Second, Phase 1 coding was conducted by a single researcher; intercoder reliability was not formally assessed. Future research incorporating independent coding and inter-rater reliability assessment would strengthen the archival analysis.

Third, the Phase 2 sample covers ten scenarios and three tools at a single point in time (March 2026). AI model behavior changes continuously; findings should not be treated as definitive characterizations of current capabilities.

Fourth, the compliance rubric involves interpretive judgments. While dimensions are grounded in regulatory text, scoring reflects the single researcher's assessment and different scorers may reach different conclusions on borderline responses.

Fifth, Phase 2 evaluates AI-generated tax advisory outputs rather than the behavior of practicing CPAs or enrolled agents. The results therefore should not be interpreted as evidence of actual practitioner noncompliance, disciplinary exposure, or client harm. In practice, professionals may revise, verify, supplement, or reject AI-generated content before communicating advice to clients. Accordingly, the Phase 2 findings are best understood as evidence about the characteristics of AI-generated outputs when evaluated against the study's Circular 230-informed rubric, not as direct evidence of practitioner conduct.

Future research could extend this work through larger scenario sets, multiple coders, longitudinal testing across model versions, and practitioner survey research examining how CPAs and EAs currently navigate AI-related Circular 230 obligations in practice.

Viii. Conclusion

This study identifies a potential regulatory alignment gap between Circular 230's professional conduct provisions and OPR's published enforcement record in the context of AI-assisted tax advisory work. Across 187 published OPR actions over six years, no case cites §10.22, §10.35, or §10.37, while the AI tools evaluated in Phase 2 received their lowest average scores on written-advice safeguards, particularly disclaimer adequacy and professional referral. These patterns are consistent with the possibility that emerging AI-assisted advisory risks are not yet fully reflected in the published enforcement record. However, the findings do not establish an enforcement deficiency, practitioner noncompliance, or a causal relationship between OPR enforcement patterns and AI-generated advisory quality.

These findings suggest that additional regulatory and professional guidance may be useful as AI-assisted tax advisory practices continue to evolve. Clarification regarding the application of §10.37 to AI-assisted written advice could help practitioners better understand their responsibilities when reviewing, modifying, or transmitting AI-generated content. Similar guidance from professional bodies such as the AICPA may also support consistency in practice. While formal rulemaking may take time, interim interpretive guidance could reduce practitioner uncertainty and provide a more practical bridge between existing professional standards and emerging AI-assisted workflows. This study provides an empirical foundation for that ongoing conversation. Future research can extend this analysis by examining practitioner behavior, expanding the AI scenario set, and assessing whether professional guidance changes how AI-assisted tax advice is reviewed and communicated.

Data Availability Statement

The Phase 1 coding instrument, coded data, and code reference guide are available as an online supplement to this article. The Phase 2 tax advisory scenarios, AI-generated responses (archived with collection dates from March 2026), and compliance rubric scoring data are available as a time-stamped supplemental appendix. Because AI-generated outputs are not reproducible given continuous model version changes, the archived responses represent a point-in-time record of tool behavior in March 2026. Supplemental materials will be made available through the journal's approved data repository upon acceptance and from the corresponding author upon reasonable request prior to acceptance.

References

American Institute of Certified Public Accountants. 2024. *Statements on Standards for Tax Services (Revised).* New York, NY: AICPA. https://www.aicpa.org.

Baldwin, R., M. Cave, and M. Lodge. 2012. *Understanding Regulation: Theory, Strategy, and Practice.* 2nd ed. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:osobl/9780199576081.001.0001

Brodeur, G. L., G. Hall, and E. Tynch. 2023. ChatGPT for legal and tax professionals. The CPA Journal, July/August. https://www.cpajournal.com/2023/10/16/chatgpt-for-legal-and-tax-professionals-2/

Cao, S., W. Jiang, B. Yang, and A. L. Zhang. 2023. How to talk when a machine is listening: Corporate disclosure in the age of AI. *The Review of Financial Studies* 36 (9): 3603–3642. https://doi.org/10.1093/rfs/hhad021

Choi, J. H., and C. L. Xie. 2026. Human + AI in accounting: Early evidence from the field. *Journal of Accounting Research* 64 (3): 1333–1373. https://doi.org/10.1111/1475-679X.70052

Clinco v. Commissioner, T.C. Memo. 2026-16, No. 8077-23 (U.S. Tax Ct. Feb. 9, 2026). Available at: https://case-law.vlex.com/vid/clinco-v-comm-r-1105094064

Coglianese, C., and D. Lehr. 2017. Regulating by robot: Administrative decision making in the machine-learning era. *Georgetown Law Journal* 105 (5): 1147–1223. https://scholarship.law.upenn.edu/faculty_scholarship/1734/

Department of the Treasury. 2014. Regulations Governing Practice Before the Internal Revenue Service. Treasury Decision 9668, 79 Fed. Reg. 33685–33695. https://www.federalregister.gov/documents/2014/06/12/2014-13739/regulations-governing-practice-before-the-internal-revenue-service

DiMaggio, P. J., and W. W. Powell. 1983. The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review* 48 (2): 147–160. https://doi.org/10.2307/2095101

Earley, C. E. 2015. Data analytics in auditing: Opportunities and challenges. *Business Horizons* 58 (5): 493–500. https://doi.org/10.1016/j.bushor.2015.05.002

Green, B. 2022. The flaws of policies requiring human oversight of government algorithms. Computer Law & Security Review 45: 105681. https://doi.org/10.1016/j.clsr.2022.105681

Hatfield, M. 2021. The Rise of Law and the Fall of Circular 230: Tax Lawyer Professional Standards, 1985–2015. Florida Tax Review 24 (2): 828–930. https://scholarship.law.ufl.edu/ftr/vol24/iss2/10

Internal Revenue Service, Office of Professional Responsibility. 2020–2025. Disciplinary actions and sanctions. *Internal Revenue Bulletin.* https://www.irs.gov/irb.

Internal Revenue Service. 2014. *Regulations Governing Practice Before the Internal Revenue Service (Treasury Department Circular No. 230, 31 C.F.R. Part 10).* Washington, DC: U.S. Government Publishing Office.

Jenkins, E. R. 2024. Tax ethics and use of generative AI systems. *The Tax Adviser* 55 (2): 42. Available at: https://www.thetaxadviser.com/issues/2024/feb/tax-ethics-and-use-of-generative-ai-systems/

Kim, A. G., M. Muhn, and V. V. Nikolaev. 2024. Financial statement analysis with large language models. *arXiv*. https://doi.org/10.48550/arXiv.2407.17866

Krippendorff, K. 2019. *Content analysis: An introduction to its methodology*. 4th ed. Thousand Oaks, CA: SAGE. https://doi.org/10.4135/9781071878781

Marchant, G. E. 2011. The growing gap between emerging technologies and the law. In *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem*, edited by G. E. Marchant, B. R. Allenby, and J. R. Herkert, 19–33. Dordrecht: Springer. https://doi.org/10.1007/978-94-007-1356-7_2

Neuendorf, K. A. 2017. *The content analysis guidebook*. 2nd ed. Thousand Oaks, CA: SAGE. https://doi.org/10.4135/9781071873045

OpenAI. 2026. GPT-5.3 Instant: Smoother, more useful everyday conversations. OpenAI, March 3. https://openai.com/index/gpt-5-3-instant/

Pasquale, F. 2020. *New Laws of Robotics: Defending Human Expertise in the Age of AI.* Cambridge, MA: Harvard University Press. https://doi.org/10.2307/j.ctv3405w6p

Percy, C., S. Dragicevic, S. Sarkar, and A. d'Avila Garcez. 2022. Accountability in AI: From principles to industry-specific accreditation. *AI Communications* 34 (3): 181–196. https://doi.org/10.3233/AIC-210080

Thomson Reuters Institute. 2025. 2025 Generative AI in professional services report. Thomson Reuters. https://www.thomsonreuters.com/en/institute/reports/2025-generative-ai-in-professional-services-report

United States v. Heppner, No. 25-cr-00503 (JSR), 2026 WL 436479 (S.D.N.Y. Feb. 17, 2026).

VanDenburgh, W. M., K. J. Tribou, and J. M. Braswell. 2025. Assessing AI from a tax perspective, part 2. Today's CPA, March/April. https://www.tx.cpa/news-publications/todays-cpa-magazine/issues/article/march-april-2025/2025/06/09/assessing-ai-from-a-tax-perspective--part-2.

Zittrain, J. 2019. The hidden costs of automated thinking. *The New Yorker* (July 23). https://www.newyorker.com/tech/annals-of-technology/the-hidden-costs-of-automated-thinking.