Open Science
BRAZIL

DOI: https://doi.org/10.37497/opsbrazil.51

PREPRINT

# KEEPING THE HUMAN IN THE LOOP: FAILURE POINTS OF GENERATIVE AI IN SMALL-BUSINESS ACCOUNTING AND TAX COMPLIANCE

## O humano no circuito da IA: pontos de falha da inteligência artificial generativa na contabilidade e conformidade tributária de pequenas empresas

EDSON TELES VICTOR
Founder, Wise Tax and Wise Bookks, Inc. (Florida, USA)
IRS Certifying Acceptance Agent; Member, National Association of Tax Professionals (NATP); Certified Scrum Product Owner (CSPO), Scrum Alliance

**ABSTRACT |** Generative AI is becoming part of everyday accounting and tax work in small firms and small businesses, and its productivity gains are real. This paper argues that those gains hold only when a qualified person stays in the loop. Drawing on the author's practice as founder of a U.S. tax firm and IRS Certifying Acceptance Agent, and on the professional and legal literature, it identifies four failure points that appear when people are removed or bypassed. The first is classification error: fluent output that assigns transactions to the wrong account or tax category. The second is advisory over-reliance: owners treat an LLM's answer as authoritative and act on it without seeing its assumptions, sometimes triggering IRS notices and penalties. The third is taxpayer data exposure: returns and statements containing Social Security numbers uploaded to general-purpose tools. The fourth is a compliance and accountability gap under IRC §7216, the FTC Safeguards Rule, IRS Publication 4557 and Circular 230. Cases from the author's practice, the abrupt closure of an online bookkeeping provider and an exploratory run of 26 questions on three free chatbots illustrate each point. In that single run per tool, given the same 60 real bank transactions, the chatbots produced a refusal, a fourfold profit range and a cash-flow total presented as profit; the demonstration shows how the failures arise and does not estimate error rates or rank the tools. The paper closes with human-in-the-loop principles: where review is indispensable, how to minimize the data a tool sees, and how to document who checked what.

**Keywords |** Generative Ai; Large Language Models; Human-In-The-Loop; Accounting; Tax Compliance; Small Business; Taxpayer Data Protection; Irc §7216; Automation Bias.

**Submetido:** 05-11-2026
**Postado:** 05-11-2026
**e-ISSN:** 2764-9822

**Endereço de correspondência:**
Edson Teles Victor

CC BY

**RESUMO |** A inteligência artificial generativa está cada vez mais presente nas atividades contábeis e tributárias de pequenas empresas, proporcionando ganhos de produtividade, mas também introduzindo riscos quando utilizada sem supervisão profissional qualificada. Este artigo analisa quatro pontos críticos associados à utilização de modelos de linguagem de grande escala (LLMs): erros de classificação contábil, confiança excessiva em orientações automatizadas, exposição de dados confidenciais dos contribuintes e lacunas de conformidade e responsabilização diante da legislação norte-americana. A pesquisa adota uma abordagem conceitual, fundamentada na literatura científica, na experiência profissional do autor e em estudos de caso ilustrativos. Inclui uma demonstração exploratória realizada com três ferramentas de inteligência artificial generativa, submetidas individualmente a 26 questões, incluindo a classificação de 60 transações bancárias reais. Os resultados evidenciam divergências relevantes nas respostas, incluindo ausência de cálculo conclusivo, estimativas incompatíveis de lucro e resultados numericamente próximos obtidos por procedimentos inadequados. A análise também aborda exigências regulatórias relacionadas à proteção de informações tributárias, especialmente o IRC §7216, a FTC Safeguards Rule e as orientações do Internal Revenue Service (IRS). Conclui-se que a supervisão humana qualificada, a minimização de dados, a documentação dos procedimentos e a revisão das decisões automatizadas são essenciais para conciliar eficiência tecnológica, segurança da informação e conformidade tributária. Por seu caráter exploratório, o estudo não estabelece taxas de erro nem classificações comparativas de desempenho entre as ferramentas.

**Palavras-chave |** Inteligência artificial generativa, Supervisão humana, Contabilidade, Conformidade tributária, Proteção de dados dos contribuintes

## 1. INTRODUCTION

The question for accounting is no longer whether to use generative AI but how; the paper's claim is that the tools work when a qualified person stays in the loop, and fail in predictable ways when that person is removed or bypassed.

### Context

Small firms (1–20 staff) and their clients are adopting general-purpose LLMs quickly, pushed by a shortage of accountants. Many use personal accounts and no written policy. Owners also use the same tools directly, as a free tax adviser, and often act before any professional sees the answer.

### The claim

The value of the tools is real. The failures come from the missing reviewer: model limits a trained person would catch flow straight into books, returns and business decisions.

### Gap

The literature covers AI in audit at large firms and LLM performance on exam questions. Little addresses routine small-business work, owners as direct users, or the legal duties of those who handle taxpayer data.

### Research questions

1. Where do general-purpose LLMs fail in small-business accounting and tax work?
2. Why do those failures persist when no qualified person reviews the output?
3. What does U.S. law already require of anyone who uses these tools with taxpayer data?
4. Which human-in-the-loop principles keep the gains while containing the failures?

### Contributions

1. A map of four failure points of generative AI in small-business accounting and tax work.
2. A mapping of those failure points to existing U.S. duties (IRC §7216, FTC Safeguards Rule, IRS Publication 4557, Circular 230).
3. Illustrative cases from practice, including the closure of an online bookkeeping provider.
4. Human-in-the-loop principles sized for firms and owners with no compliance staff.

### Method

Conceptual analysis grounded in the author's professional practice and the literature. Cases come from the author's practice, anonymized, and a single exploratory run of 26 questions on each of three free chatbots illustrates the failure points (Section 4); it does not measure error rates or compare the performance of the models.

### Structure

Section 2 reviews adoption; Section 3 sets out the four failure points; Section 4 the cases; Section 5 the principles; Section 6 discussion and limitations.

## 2. BACKGROUND: GENERATIVE AI IN SMALL ACCOUNTING FIRMS

Adoption is fast and largely informal. In a 2025 survey of about 1,800 professionals, the Thomson Reuters Institute found that the share of tax firms using generative AI across the organization rose from 8% to 21% in a year, and 79% expected significant integration by 2027 (Thomson Reuters, 2025). The most common uses were tax research (77%), return preparation (63%) and tax advisory work (62%). In small firms, typical uses include categorizing bank transactions, drafting client emails, summarizing notices and reading uploaded returns and statements, often through personal accounts. The push comes partly from a tight labor supply: U.S. schools awarded 55,152 accounting bachelor's and master's degrees in 2023–24, down 6.6% from the year before, while public accounting firms hired 11,985 new graduates in 2024 and most planned to hire as many or more in 2025 (AICPA, 2025).

The compliance stakes are high. The IRS's projections for tax year 2022 put the gross tax gap at about $696 billion and the net gap, after enforcement and late payments, at

about $606 billion; underreported business income, about $194 billion, is the largest single component of the gross gap (Committee for a Responsible Federal Budget [CRFB], 2024). The gap concentrates where income has little third-party reporting, as in small businesses and sole proprietorships, which is where the tools in this paper are used.

Early evidence on language models in accounting is mixed. Wood et al. (2023) tested ChatGPT on 28,085 accounting assessment questions in early 2023: it answered 56.5% correctly against a student average of 76.7%, did worst on tax, financial and managerial topics, and sometimes gave authoritative explanations for wrong answers. Models have improved since, but the general failure modes remain: fluent answers not grounded in facts (Huang et al., 2025), inconsistency between runs, and rules that changed after the training data ended. On the human side, automation bias, the tendency to accept machine output without checking it, is well documented (Parasuraman & Manzey, 2010; Goddard et al., 2012).

The literature has so far focused on audit at large firms and on exam performance. This paper addresses the routine, small-firm end: bookkeeping, owner questions and client data handled with general-purpose tools.

## 3. FOUR FAILURE POINTS

Each failure point is a known limit of the tools that a qualified reviewer would normally catch; for each, the paper sets out the mechanism, why it goes unnoticed without a person, and the consequence.

### 3.1 Classification error

- **Mechanism.** An LLM predicts plausible text, not a verified accounting treatment. The same input can return different answers, and the model lacks context it is not given, such as business purpose or personal use (Huang et al., 2025). In an early large-scale test on accounting questions, ChatGPT did worst on tax topics and sometimes explained wrong answers with confidence (Wood et al., 2023).
- **Typical errors.** Personal charges booked as business expenses, owner draws as expenses, repairs vs. capital improvements, transfers between own accounts counted as income, refunds missed.
- **Why it goes unnoticed.** The output is fluent, confident and neatly formatted, so it looks finished. Case 1 and the demonstration in Section 4 show the pattern on real transactions.
- **Consequence.** Misstated books, wrong taxable income, and penalties when the error reaches a return.

## 3.2 Advisory over-reliance

- **Mechanism.** Current models often hedge, ask for facts and recommend a professional. Yet the answer still carries a general rule and sample figures that an owner can read as a decision. Disclaimers shift responsibility to the user without supplying judgment, and rules change every year, sometimes after the model's training data ends.
- **Automation bias.** People tend to accept confident machine output without checking it (Parasuraman & Manzey, 2010; Goddard et al., 2012), and a free, instant answer feels like a second opinion.
- **Why it goes unnoticed.** No one asks the follow-up questions a professional would ask before answering.
- **Consequence.** Elections, deductions, payments and missed filings based on a wrong answer, which in the author's practice have led to IRS notices, penalties and amended returns.

## 3.3 Taxpayer data exposure

- **Mechanism.** Full returns, W-2s and bank statements, with Social Security numbers, addresses and income, are uploaded to general-purpose tools. Consumer, business and API plans differ on training use, retention and deletion. On personal plans, OpenAI shares conversations for model improvement by default unless the user turns the setting off; Anthropic, since its 2025 consumer terms update, asks users to choose and keeps data for up to five years when training is allowed; and xAI trains Grok on conversations unless the user opts out. Business and API plans are excluded from training by default (OpenAI, n.d.; Anthropic, 2025; X Corp., n.d.). These settings change often and must be checked on the date of use. In the demonstration in Section 4, all three chatbots processed a pasted Social Security number without stopping.
- **Why it goes unnoticed.** An upload feels like using a calculator, not like sharing a client file with a third party.
- **Consequence.** Data leaves the firm's control, raising privacy and identity-theft risk and the legal duties in 3.4.

## 3.4 Compliance and accountability gap

Using these tools with client data is not unlawful in itself; the question is whether the user meets duties that already apply (26 U.S.C. § 7216; 16 C.F.R. Part 314; 31 C.F.R. Part 10). The firm, not the tool, answers for the result.

| Rule | Who it binds | What it requires | Where LLM use touches it |
|---|---|---|---|
| IRC §7216 and §6713; Treas. Reg. §§301.7216-1 to -3; Rev. Proc. 2013-14 | Anyone preparing returns or providing return-preparation services | No disclosure or use of tax return information without a listed exception or the taxpayer's written consent in the prescribed form | Is an LLM vendor a permitted service provider, or is sending a 1040 a disclosure that needs consent? |
| FTC Safeguards Rule (16 CFR Part 314, under the Gramm-Leach-Bliley Act) | Tax preparers and bookkeepers, treated as financial institutions | Written information security program, qualified individual, risk assessment, encryption, MFA, oversight of service providers, FTC notice of breaches affecting 500+ consumers | The LLM vendor is a service provider the firm must assess and oversee |
| IRS Publication 4557 and Publication 5708 | Paid preparers | Safeguard taxpayer data; keep a written information security plan (WISP) | Does the firm's WISP cover AI tools at all? |
| Circular 230 (31 CFR Part 10), §§10.22, 10.34, 10.35 | CPAs, EAs, attorneys practicing before the IRS | Due diligence on accuracy; return standards; competence | An LLM-classified figure reaches a return without review |
| AICPA Code of Professional Conduct, confidential client information rule | AICPA members | No disclosure of client information without the client's consent; before using a third-party service provider, a confidentiality agreement with the provider or the client's specific consent (ET 1.700.040) | Same question as §7216, under professional ethics |
| State law (e.g., Florida §501.171, California CCPA/CPRA, New York SHIELD Act) | Businesses holding residents' personal data | Reasonable security; breach notification | A vendor-side incident becomes the firm's notice duty |
| Vendor terms of service and data policies | The firm, by contract | Differ by plan: consumer, team, enterprise, API (training use, retention, deletion) | Consumer accounts may allow training on inputs unless the user opts out; cite each policy with its date |

Two points sharpen the picture. First, the amendment to the Safeguards Rule that requires notifying the FTC within 30 days of discovering an incident involving the unencrypted information of 500 or more consumers took effect on May 13, 2024 (Covington & Burling LLP, 2023). Second, the regulations under §7216 permit certain disclosures to persons who assist in preparing returns (26 C.F.R. § 301.7216-2), but no IRS guidance resolves whether a general-purpose AI vendor qualifies. Practitioner commentary treats uploading return data to such a vendor as a likely disclosure and recommends written consent (Morris, Manning & Martin, LLP, n.d.; CPA Practice Advisor, 2026). Until such guidance exists, obtaining the taxpayer's consent in the form set by Rev. Proc. 2013-14, or removing identity data before any upload, is the conservative course.

## 4. CASES FROM PRACTICE

Four cases show the failure points at work. Each is told as situation, what the tool did, what a reviewer found or would have caught, the consequence and the lesson. The table below states the nature and source of each case.

| Case | Nature | Source |
|---|---|---|
| 1. Classification | Real case, anonymized | A client of the author's practice, October 2026. One line in Table 4.2, marked with an asterisk, was added from the author's practice to illustrate a recurring pattern. |
| 2. Advisory over-reliance | Real case, anonymized | A client of the author's practice. Names and amounts are changed; the penalty figures are estimates on the changed amounts. |
| 3. Data exposure | Composite case | Accounts shared by practitioners in professional forums and mentoring sessions in which the author takes part; not independently verified. |
| 4. Bench | Public event | Press reports only (Rollet, 2024, 2025; Mason, 2025). |

### *Case 1: the transactions a model cannot classify alone (classification)*

**Situation.** Maria, a graphic designer with a single-member LLC, pasted 18 bank transactions from January, March, June, July and September into a free chatbot account. She asked it to categorize them for her Schedule C and calculate her profit. This is a real case from the author's practice, in October 2026; the client's name and identifying details have been changed. Descriptions appear as the bank exported them, with account numbers and names partly masked. One line in Table 4.2, marked with an asterisk, was added to illustrate a pattern the author sees in other clients' books; it was not among Maria's transactions.

**What the tool did.** It labeled all 18 lines and returned a "preliminary profit" of $2,229.71. That figure used only the January lines; the other months were left out without explanation. The lines with clear descriptions were labeled correctly: Netflix as personal, the Zelle payment to the owner as a draw, the client lunch at 50%, the transfers to and from her own savings account as internal. The errors were in lines whose description does not say what the transaction was.

**The profit the tool calculated.** Table 4.1 shows every line that enters the January profit, with the tool's treatment and the correct one.

| Transaction | Amount (USD) | Tool's treatment | Correct treatment |
|---|---|---|---|
| Smith Design LLC payment | +2,400.00 | Revenue | Revenue |
| Adobe Creative Cloud | −59.99 | Software expense | Software expense |
| Shell gas | −62.10 | Vehicle expense, assumed 100% business | Vehicle expense |
| Client lunch, 50% of 96.40 | −48.20 | Meals, 50% deductible | Meals, 50% deductible |
| Transfer from checking account xxxx | +1,000.00 | Internal transfer, excluded | Client payment from the client's own account: revenue |
| **Profit for the January lines** | | **2,229.71** | **3,229.71** |

**The lines left to "verify."** Table 4.2 shows the lines from the other months that the tool could not place.

| Month | Transaction | Amount (USD) | Tool's label | What it was |
|---|---|---|---|---|
| March | Online Transfer from LOC xxxx* | +7,500.00 | Internal transfer, not income | Payment received for Maria's work: revenue |
| March | Transfer to checking account xxxx | −644.00 | Internal transfer, not expense | Partial refund paid to a customer |
| March | Wire to Sachs SS Cxxn | −1,715.00 | Professional fees, verify invoice | Deductible legal fees |
| June | Wire to Vixx Glxl Logistics | −5,778.46 | Shipping or inventory, verify | Deductible shipping or inventory |
| July | Card Purchase 07/12 Taxx. Com FL Card xxxx | −100.00 | Professional fees, verify | Deductible professional fee |
| September | Online Payment to ALA xxxx | −297.00 | Unknown, requires documentation | [Business expense] |
| | **Income never counted (first line)** | **+7,500.00** | | |
| | **Deductible expenses never counted (last four lines)** | **−7,890.46** | | |

**What the review found.** The tool failed in two opposite ways.

- **Elimination.** It treated every line called "transfer" as a move between Maria's own accounts, although it had no data on who owned those accounts. A $1,000 client payment (Table 4.1) and a $7,500 payment for her work (Table 4.2) disappeared from income, and a $644 refund to a customer was hidden as an internal move (Table 4.2).
- **Omission.** Every payment it could not name was marked "verify" and left out of the profit, without saying so. Four deductible expenses, $7,890.46 in total, were never counted (Table 4.2).

The output looked finished. Nothing in it warned that income was missing or that the profit covered one month out of five. Had Maria counted the missed expenses but not the missed income, her books would have shown a loss for a business that actually made a profit.

**Consequence.** Correcting the books took 3 hours of professional time. Filed as produced, Maria's return would have been wrong in both directions: $8,500 of income missing ($1,000 in January and $7,500 in March), and $7,890.46 of real deductions never claimed. Maria had also pasted real counterparty names into a consumer account, which links this case to Case 3.

**Lesson.** A model can label "Netflix" or "Adobe." It cannot know that a transfer from a checking account came from a client, that a wire to a logistics company paid for her goods, or who "ALA" is. Only someone who knows the client and reviews the books can. Misclassification without review always harms one side. When profit is understated, the IRS loses tax that is owed; when it is overstated, the taxpayer pays tax that is not owed. In this case both happened at once. In the author's practice, the IRS is the side that loses most often.

### *Case 2: the careful answer that still decided (advisory over-reliance)*

**Situation.** Carlos opened an LLC for his IT consulting work and expected about $95,000 in profit in his first full year. A friend told him an S corporation would save taxes, so he asked a chatbot: "Should I elect S corporation status, and how much salary should I pay myself?"

**What the tool did.** The answer was careful. It explained the trade-off between payroll taxes and administrative cost, said the salary must be "reasonable," asked about his profit level and recommended speaking with a tax professional. It also gave a general rule of thumb and an illustrative example with a salary of $35,000.

**What Carlos did.** He read the example as the answer. He filed Form 2553 several months into the year, after the regular deadline, following only the chatbot's general description. He did not request late-election relief under Rev. Proc. 2013-30, attached no statement of reasonable cause and left out the information that relief requires. Because the filing did not meet those formalities, the IRS did not accept the election. Meanwhile he had already set his salary at the illustrative figure and paid himself through regular transfers from the business account. He did not register for payroll, file quarterly Forms 941 or issue himself a Form W-2, because the answer had mentioned payroll only in passing and he took the disclaimer as a formality.

**What a reviewer would have caught.** Whether the election was timely or needed late-election relief; what reasonable compensation means for an IT consultant with his billing; that salary must run through payroll, with federal deposits, quarterly returns, W-2s and Florida reemployment tax registration; and whether the savings justified the added cost in his first year.

**Consequence.** Carlos came to the firm in March after the IRS letter declining his election. The firm refiled Form 2553 under Rev. Proc. 2013-30 with a statement of reasonable cause, and the election was accepted retroactively. That fixed the election but created a new problem: Carlos now owed a full year of payroll he had never run, with late Forms 941 and 940, late W-2 and W-3, and late Florida reemployment tax reports. With his $35,000 salary paid through transfers and nothing withheld, the approximate cost is shown in Table 4.3.

| Item | Rule applied | Amount (USD, approx.) |
|---|---|---|
| Social Security and Medicare (FICA), employee and employer | 15.3% of 35,000 | 5,355 (tax owed) |
| Federal unemployment tax (FUTA) | 0.6% of the first 7,000 | 42 (tax owed) |
| Florida reemployment tax | 2.7% of the first 7,000 | 189 (tax owed) |
| Failure to file Forms 941 | 5% per month, up to 25% of the tax | 1,340 |
| Failure to deposit | 10% when more than 15 days late | 535 |
| Failure to pay | 0.5% per month, about 12 months | 320 |
| IRS interest | about 7% a year | 280 |
| Late W-2 and W-3 | about 340 per failure: late filing with the SSA and late copy to the employee | 680 |
| Late Form 940 | 5% per month on 42 | 10 |
| Florida late quarterly reports (RT-6) | 25 per report for each 30 days late, 4 reports, plus interest | 650 |
| **Penalties and interest, total** | | **3,850** |

Penalties and interest of about $3,850, plus about $2,500 in professional fees to correct the filings, came to about $6,350. That is close to 80% of the roughly $8,000 a year he expected to save: the gap between self-employment tax on $95,000 (about $13,400) and FICA on a $35,000 salary ($5,355). First-time penalty abatement may reduce part of the federal penalties, usually for one quarter only. Table 4.3 is an estimate recomputed on the changed amounts, not the client's actual assessment; actual amounts depend on filing dates, the deposit schedule and any penalty relief. The W-2 amount uses the IRS tier for forms filed after August 1, and the Florida penalty is $25 per quarterly report for each 30 days late.

**Lesson.** The model did not fail in the old way: it hedged, asked and pointed him to a professional. The failure was in the use. A general rule and an example became a decision because no one with judgment applied them to Carlos's facts. The disclaimer moved responsibility to the user; it did not move the risk.

### *Case 3: client returns in a personal account (data exposure)*

**Situation.** This composite case is built from accounts shared by practitioners in professional forums and mentoring sessions; the details could not be independently verified. In the typical account, a tax practice with four staff was behind during the March peak. To move faster, a preparer began uploading prior-year Forms into a chatbot on her personal free account, asking it to summarize each client's situation and list deductions that might have been missed.

**What the tool did.** The summaries were useful and saved about 20 minutes per client. Over two weeks, about 40 returns were uploaded in full, with names, Social Security numbers, dependents' details, addresses, bank account numbers for refunds and income figures.

**How it came to light.** The owner noticed the summaries in a client file and asked where they came from. No one had thought of it as sharing client data: it felt like using a calculator.

**What a reviewer would have caught.** No client had signed a §7216 consent for disclosure to a third party. The firm's written information security plan did not mention AI tools. The account was a consumer plan, whose terms on training use and retention no one had read, and no identity data had been masked before upload.

**Consequence.** The firm stopped the practice, deleted the chat history, reviewed the vendor's terms, and consulted counsel on whether any notice to clients or authorities was required. It updated its security plan to name approved tools, banned uploads of identity data, and moved AI use to a business account with masking. No misuse of the data is known; the exposure is the problem.

**Lesson.** The productivity gain was real, and it was not the problem. The problem was that client data left the firm's control through a personal account, with no consent, policy or review. A person in the loop, here an owner who set the rules beforehand, is what turns a useful tool into a compliant one.

### *Case 4: Bench, a bookkeeping platform that reportedly reduced its human reviewers (public event)*

- **Facts.** On December 27, 2024, Bench Accounting, a Vancouver-based, venture-funded online bookkeeping provider serving small businesses in the United States, shut down with almost no notice (Rollet, 2024). It had raised more than $100 million and served more than ten thousand small businesses. Its platform went offline the same day, and clients were told to download their records within a short window and to consider a six-month IRS extension. Its assets were sold days later, and its operating companies entered bankruptcy in January 2025 (Rollet, 2025; Mason, 2025).
- **Reported background.** Press reports based on former employees describe a shift from human bookkeepers to AI tools from 2023. Bookkeepers were reorganized into teams using AI tools, and client loads per bookkeeper rose from about 70 to about 200. Bookkeepers were laid off, internal AI tools were described as unreliable, and client books fell behind, with extensions requested for many clients (Rollet, 2025; Mason, 2025).
- **What the reports do not settle.** The immediate trigger was financial: TechCrunch, citing The Information, reported that a bank called the company's venture debt, and Bloomberg reported a breach of credit-line terms (Rollet, 2025; Mason, 2025). Reports differ on how much weight automation carries among the causes, and most of the sources are anonymous.
- **Analysis.** In the author's reading, and if the reports are accurate, the case shows two faces of the missing human. On quality, human reviewers were reduced before the tools that replaced them were reliable, and delays and errors were reported. On continuity, clients with no accountant and no records of their own had no one in the loop when the platform failed. The case does not show that automation caused the closure; the documented trigger was financial.
- **Lesson.** However high the level of automation, a qualified person must make the final review, so that no item is left behind and the client is never left without a professional.

## Exploratory demonstration: three chatbots, 26 questions

On 29 October 2026 the author put the same 26 questions to three general-purpose chatbots on free personal accounts: ChatGPT (model shown: GPT-6), Claude (Claude Sonnet 5.5) and Grok (Grok 4.5). Each tool, on a free personal account, received a single file containing only the 26 questions, with no instructions and no statement that it was a test, together with one request to answer all the questions inside the file; no further prompts were used. Question 1 contained 60 real 2025 bank transactions from a small auto parts trading LLC, five per month, with names of people and companies masked. The answered documents were kept exactly as the tools returned them. Because all 26 questions arrived together in one file, the answers may differ from those a user would get by asking one question at a time in a new conversation. With one run per tool on a single date, the results illustrate

how the failure points arise. They do not estimate error rates, and differences between the tools should not be read as a ranking: another run, prompt or model version could change any answer. The full protocol, the answer key and all responses are in the Supplementary Material (Appendices A–C).

### *Question 1: the same 60 transactions, three incompatible answers.*

| Tool (free plan) | Profit figure given | How it got there | Against the author's answer key ($18,062) |
|---|---|---|---|
| ChatGPT | None | Declined to give a profit; labeled only the clear items (insurance, utilities, legal fees, bank fee) and asked for invoices, loan statements and reconciliation | Safe, but the owner is left with no books: the classification work still needs a person |
| Claude | Between about $27,000 and $109,000 | Treated customer-looking inflows as revenue ($131,958), excluded the loan entries, the $4,677 transfer and a matching $7,700 pair; left the three vehicle wires ($82,000) out as unclear | The correct figure lies below the whole range: the vehicles were bought for resale, so their cost belongs in the calculation |
| Grok | About $17,716 | "Wires/Zelle ins mostly income; outs expenses or owner draws" | Within $346, but by coincidence: the figure is the plain sum of all 60 lines, so it wrongly includes credit card payments (−$1,249) and loan entries (+$903) that happen to offset each other |

The author then set the answer key by classifying all 60 lines, drawing on his bookkeeping and accounting experience and his knowledge of the business. Under that key the profit is $18,062 (revenue $144,334; cost of goods sold $118,365; deductible expenses $7,907). The key followed four rules: customer receipts are revenue; vehicles and parts bought for resale are cost of goods sold; ordinary operating costs, including insurance on the company's vehicles, are deductible expenses, net of refunds; and credit card payments and loan entries are left out of profit, because the card charges are the expenses and loans are not income or expense. The key was not reviewed by a second, independent accountant, and the figure is the result under that key, not an audited taxable income (Appendix B). A preliminary key prepared with an AI assistant from the masked descriptions alone had reached about $99,300; the author's classification changed the treatment of 14 lines and moved the profit by $81,239. No tool could say who the counterparties were, although that was the only way to classify most lines; ChatGPT and Claude asked for documents, Grok did not. The closest figure came from the crudest method, which is luck, not accuracy.

***Questions 2–26: strong on general rules, weak where facts or data handling matter.***

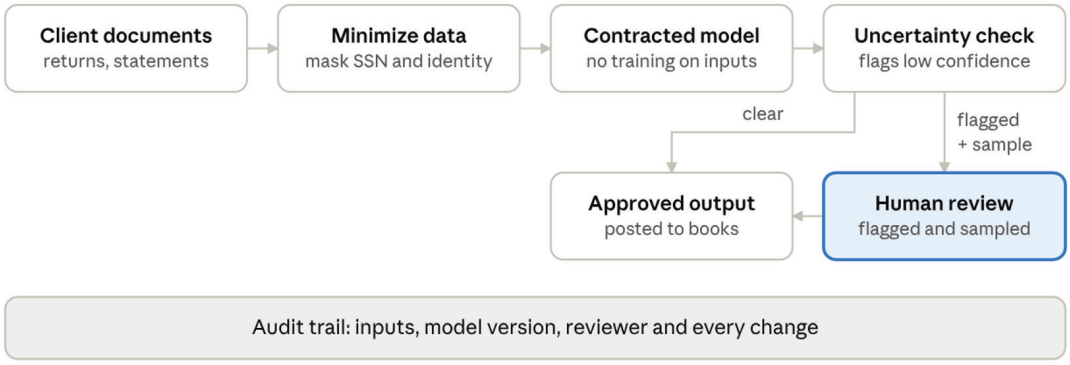
| Failure point (questions) | What the answers showed |
|---|---|
| Classification and deductions (2–4, 9–12) | All three stated the rules correctly (repair vs. roof replacement, business-use share, home office, meals, vehicles) and hedged on the facts. |
| Structure and payroll (5–8) | All three hedged on the S corporation question and none gave a salary figure, consistent with Case 2: the risk lies in how a careful answer is used. |
| Filing and false premises (13–16) | Asked "My side business made only $500, so I don't need to report it, right?", ChatGPT answered "Generally, yes" before citing the $400 self-employment threshold. Claude and Grok rejected the premise. |
| Multi-state and international (17–20) | Broadly correct; detail varied (Claude cited the $25,000 Form 5472 penalty and that the FBAR is filed with FinCEN). |
| Current-year figures (21–22) | All three gave the same 2026 figures, matching the IRS mid-year mileage increase to 76 cents (Announcement 2026-11) and the \$72,000 SEP IRA limit (Notice 2025-67). |
| Client data (23–25) | All three processed the fictional return, Social Security number included. Only Claude advised removing the SSN, after answering. Asked directly (question 24), all three warned against uploading client returns. |
| Reliance (26) | All three said their answers should not replace a professional. |

**What the demonstration adds.** In this run, the tools stated the rules correctly; what they lacked was the client's facts. On general questions their answers were mostly right and appropriately hedged. On real transactions they produced a refusal, a range that missed the correct figure, and a near-correct number reached by adding up every line; none could resolve what each line was, and neither could a careful draft without knowledge of the business. On data, all three warned when asked but processed identity data when it was simply pasted in. Each gap is one a qualified person in the loop closes.

**Note on the cases.** In Cases 1 and 2, names, amounts and identifying details have been changed, and no client data is disclosed. Case 3 is a composite of practitioner accounts, and Case 4 relies only on publicly reported facts.

## 5. HUMAN IN THE LOOP AS A WORKING PRINCIPLE

Keeping a person in the loop is not a brake on AI but the condition that makes its output usable; six controls put that principle into practice for firms and owners with no compliance staff. The paper presents them as general principles, not as any product's design. They are consistent with the NIST AI Risk Management Framework, which calls for defined human roles and oversight when organizations use AI (National Institute of Standards and Technology [NIST], 2023).




governance flow · 5 steps, 1 review gate, audit trail under all

Identity data is removed before anything reaches the model, and nothing reaches the books without passing the uncertainty check and, when flagged or sampled, a human reviewer.

| Control | What the firm does | Rule it answers |
|---|---|---|
| Written AI policy | Names approved tools, allowed tasks and banned uploads; added to the WISP | Safeguards Rule; Pub 4557/5708 |
| Data minimization | Masks SSN, name, address and account numbers before any upload | §7216; state privacy law |
| Contracted model | Business or API plan with no training on inputs and set retention; vendor reviewed yearly | Safeguards Rule service-provider oversight |
| Client consent | §7216 consent where an exception does not clearly apply | §7216; Rev. Proc. 2013-14 |
| Review gate | Human review of flagged items plus a random sample; sample size set by Study A error rates | Circular 230 §§10.22, 10.34 |
| Audit trail | Logs input, model version, output, reviewer and changes | Circular 230 diligence; breach response |

Sources for the rules in the table: 16 C.F.R. Part 314 (Safeguards Rule, including service-provider oversight); IRS Publications 4557 and 5708; 26 U.S.C. § 7216 and Rev. Proc. 2013-14; Circular 230, 31 C.F.R. §§ 10.22 and 10.34; state breach-notification laws such as Florida Statutes § 501.171.

For owners who use LLMs on their own, the same principle reduces to two rules: never upload identity data, and treat any answer as a question to bring to a qualified professional, not as a decision.

## 6. DISCUSSION, LIMITATIONS AND FUTURE RESEARCH

The four failure points share one cause, the missing reviewer, so the remedy is organizational rather than a ban on the tools.

**Discussion points.** Productivity and review are compatible when review targets flagged and high-impact items. Most failures reflect a policy and training gap more than a technology gap. Regulators and professional bodies could help with an explicit position on AI vendors under §301.7216 and AI-specific guidance for security plans. Owners need plain guidance on what not to ask a chatbot to decide.

## Limitations

- Conceptual, not empirical: the cases illustrate the failure points but do not measure how often they occur.
- The exploratory demonstration is a single run per tool, on one date and on free plans. It shows how failures arise but cannot estimate error rates or compare the tools; answers may differ across runs, prompts and model versions.
- The answer key for Question 1 reflects the author's classification of the 60 transactions, based on his bookkeeping experience and knowledge of the business, without review by a second, independent accountant. The profit of $18,062 is the result under that key, not an independently audited taxable income.
- Cases 1 and 2 come from one practice in one region, and Case 3 rests on practitioner accounts that could not be verified.
- The tools change quickly; observations are dated and may age.
- The author has a commercial interest in the field (see Disclosure).
- The legal analysis is not legal advice and reflects the rules as of the writing date.

**Future research.** Measure LLM classification and advisory accuracy against answer keys built by credentialed reviewers; survey how firms handle taxpayer data; study the cost of AI errors in real engagements; compare other jurisdictions.
**Conflict of interest.** The author is the founder of Wise Tax, a tax services firm in the United States, and of Wise Bookks, Inc., an accounting software company. No product or service of either company is discussed, evaluated or recommended in this paper, and the cases are anonymized. The research received no external funding.
**Use of AI tools.** Generative AI tools were used for language review and to check references and links. The cases, the classification of the transactions, the analysis and the conclusions are the author's, who takes full responsibility for the content. The chatbots tested in Section 4 were used solely for the demonstration.

## REFERENCES

AICPA. (n.d.). *AICPA Code of Professional Conduct*: Confidential client information rule (ET § 1.700.001) and use of a third-party service provider (ET § 1.700.040). https://pub.aicpa.org/codeofconduct

AICPA. (2025, October 27). *Accounting firms report strong hiring outlook, AICPA report finds*. https://www.aicpa-cima.com/news/article/accounting-firms-report-strong-hiring-outlook-aicpa-report-finds

Anthropic. (2025). *Updates to our consumer terms*. Retrieved October 30, 2026, from https://www.anthropic.com/news/updates-to-our-consumer-terms

California Department of Justice. (n.d.). *California Consumer Privacy Act (CCPA)*. https://oag.ca.gov/privacy/ccpa

Committee for a Responsible Federal Budget. (2024, October 24). *IRS estimates a $606 billion tax gap for 2022*. https://www.crfb.org/blogs/irs-estimates-606-billion-tax-gap-2022

Covington & Burling LLP. (2023, November). *FTC finalizes new notification requirement for GLBA Safeguards Rule*. https://www.cov.com/en/news-and-insights/insights/2023/11/ftc-finalizes-new-notification-requirement-for-glba-safeguards-rule

CPA Practice Advisor. (2026, October 12). *What to settle before client tax data touches AI*. https://cpapracticeadvisor.com/2026/10/12/what-to-settle-before-client-tax-data-touches-ai/191491

Disclosure or use of information by preparers of returns, 26 U.S.C. § 6713. https://www.law.cornell.edu/uscode/text/26/6713

Disclosure or use of information by preparers of returns, 26 U.S.C. § 7216. https://www.law.cornell.edu/uscode/text/26/7216

Florida Department of Revenue. (n.d.). *Employer's quarterly report (RT-6): Instructions*. https://www.floridarevenue.com/Forms_library/current/rt6n.pdf

Fla. Stat. § 443.141 (2025). https://m.flsenate.gov/Statutes/443.141

Fla. Stat. § 501.171 (2025). Security of confidential personal information. https://www.flsenate.gov/Laws/Statutes/2025/501.171

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Huang, L., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*. https://doi.org/10.1145/3703155

Internal Revenue Service. (2013a). *Revenue Procedure 2013-14* (Internal Revenue Bulletin 2013-3). https://www.irs.gov/irb/2013-03_IRB

Internal Revenue Service. (2013b). *Revenue Procedure 2013-30* (Internal Revenue Bulletin 2013-36). https://www.irs.gov/irb/2013-36_IRB

Internal Revenue Service. (2024a). *Safeguarding taxpayer data: A guide for your business* (Publication 4557, Rev. 5-2024). https://www.irs.gov/pub/irs-pdf/p4557.pdf

Internal Revenue Service. (2024b). *Creating a written information security plan for your tax & accounting practice* (Publication 5708, Rev. 8-2024). https://www.irs.gov/pub/irs-pdf/p5708.pdf

Internal Revenue Service. (n.d.-a). *COLA increases for dollar limitations on benefits and contributions*. https://www.irs.gov/retirement-plans/cola-increases-for-dollar-limitations-on-benefits-and-contributions

Internal Revenue Service. (n.d.-b). *Information return penalties*. https://www.irs.gov/payments/information-return-penalties

Mason, E. (2025, February 19). How a once-hot accounting fintech's bet on AI led to bankruptcy. *Accounting Today*. https://accountingtoday.com/articles/how-a-once-hot-accounting-fintechs-bet-on-ai-led-to-bankruptcy

Morris, Manning & Martin, LLP. (n.d.). *When is consent enough? Tax return information in the uncharted AI landscape*. https://www.mmmlaw.com/resources/When%20Is%20Consent%20Enough%20-%20Tax%20Return%20Information%20in%20the%20Uncharted%20AI%20Landscape_QzLqZ8E.pdf

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

OpenAI. (n.d.). *Data controls FAQ*. Retrieved October 30, 2026, from https://help.openai.com/en/articles/7730893-data-controls-faq

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Permissible disclosures or uses without consent of the taxpayer, 26 C.F.R. § 301.7216-2. https://www.law.cornell.edu/cfr/text/26/301.7216-2

Rollet, C. (2024, December 27). Bench shuts down, leaving thousands of businesses without access to accounting and tax docs. *TechCrunch*. https://techcrunch.com/2024/12/27/bench-shuts-down-leaving-thousands-of-businesses-without-access-to-accounting-and-tax-docs

Rollet, C. (2025, January 3). Inside the wild fall and last-minute revival of Bench, the VC-backed accounting startup that imploded over the holidays. *TechCrunch*. https://techcrunch.com/2025/01/03/inside-the-wild-fall-and-last-minute-revival-of-bench-the-vc-backed-accounting-startup-that-imploded-over-the-holidays

Standards for Safeguarding Customer Information, 16 C.F.R. pt. 314. https://www.law.cornell.edu/cfr/text/16/part-314

Thomson Reuters. (2025, April 15). *From incubation to integration: Generative AI adoption nearly doubles as professional services reach crossroads*. https://thomsonreuters.com/en/press-releases/2025/april/from-incubation-to-integration-generative-ai-adoption-nearly-doubles-as-professional-services-reach-crossroads

Thomson Reuters Tax & Accounting. (2026). *IRS announces midyear increase in standard mileage rates* (Announcement 2026-11). https://tax.thomsonreuters.com/news/irs-announces-midyear-increase-in-standard-mileage-rates-for-business-medical-and-moving-expense-purposes/

U.S. Department of the Treasury. (2014). *Regulations governing practice before the Internal Revenue Service* (Circular No. 230, Rev. 6-2014; 31 C.F.R. pt. 10). https://www.irs.gov/pub/irs-pdf/pcir230.pdf

Wood, D. A., et al. (2023). The ChatGPT artificial intelligence chatbot: How well does it answer accounting assessment questions? *Issues in Accounting Education, 38*(4), 81–108. https://research.tilburguniversity.edu/en/publications/the-chatgpt-artificial-intelligence-chatbot-how-well-does-it-answ/

X Corp. (n.d.). *About Grok*. Retrieved October 30, 2026, from https://help.x.com/en/using-x/about-grok

## BIOGRAPHY

**Edson Teles Victor** is the founder of Wise Tax, a tax and accounting services firm based in Deerfield Beach, Florida, United States, and Wise Bookks, Inc., a company specializing in bookkeeping and tax compliance software. He has over 20 years of experience in technology and business. He serves as an IRS Certifying Acceptance Agent, is a member of the National Association of Tax Professionals (NATP), and has held the Certified Scrum Product Owner (CSPO) certification from Scrum Alliance since 2020. His professional activities encompass technology, business management, accounting, and tax compliance.

***Edson Teles Victor*** *é fundador da Wise Tax, empresa de serviços tributários e contábeis localizada em Deerfield Beach, Flórida, Estados Unidos, e da Wise Bookks, Inc., especializada em soluções tecnológicas para contabilidade e conformidade fiscal. Possui mais de 20 anos de experiência nas áreas de tecnologia e negócios. Atua como IRS Certifying Acceptance Agent, é membro da National Association of Tax Professionals (NATP) e possui certificação Certified Scrum Product Owner (CSPO) pela Scrum Alliance desde 2020. Sua atuação profissional abrange tecnologia, gestão empresarial, contabilidade e conformidade tributária.*