DANNY WERFEL, FORMER IRS COMMISSIONER

Appendix A
AI Risk Register for Tax Authorities

# RISK	DESCRIPTION	SCENARIO	ILLUSTRATIVE CONTROLS
1 Hallucination	The AI generates confident, specific, and factually incorrect output, whether delivered to a taxpayer seeking guidance or embedded in an authority's internal workflow.	*A public-facing chatbot confidently tells a taxpayer she qualifies for the earned income credit (EITC). She doesn't. She files, claims it, and receives a penalty notice six months later.*	Ground all outputs to authoritative IRC/Treasury sources; require source citations on every AI response; human review sampling before and after deployment; escalation triggers when system encounters uncertainty.
2 Drift	The AI's outputs gradually shift away from accurate tax interpretation over time, without any single failure triggering a review.	*A chatbot deployed for EITC guidance begins producing subtly different eligibility interpretations over 18 months as the underlying model is updated, without any single output being obviously wrong.*	Version control on model updates; periodic accuracy benchmarking against authoritative sources; output consistency monitoring over time; formal revalidation before and after model updates.

# RISK	DESCRIPTION	SCENARIO	ILLUSTRATIVE CONTROLS
3 Explainability	The AI cannot trace or explain why it produced a given output, making it impossible to verify, audit, or defend. Explainability also encompasses the inability to document what data was used to train the model, what AI approach was employed, and how the application was tested — all of which are prerequisites for meaningful governance. Distinct from Lack of Auditability (see No. 18), which concerns the historical record of system behavior, vs. a single output at the moment it is produced.	*A taxpayer appeals a determination and asks the authority to explain the AI reasoning. The authority cannot reconstruct the output or explain what inputs drove it.*	Require source citations on all outputs; build explanation features into system design; maintain audit logs linking outputs to inputs; legal review of explainability requirements before deployment; documentation of training data sources, AI methodology, and testing procedures as a baseline explainability requirement; validation that training data is fit for purpose for the specific tax administration context.
4 Over-Trust / Unauthorized Reliance	Staff, taxpayers, or preparers treat AI output as authoritative, removing the human check that catches errors — and in the most consequential cases, a taxpayer relies on AI-generated guidance as if it were an official legal determination, creating potential unintended estoppel claims.	A taxpayer asks the authority's chatbot whether they qualify for a home office deduction. The chatbot answers “Yes, you qualify.” The taxpayer claims the deduction, which is later disallowed on audit. The taxpayer argues they relied in good faith on official guidance.	Clear disclaimers that AI outputs do not constitute authoritative guidance; source citations on all outputs; escalation to human support for scenario-specific questions; staff training on appropriate AI use; legal review of chatbot scope before deployment.

# RISK	DESCRIPTION	SCENARIO	ILLUSTRATIVE CONTROLS
5 Policy Drift	The AI subtly reshapes how tax rules are interpreted through accumulated outputs, effectively rewriting policy without any formal change process.	*An AI-powered taxpayer assistance tool begins providing slightly different interpretations of eligibility criteria for education credits compared to official published guidance, gradually shifting taxpayer understanding of the rules.*	Authoritative grounding; policy alignment validation; change management controls on model updates; output constraints limiting interpretation; human oversight on guidance outputs.

# RISK	DESCRIPTION	SCENARIO	ILLUSTRATIVE CONTROLS
6 Bias / Neutrality	The AI produces systematically different outcomes for different taxpayer groups based on income, language, geography, or other characteristics unrelated to legal merit.	*A multilingual chatbot provides less accurate guidance to non-English speakers, creating unequal access to accurate information. Bias risk extends to enforcement outcomes and to variation in the quality, tone, or clarity of AI-generated communications that disadvantage certain taxpayers.*	Bias testing across demographic and language groups before deployment; neutrality review of outputs; multilingual testing and validation; ongoing disparity monitoring; fairness audits with documented findings.
7 Audit Targeting	The AI introduces or amplifies bias in how taxpayers are selected for audit, producing disparate impact across industries, demographics, or geographic groups.	*An AI audit selection model trained on historical IRS data disproportionately flags returns from self-employed filers in certain ZIP codes. Independent researchers later demonstrate the pattern correlates with race.*	Pre-deployment fairness testing across demographic groups; periodic bias audits with documented findings and remediation; human review required for high-impact audit selection decisions; disparity monitoring as an ongoing performance indicator; recognition that tax data systems may not collect demographic variables needed for bias testing, requiring acquisition of third-party data or deliberate additions to future data collections to support this testing.

# RISK	DESCRIPTION	SCENARIO	ILLUSTRATIVE CONTROLS
8 Trust Shock / Compliance Norm Erosion	A visible AI failure triggers rapid loss of public confidence in the tax authority — and repeated smaller failures cumulatively erode the public's belief that the tax system is fair, gradually degrading voluntary compliance.	*Over 18 months, minor recurring AI errors — inconsistent guidance, occasional incorrect penalty calculations, audit-targeting patterns correlated with industry type — produce no single news cycle but cause trust surveys to decline and compliance rates to edge downward.*	Proactive public communication about AI governance; transparency reporting on AI use and error rates; taxpayer trust and confidence monitoring distinct from complaint volume; fairness audits; response plan for visible failures.
9 Surveillance Perception	Expanded AI use creates public perception that the authority is conducting mass surveillance, eroding trust even where the underlying use is legitimate.	*A state tax authority deploys AI to cross-reference third-party data sources for compliance purposes. Even though the use is legally authorized, public perception of mass surveillance triggers advocacy campaigns and declining cooperation with voluntary disclosure programs.*	Proactive transparency about AI use and its limits; public communications strategy; privacy impact assessments before deployment; clear statutory authority documentation for each AI use case.

# RISK	DESCRIPTION	SCENARIO	
10 Prompt Injection	A malicious actor manipulates inputs to the AI system to bypass its safeguards, extract sensitive information, or produce harmful outputs.	*A taxpayer submits a query with embedded instructions that cause the chatbot to bypass its topic restrictions, revealing info about the authority's detection models — and as a result the deployment expands the external attack surface available to bad actors.*	Prompt filtering and sanitization; red-teaming and adversarial testing before deployment; monitoring for injection patterns; strict output constraints; regular security testing.
11 Data Leakage	Taxpayer data entered into or processed by AI systems is exposed, accessed without authorization, or retained and reused by vendors in ways that violate privacy obligations — and the aggregation of that data creates concentrated breach targets.	*A chatbot vendor retains anonymized conversation transcripts in its training pipeline. A researcher later demonstrates that combining those transcripts with public records re-identifies specific taxpayers — a privacy violation the authority never anticipated or prohibited in the contract.*	Contract prohibition on vendor data retention and reuse; data minimization; annual vendor security review with audit rights; re-identification risk assessment before any data sharing.
12 Re-identification	Anonymized data processed by AI systems are combined with other information to identify specific individuals, creating unanticipated privacy exposure — and AI accelerates the speed and scale at which re-identified data is weaponized for fraud.	*An authority shares anonymized taxpayer interaction data with a research partner. The researcher combines it with public records and successfully re-identifies a subset of taxpayers, exposing sensitive filing information.*	Re-identification risk assessment before any data sharing; data minimization; privacy impact assessment for all AI deployments involving taxpayer data; contractual prohibitions on data combination.

# RISK	DESCRIPTION	SCENARIO	ILLUSTRATIVE CONTROLS
13 Adversarial Manipulation	Sophisticated actors deliberately craft tax filings or inputs to exploit known weaknesses in AI detection models, evading compliance enforcement.	*A fraud ring reverse-engineers the authority's AI audit selection model and structures filings to fall below detection thresholds, evading detection while filing fraudulent returns at scale.*	Adversarial testing of detection models; model opacity to prevent reverse engineering; periodic model refresh; human review of anomaly patterns; red-team testing.
14 Agentic AI Risk	An AI system with autonomous action capabilities takes irreversible actions before a human can intervene (or that impose significant burden on taxpayers even when reversed), is manipulated into accessing systems or data beyond its intended scope, or causes cascading failures across connected legacy systems at a speed that outpaces human oversight.	*An agentic AI managing correspondence is manipulated through a crafted input into issuing penalty notices to thousands of taxpayers based on incorrect eligibility determinations. Notices are dispatched before any human reviewer flags the anomaly.*	Mandatory human approval for irreversible, financial, or rights-affecting agent actions; least-privilege design; kill-switch capability; sandboxed execution; real-time anomaly detection; governance review for any capability expansion.

# RISK	DESCRIPTION	SCENARIO	ILLUSTRATIVE CONTROLS
15 Workforce and Expertise Erosion	AI adoption reduces hiring and development of human tax expertise while senior institutional knowledge departs, leaving the authority less capable of governing the systems it has deployed.	*After three years of AI-assisted audit selection, an agency has not hired experienced auditors at historical rates. When the model fails a mandatory equity review, no one on staff has the expertise to rebuild the methodology or validate a replacement.*	Minimum human expertise thresholds in critical roles; AI supervision career tracks; workforce capability assessments before and after deployment; knowledge retention programs; hiring floors in critical roles.
16 Scope Creep	AI systems are quietly expanded beyond their originally approved use cases without corresponding updates to controls, governance, or validation.	*A chatbot deployed for filing deadlines and refund-status questions is expanded to answer credit eligibility questions before the expanded use case is fully tested and original controls are updated.*	Formal use-case approval process required before any expansion; phased rollout with re-validation before scope expansion; change-management review; updated guardrails for each expanded use case.
17 Control Dilution	Existing review and oversight processes fail to scale with AI-driven increases in volume and speed, resulting in weaker governance even if all controls remain nominally unchanged.	*An AI-enabled correspondence drafting tool allows staff to produce far more taxpayer communications, but legacy review processes do not adapt — resulting in weaker quality checks as volume increases.*	Control redesign for AI-enabled workflows; sampling and secondary review scaled to volume; volume-based escalation thresholds; audit trails; training for staff and supervisors; periodic testing of whether human oversight is genuinely active.

# RISK	DESCRIPTION	SCENARIO	ILLUSTRATIVE CONTROLS
18 Lack of Auditability	The authority cannot reconstruct what its AI systems did, making appeal response and error identification implausible. This risk is acute with commercial off-the-shelf solutions, which do not automatically generate artifacts useful for reconstruction — auditability must be intentionally designed into the system from the outset.	*A taxpayer appeals an AI-assisted determination. The authority cannot reconstruct the output, identify what inputs drove it, and thus show the determination was consistent with comparable cases.*	Auditability-by-design as a pre-deployment requirement — specifying what artifacts must be generated and retained before procurement or build decisions are made; tamper-proof logging of all AI outputs, inputs, and human review decisions; audit trail access independent of vendor cooperation; log retention commensurate with appeals window.
19 Regulatory & Policy Noncompliance	The authority deploys or operates AI in ways that are inconsistent with applicable laws, regulations, or policies.	*A federal tax authority deploys an AI system under prior-administration requirements. A policy change imposes new transparency and human oversight requirements. The authority lacks a monitoring function and does not identify the gap.*	Designated function monitoring AI-related legal and regulatory requirements; periodic compliance reviews of all active deployments; AI inventory with current compliance status; legal review triggered by any change in executive orders or OMB guidance.

# RISK	DESCRIPTION	SCENARIO	ILLUSTRATIVE CONTROLS
20 Enterprise Governance Gap / Gradual Autonomy Creep	No single function within the authority is responsible for managing AI investments across the organization — and AI tools gain autonomous capabilities incrementally through vendor updates without any single change triggering a formal governance review.	*A tax authority operates 14 separate AI systems each with local governance and no enterprise approach. One system — a correspondence tool — has acquired memory and routing capabilities through vendor updates, becoming functionally agentic without any formal governance decision.*	Designated Chief AI Officer with enterprise-level governance authority; centralized AI use case inventory; vendor contracts requiring notification before autonomous capability additions; annual reclassification of all AI tools; governance review triggered by any capability expansion.

Appendix B
Applying the AI Risk Register for Tax Authorities
Case Study #1

Case Study 1 — Tax Authority

The Scenario

A vendor proposes an AI-powered chatbot for the Department's public website, capable of answering taxpayer queries in natural language. The Department's goal: deploy responsibly for fact-based questions while ensuring the chatbot never ventures into tax advice, eligibility determinations, or definitive conclusions about what a taxpayer qualifies for.

The Eight-Step Methodology at a Glance

Step	Action	What It Produces
1	**Frame the Deployment**	Scope, function, user population, process stage, and deployment boundaries defined.
2	**Screen for Relevance**	A filtered list: Which of the 20 risks in the register apply to this deployment.
3	**Score Each Relevant Risk**	High / Medium / Low for each risk, using the Likelihood × Impact matrix and tax AI push factors.
4	**Determine the Tier**	Tier 1 (Deployment Blocker), Tier 2 (Active Governance), or Tier 3 (Monitor and Mitigate).
5	**Apply the Intensity Checklist**	Seven control dimensions, calibrated to the score — defines what must be in place before deployment.
6	**Conduct Vendor Diligence**	Specific questions and contractual requirements derived from the risk profile.
7	**Design the Deployment Strategy**	Phased rollout, monitoring gates, and expansion criteria.
8	**Establish Ongoing Governance**	Escalation thresholds, review cadence, and reclassification triggers.

WHY THIS STEP MATTERS

The risk profile of an AI system depends on what it does and whom it serves. Framing the deployment precisely is what makes subsequent steps meaningful rather than generic.

This step produces a Deployment Profile — the reference document the risk team returns to at every subsequent step.

*** Service Level Agreements (SLAs) in technology procurement have traditionally focused on operational performance metrics such as system uptime, response time, and availability. This framework proposes extending the SLA construct to accuracy — requiring vendors to make contractually enforceable commitments about output correctness, not merely system performance.**

DEPLOYMENT PROFILE — CASE APPLICATION

Dimension	Definition / Boundary for This Deployment
Core function	Answers taxpayer questions in natural language via LLM on the Department's public website. Authorized for fact-based queries only: filing deadlines, payment status, form identification, office locations, and general process questions.
User population	General public — anonymous, unauthenticated. Projected 8k–10k queries per week.
Deployment scope	Fact-based queries only. The chatbot must not answer eligibility questions, offer tax advice, or reach definitive conclusions about what a taxpayer qualifies for. Scope enforcement is the central governance challenge.
Output type	Information only — no transaction initiation. Outputs carry implicit authority because they originate from the Department's official website.
Human oversight	None specified in the vendor proposal. Outputs are generated and delivered in real time without human review.
Data handled	User query text only. No taxpayer account or ID data. Query logs retained by vendor — terms not yet reviewed.
Vendor relationship	Third-party LLM vendor. Vendor represents the system is grounded to state tax guidance. Contractual SLA for accuracy requirements ("Accuracy SLA")* and data terms not reviewed.
Legal context	HIGH. Official government website creates strong presumption of authority.
Process stage	Filing: Customer Service (Taxpayer Assistance) — the stage of the tax lifecycle where this deployment operates. Examples: Authentication; Filing Preparation; Electronic Screening; Audit Selection; Assessment & Enforcement; Collections.

For each of the 20 risks, ask: Does a plausible failure pathway exist for this risk in this deployment? Document why risks are excluded.

ACTIVATED RISKS — PROCEED TO SCORING

Risk	Relevance Rationale
Hallucination	Core function requires accurate guidance. Incorrect output reaches taxpayers at scale with no human filter.
Scope Creep	No topic restriction specified. Without scope enforcement, the chatbot will drift into tax advice over time — the central risk of this deployment.
Policy Drift	Accumulated outputs may subtly reshape taxpayer understanding of rules, diverging from official guidance.
Over-Trust / Unauthorized Reliance	Official website creates presumption of authority. Legal reliance risk is live from the first query.
Trust Shock	A visible AI error at scale could cause disproportionate, irreversible damage to public trust.
Lack of Auditability	Department cannot independently reconstruct outputs without vendor cooperation.
Control Dilution	No human review process specified. The proposed deployment has no control environment.
Data Leakage	Query content enters vendor pipeline. Data retention terms unreviewed.
Enterprise Governance Gap / Autonomy Creep	No designated governance owner identified. No cross-functional review structure proposed.

LIMITED OR NOT ACTIVATED

Risk	Status
Explainability	*LIMITED*
Bias / Neutrality	*LIMITED*
Prompt Injection	*LIMITED*
Re-identification	*NOT ACTIVATED*
Workforce and Expertise Erosion	*LIMITED — monitor if expands*
Audit Targeting	*NOT ACTIVATED*
Adversarial Manipulation	*NOT ACTIVATED*
Agentic AI Risk	*NOT ACTIVATED*

LIKELIHOOD × IMPACT MATRIX

IMPACT ↓ / LIKELIHOOD →	LOW	MEDIUM	HIGH
HIGH Impact	MEDIUM	**HIGH**	**HIGH**
MEDIUM Impact	LOW	MEDIUM	**HIGH**
LOW Impact	LOW	LOW	MEDIUM

CONSIDERATIONS WHEN SCORING IN TAX AI CONTEXTS

Scale — Tax AI systems operate at millions of interactions. Even low-probability failures produce large numbers of harmed taxpayers. Weigh volume of exposure when scoring likelihood and impact.

Irreversibility — A penalty incurred, an audit triggered, a return filed on wrong guidance cannot be undone with an apology. Irreversible harms warrant a higher score regardless of probability.

Legal Exposure — Where failure creates legal liability — unauthorized reliance, estoppel, regulatory noncompliance — the impact floor is High. Legal consequences exist independent of how often the failure occurs.

Self-Reinforcing Failure — Some failures become harder to detect over time. Hallucination that erodes trust reduces scrutiny, which allows more hallucination. Treat escalating risk trajectories as higher-severity than static ones.

SCORES — CASE APPLICATION

Risk	Matrix	Score	Key Rationale
Hallucination	High / High	**HIGH**	Scale, irreversibility, and legal exposure
Scope Creep	Med / High	**HIGH**	No scope boundary defined. Self-reinforcing: drift normalizes over time.
Unauthorized Reliance	Med / High	**HIGH**	Legal exposure push factor creates impact floor at High.
Control Dilution	High / High	**HIGH**	No human review process exists — the risk is not that controls will be diluted but that there are none to dilute.
Lack of Auditability	High / High	**HIGH**	Dept. cannot respond to disputes without vendor cooperation. Legal exposure applies.
Trust Shock	Low / High	**HIGH**	Scale + irreversibility issues push Low/High matrix position to HIGH.
Policy Drift via AI	Low / Med	**MEDIUM**	Cumulative; slow-moving.
Vendor Data Leakage	Low / Med	**MEDIUM**	Scale concern applies partially; no account data in queries.
Enterprise Governance Gap	Med / Med	**MEDIUM**	Elevates all other risks.

Tier Assignments for This Deployment

TIER 1 — Deployment Blocker

Controls must be demonstrated adequate before deployment. If a Tier 1 risk cannot be adequately controlled, deployment does not proceed.

TIER 2 — Active Governance Required

Can deploy if baseline controls are in place, but active ongoing governance is required. Shutdown is available as a last resort.

TIER 3 — Monitor and Mitigate

Standard controls and periodic monitoring. If risk escalates materially, reclassify to Tier 2.

Risk	Tier	Rationale
Hallucination	T1	Incorrect outputs on fact-based queries will reach taxpayers at scale with no human filter. Grounding verification and human review sampling are gate conditions.
Scope Creep	T1	A chatbot without a technically enforced scope boundary is ungovernable. Formal scope definition — what the chatbot can and cannot answer — is a pre-deployment structural requirement.
Over-Trust / Unauthorized Reliance	T1	Official website + anonymous users + no disclaimer = immediate legal exposure. General Counsel review of disclaimer language is a predeployment gate condition.
Control Dilution	T1	No human review process is defined. A deployment with no control environment cannot be governed. Establishing minimum oversight structure is required before launch.
Lack of Auditability	T1	Without independent Department access to output logs, the Department cannot respond to disputes or detect systematic errors. A pre-deployment requirement.
Trust Shock	T1	A public trust event, once triggered, is not correctable with additional controls. A communications response plan is required before deployment.
Policy Drift	T2	Requires ongoing monitoring of output patterns for consistency with current guidance. Not a gate condition but requires a defined governance cadence from Day 1.
Data Leakage	T2	Contractual gaps require predeployment remediation and ongoing vendor monitoring.
Enterprise Governance Gap / Autonomy Creep	T2	Designating a governance owner is a concurrent-with-deployment requirement, not a gate condition.

Note: For Tier 1 risks that are scored HIGH, the full intensity column is a predeployment requirement.

Control Dimension	HIGH — Full Intensity	MEDIUM — Standard Intensity	LOW — Baseline Intensity
Human Oversight	Designated reviewer role with explicit authority to flag, hold, or escalate any output before or after delivery. Senior sign-off required before phase expansion. Reviewer is accountable for output quality — not merely for having processed a volume of reviews. No AI output category may be treated as preapproved.	Documented review policy with designated reviewer and sign-off requirement. Director-level authority to restrict deployment if quality indicators breach threshold. Escalation path to senior leadership defined and tested.	Firm policy designating who is responsible for AI output quality. Standard management escalation chain applies. Exception-based escalation if anomaly or complaint threshold is breached.
Review Frequency	Weekly output quality review. Daily monitoring of error flags and complaint volume.	Monthly output quality review. Real-time error flag monitoring. Quarterly governance check-in.	Quarterly spot-check. Automated error flag monitoring. Annual governance review.
Coverage & Sampling	100% logging of all outputs. Statistically valid random sample of ≥25% reviewed by designated reviewer weekly. Sample is drawn randomly — not from flagged or complained-about outputs only. If error rate in sample exceeds threshold, automatic escalation and deployment pause triggered. All complaints reviewed individually regardless of sample.	100% logging. Random sample of ≥10% reviewed monthly. High-severity complaints reviewed individually; others triaged by severity. If two consecutive monthly samples show error rate above threshold, escalation to director level.	100% logging. Spot-check sample reviewed quarterly. Aggregate complaint review unless high-severity. Escalation triggered by significant threshold breach.
Escalation Threshold	Immediate escalation on any single anomaly. C-suite and legal notification. Automatic deployment pause if threshold exceeded.	Escalation after confirmed pattern (3+ incidents). Director-level notification. Deployment review triggered.	Escalation after significant breach. Management notification. Next governance cycle review.

Control Requirements by Dimension (Part 2 of 2)

Control Dimension	HIGH — Full Intensity	MEDIUM — Standard Intensity	LOW — Baseline Intensity
Documentation & Audit Trail	Tamper-proof logging of every output, review decision, and escalation. Retained 7 years. Audit-ready at all times.	Full output log; detailed records for reviewed and flagged outputs. Retained 5 years.	Summary log. Detailed records for escalated items. Retained 3 years.
Vendor Accountability	Contractual audit rights. Third-party accuracy testing annually. Vendor liability provisions. SLA with financial penalties.	SLA with accuracy metrics. Quarterly vendor performance reports. Right to terminate for failure.	Standard SLA. Annual vendor summary. Right to terminate for material breach.
Deployment Scope & Phasing	Highly restricted Phase 1. Documented scope definition technically enforced. Senior sign-off before each phase expansion. Full audit before Phase 3.	Moderate Phase 1 with monitoring gates. Director-level sign-off before expansion.	Standard phased rollout. Management sign-off before expansion.

For Tier 1 risks, vendor diligence is a gate condition. Unsatisfactory answers are grounds to pause the procurement — not to proceed with caveats.

ASK YOUR VENDOR — BEFORE SIGNING	
Relates To	**Question**
Hallucination	What is the documented hallucination rate on state-specific tax content — not general benchmarks? Can you demonstrate this on our content corpus?
Hallucination	How is the model grounded to authoritative state tax sources, and what is the update process when state law or guidance changes?
Hallucination / Scope	What does the system do when a taxpayer question approaches or crosses the line into tax advice or eligibility determinations? Does it decline, escalate, or answer anyway?
Scope Enforcement	Can the system be technically configured to refuse out-of-scope questions — and how is that restriction enforced in the model, not just in the interface?
Auditability	Can the Department independently access a tamper-proof log of every output the system generates, without relying on vendor cooperation or a data request process?
Data Leakage	Is taxpayer query content retained in vendor systems? For what period, under what access controls, and is it used for model training?
Unauthorized Reliance	Has legal review been conducted on disclaimer requirements for a public-sector tax guidance deployment? Will you support the Department in implementing compliant disclaimers?
Context Retention	How does the system handle earlier context as a conversation extends – does it retain, summarize, or lose prior exchanges (sometimes referred to as the “goldfish” challenge)? Are users alerted when earlier context is no longer fully retained?
Performance Standards	What Accuracy SLA applies? What financial remedy is available if systematic errors exceed the agreed threshold?

ILLUSTRATIVE CONTRACT TERMS	
Requirement	**Substance**
Accuracy SLA	Vendor warrants outputs consistent with authoritative state tax guidance within a defined error tolerance. Material accuracy failure triggers financial remedy and right to immediate termination.
Data Retention & Use Prohibition	Vendor may not retain query content beyond 90 days and may not use query content for model training without explicit written consent.
Scope Enforcement Warranty	Vendor warrants that the system will technically decline or escalate questions that exceed the Department's authorized scope definition and will document how that enforcement operates.
Independent Audit Rights	Department retains the right to audit vendor systems, testing records, and grounding methodology with 30 days' notice. Vendor must submit to annual third-party accuracy testing on state-specific content.
Audit Trail Access	Vendor must provide the Department with independent, direct access to a tamper-proof log of every output, not contingent on vendor cooperation or data request processes.
Incident Notification	Vendor must notify the Department within 24 hours of discovering any systematic output error, grounding failure, scope boundary breach, or data access incident.
Change Notification	Vendor must notify the Department 60 days before any material change to the model, training data, grounding methodology, or scope enforcement mechanism.
Context Retention Disclosure	Vendor must document how the system manages conversational context across multi-turn interactions, including any degradation in retention as session length increases, and must notify Department of any changes to context handling in model updates.

PHASE 1 — Restricted Launch

AUTHORIZED SCOPE

Fact-based queries only: filing deadlines, payment due dates, office locations, general form identification. No credit eligibility, no scenario-based guidance, no questions requiring interpretation of tax law.

GATE CONDITIONS FOR EXPANSION

- Hallucination rate ≤1% on authorized query types in independent testing
- Scope enforcement demonstrated — system declines out-of-scope questions in testing
- 100% logging with Dept-accessible audit trail confirmed
- Legal disclaimer approved by General Counsel
- Escalation protocol tested and operational
- Deputy Commissioner sign-off

REQUIRED CONTROLS

100% logging. ≥20% weekly human sample. Real-time error flag monitoring. Disclaimer on every output. Scope enforcement tested monthly.

PHASE 2 — Expanded Scope

AUTHORIZED SCOPE

Add common procedural and process questions (how to file an amended return, how to request an extension). Still no credit eligibility, no interpretation of qualifying criteria, no scenario-based guidance.

GATE CONDITIONS FOR EXPANSION

- Phase 1 hallucination rate maintained at ≤1% for 90 consecutive days
- Scope boundary holding — no out-of-scope responses detected in Phase 1 monitoring
- Director-level sign-off
- Legal counsel confirms disclaimer adequacy for expanded scope

REQUIRED CONTROLS

All Phase 1 controls. Monthly output quality review. Weekly scope drift monitoring — flag any output approaching eligibility guidance.

PHASE 3 — Full Authorized Scope

AUTHORIZED SCOPE

Expand to full authorized scope as defined in the deployment profile. Any further scope expansion beyond this point requires restarting the relevance screening and scoring process from Step 2.

GATE CONDITIONS FOR EXPANSION

- Phase 2 hallucination rate ≤0.5% for 90 consecutive days
- No scope boundary breaches in Phase 2
- Trust monitoring shows no adverse public signals
- Full audit of Phase 2 operations completed
- Deputy Commissioner and General Counsel sign-off

REQUIRED CONTROLS

All prior controls. ≥20% weekly sampling. Quarterly vendor audit. Automatic pause if hallucination rate exceeds 1% in any 30-day window.

ELEMENT 1 — ACCOUNTABILITY

Role	Responsibility
Chief Information Officer/Designated AI Governance Owner	Accountable for overall risk profile and pre-deployment verification sign-off for each phase. Must have access to expertise in data science and statistical methodology, not only technology infrastructure; a Chief Data Officer or equivalent quantitative role should have authority over model evaluation, training data fitness, and bias assessment.
Deputy Commissioner, Tax Administration	Accountable for tax content accuracy and for authorizing phase expansions based on gate condition evidence.
General Counsel	Accountable for Over-Trust / Unauthorized Reliance risk and disclaimer adequacy. Required sign-off before Phase 2 and Phase 3.
IT Risk Manager	Accountable for vendor controls, audit trail access, error flag monitoring, and weekly technical governance review.
Tax Law Review Team	Accountable for human review sampling in Phase 2 and Phase 3 and for escalation of out-of-scope outputs.
Director of Communications	Accountable for Trust Shock monitoring — monthly review of public channels and escalation if adverse coverage emerges.

ELEMENT 2 — INDICATORS AND THRESHOLDS

Indicator	Threshold and Response
Hallucination rate (weekly sample)	>1% in any 30-day window → immediate deployment pause and root cause review. >0.5% for two consecutive weeks → governance review and scope restriction.
Scope boundary breaches	Any output that approaches or crosses into eligibility guidance or tax advice → immediate review of all outputs on that topic. Repeated breaches → Phase restriction.
Complaint volume citing chatbot	>5 complaints in any week → escalation to Deputy Commissioner. >20 in any month → legal notification.
Audit findings contradicting output	Any finding contradicting a chatbot answer → review of all outputs on that topic area and potential scope restriction.
Trust monitoring	Monthly review of public media and social channels. Any negative coverage → Communications Director notification within 24 hours.
Vendor performance metrics	Two consecutive months below SLA accuracy threshold → formal vendor notice and consideration of contract termination.
Scope expansion requests	Any proposal to expand the chatbot's authorized query scope → new relevance screening and scoring exercise before approval.

ELEMENT 3 — ESCALATION PROTOCOL

Level 1 IT Risk Manager	Trigger: Any single anomaly or indicator threshold breach. **Response: Response within 24 hours.**
Level 2 CIO + Deputy Commissioner	Trigger: Confirmed pattern (3+ incidents or statistical anomaly) or Level 1 unresolved within 48 hours. **Response: Deployment scope restriction considered.**
Level 3 Commissioner + General Counsel	Trigger: Level 2 unresolved, legal claim received, out-of-scope output reported publicly, or media coverage of chatbot error. **Response: Deployment pause or termination considered.**
Level 4 Deployment Shutdown	Trigger: Tier 1 risk materializes and controls cannot be made adequate. **Response: This is the defined protocol — not a last resort.**

ELEMENT 4 — LIFECYCLE TRIGGERS

Trigger	What It Requires
Tier 3 → Tier 2	Triggered when a Tier 3 risk shows sustained threshold breaches across two review cycles.
Tier 2 → Tier 1	Triggered when a Tier 2 risk develops a failure mode that cannot be adequately controlled through intensified governance — for example, Data Leakage if a confirmed data incident occurs.
Phase Expansion	Requires completion of all gate conditions for the current phase and formal documentation of evidence before the governance owner approves expansion. Gate conditions are not negotiable.
Scope Expansion	Any proposed expansion of the chatbot's authorized query scope — beyond the deployment profile established in Step 1 — triggers a full restart of Steps 2 through 4 before approval.
Deployment Shutdown (Post-Launch)	Required if a Tier 1 risk materializes and controls cannot be made adequate. Determination made jointly by CIO, Deputy Commissioner, and General Counsel. Commissioner notified immediately.

Appendix C
AI Risk Register for Tax Preparers

PRACTICE RISKS

# RISK	DESCRIPTION	SCENARIO	ILLUSTRATIVE CONTROLS
1 Hallucinated Advice	The AI suggests a tax position, deduction, or interpretation that is confidently stated but factually wrong or legally unsupported. The error enters the return because it sounds authoritative.	*An AI tool suggests a home office deduction that doesn't meet the exclusive-use test under IRC §280A. The preparer incorporates it without verifying the underlying requirement. The deduction is disallowed on audit.*	Mandatory human validation of all AI-suggested positions before filing; require AI outputs to cite specific IRC section or authority; documentation of which suggestions were accepted, modified, or rejected.
2 Client Data Leakage	Sensitive client information — income, identity, business details — is exposed to or retained by the AI vendor through prompts, logs, or training pipelines in ways the client never consented to and the preparer may be unaware of — and the aggregation of that data creates concentrated breach targets.	*A preparer enters a client's full Schedule C details into a commercial AI tool. The vendor's terms permit use of query content for model training. The client's business financials become part of the vendor's training data without consent.*	Vet AI vendors for data retention and training pipeline practices before use; contractual prohibition on vendor use of client data; data minimization — enter only what is necessary; client disclosure of AI tools used.
3 Incorrect Document Extraction	The AI misreads, misclassifies, or omits information from source documents — Forms W-2, 1099, K-1 — producing errors that flow silently into the return without triggering any obvious flag.	*An AI tool misclassifies a box 12 code on a W-2, treating a pre-tax health premium as taxable wages. The error flows into the return without any alert. The discrepancy is discovered only when the taxpayer receives a notice.*	Independent verification of AI-extracted figures against source documents for all returns; spot-check audits of AI document extraction accuracy; escalation protocol when source documents are complex or ambiguous.

# RISK	DESCRIPTION	SCENARIO	ILLUSTRATIVE CONTROLS
4 Failure of Professional Judgment	The preparer accepts an AI output without independent verification, over-trusting the confident tone of the response rather than checking the underlying authority.	*A junior associate accepts an AI-generated Schedule C deduction list because the output sounds confident and is well-formatted. He does not independently verify any of the cited authorities. Two positions are unsupportable and are disallowed on audit.*	Mandatory independent verification policy (AI is assistive vs. determinative); training on AI limitations and common hallucination patterns in tax; escalation rules for novel positions regardless of AI confidence; firm policy that AI output is a starting point, not a conclusion.
5 Misplaced Authority / Over-Trust	The preparer or client attributes more legal weight to AI output than it warrants — treating a probabilistic suggestion as a definitive answer and failing to recognize the limits of what the tool can reliably determine.	*A junior preparer accepts an AI-generated deduction recommendation because it appears confident and well-worded, without independently verifying the authority. The client, informed of the recommendation, treats it as settled guidance.*	Mandatory human validation; source-linked explanations required; training on AI limitations; escalation rules for uncertain or novel positions; clear policy that AI is assistive, not determinative.
6 Control Dilution	The speed and volume AI introduces causes existing review and sign-off processes to become less rigorous in practice, even if they remain unchanged on paper. More returns get processed with less scrutiny per return.	*AI tools accelerate return preparation, but existing partner review processes do not adapt. Partners spend less time per return reviewing AI-assisted work, increasing the likelihood that unsupported positions go undetected.*	Redesigned review controls for AI-enabled workflows; sampling and spot checks scaled to volume; documented sign-off requirements that require genuine analysis, not just approval; workload thresholds.

LEGAL & REGULATORY RISKS

# RISK	DESCRIPTION	SCENARIO	ILLUSTRATIVE CONTROLS
7 Malpractice Exposure	A filing error traceable to AI output results in penalties, interest, or audit findings, and the preparer faces professional liability for a return they signed but did not fully verify.	*Penalties are assessed after audit. The AI vendor's terms disclaim liability for tax advice. The client argues the preparer — who signed the return — is responsible. The preparer has no documentation showing the AI was used or how it was reviewed.*	Full audit trail of AI suggestions and preparer decisions; engagement letter language addressing AI use and responsibility allocation; vendor contracts clarifying liability; signing preparer independently verifies all AI-drafted positions before filing.
8 Liability Ambiguity	When an AI-assisted position is challenged, the allocation of responsibility between the signing preparer and the AI vendor is legally unclear.	*A preparer uses an AI tool to identify eligible deductions. The system recommends a deduction that is disallowed on audit. When penalties are assessed, it is unclear whether responsibility lies with the preparer who signed the return or the vendor whose system generated the recommendation.*	Engagement letters explicitly address AI use and responsibility allocation; vendor contracts clarify vendor liability for AI-generated recommendations; documentation of preparer's independent review of all AI-suggested positions.

LEGAL & REGULATORY RISKS

# RISK	DESCRIPTION	SCENARIO	ILLUSTRATIVE CONTROLS
9 Regulatory Noncompliance	The firm's use of AI violates professional standards, licensing rules, or engagement obligations — including where unlicensed staff rely on AI for legal interpretations that constitute unauthorized practice, or where the tools used fail to meet applicable professional conduct requirements.	*Unlicensed staff use an AI research tool to answer client questions about tax law interpretation, relying on AI output as a substitute for professional judgment. A state board of accountancy later determines this constitutes unauthorized practice.*	Compliance review of AI tools against applicable professional standards before deployment; supervision requirements for unlicensed staff using AI; training on scope of authorized practice; annual review of AI use practices against evolving regulatory guidance.

# RISK	DESCRIPTION	SCENARIO	ILLUSTRATIVE CONTROLS
10 Client Trust Erosion	The client discovers AI was used in preparing their return after a disallowed position or error without prior disclosure, damaging the relationship and potentially triggering a legal or fee dispute.	*After a disallowed deduction, the client learns through the appeals process that an AI tool generated the position. No one had disclosed this. She terminates the engagement, files a bar complaint about nondisclosure, and disputes the fee for the affected year.*	Client disclosure policy — inform clients when AI materially contributes to their return; clear role definition: AI assists the preparer, preparer is responsible; proactive communication protocol before AI tools are deployed.
11 Reputational Harm	A pattern of AI-assisted errors or a single high-profile failure becomes known in the market, damaging the firm's standing with existing and prospective clients beyond the individual affected engagement.	*A firm's AI-assisted errors result in a state board investigation that becomes public. The firm's name appears in a trade publication in connection with AI-related malpractice.*	Quality control systems that identify error patterns before they accumulate; incident response protocol for high-profile failures; proactive comms to affected clients; monitoring of professional board activity and trade coverage.
12 Lack of Transparency	The firm has no clear policy on how AI is used, disclosed to clients, or documented internally — leaving clients uninformed, the firm's decision-making unrecorded, and the firm exposed if AI use later becomes a point of contention.	*A firm uses AI across all return preparation but has no disclosure policy, no documentation standard for AI suggestions, and no client communication protocol. When an AI-assisted position is challenged, the firm cannot demonstrate what was AI recommended vs. preparer verified.*	Documented AI use policy covering disclosure, review requirements, and documentation standards; client communication protocol; internal documentation requirements for AI suggestions accepted, modified, or rejected; periodic review against evolving professional standards.

WORKFORCE & STRATEGIC RISKS

# RISK	DESCRIPTION	SCENARIO	ILLUSTRATIVE CONTROLS
13 Skill Erosion	Staff develop expertise in operating AI tools rather than in the underlying tax law those tools are applying. Loss of independent analytical capability to catch AI errors or handle complex matters without AI assistance.	*Three years into using an AI research tool, the firm's junior associates cannot independently analyze a novel partnership allocation issue. When the vendor discontinues the product, the firm has a capability gap it cannot quickly close.*	Mandatory training on underlying tax law, not just tool operation; annual capability assessment — can the firm handle complex matters without this tool; explicit professional development requirements independent of AI tool use.
14 Vendor Dependence	The firm becomes operationally reliant on a single AI vendor to the point where a vendor outage, pricing change, or contract termination would materially disrupt client service. The firm no longer has the internal capability to operate without the tool.	*A vendor discontinues a core AI research product with 30 days notice. The firm has no alternative workflow and no staff trained to perform the function manually. Three weeks of client service disruption follow before a replacement is found.*	Vendor diversification — avoid single-vendor dependency for core functions; maintain documented manual workflows for all AI-assisted processes; annual assessment of vendor dependency risk; contractual protections including data portability and reasonable notice periods.
15 Institutional Knowledge Loss	Experienced preparers retire or leave earlier than they otherwise would as AI reduces demand for their expertise. The tacit knowledge they carry — about complex clients, edge cases, and professional judgment — leaves with them and is not replaced.	*Senior staff retire early as AI handles work previously requiring their expertise. When a long-tenured client faces a complex restructuring, no one at the firm has the institutional knowledge of the client's history or the independent judgment to advise without AI assistance.*	Knowledge retention programs; mentorship structures that transfer tacit knowledge before senior staff depart; succession planning for complex client relationships; retention incentives for senior staff in critical roles.

Appendix D
Applying the AI Risk Register for Tax Preparers
Case Study #2

Case Study 2 — Tax Preparer

The Scenario

A vendor proposes an AI tool integrated into the firm's existing document management workflow. The system is capable of:

- Ingesting and extracting data from client financial documents, W-2s, 1099s, and prior-year returns
- Suggesting itemized deductions and credits based on document analysis
- Drafting initial tax positions and return sections in natural language
- Generating natural language explanations of suggested positions for preparer review

The Eight-Step Methodology at a Glance

Step	Action	What It Produces
1	**Frame the Deployment**	Scope, function, client population, and deployment boundaries defined
2	**Screen for Relevance**	A filtered list: which of the 15 risks apply to this deployment
3	**Score Each Relevant Risk**	High / Medium / Low for each risk within each stack layer, using the Likelihood × Impact matrix and push factors
4	**Assess the Cascade ★**	A layer-by-layer view of how uncontrolled risk at one level propagates upward — replaces tier classification
5	**Apply the Intensity Checklist ★**	Control requirements per stack layer, calibrated to layer score — what must be in place before any client returns are processed
6	**Conduct Firm and Vendor Diligence**	Questions and contractual requirements for the firm's own readiness and the vendor's product
7	**Design the Adoption Strategy ★**	Phased rollout anchored by the research-only / position-drafting boundary; gate conditions not timelines
8	**Establish Ongoing Governance**	Escalation thresholds, review cadence, and reclassification triggers

★ *Steps 4, 5, and 7 are adapted for the preparer context. No Tier 1 / Tier 2 / Tier 3 designations as used in the Tax Authority methodology are used in this framework.*

WHY THIS STEP MATTERS

The risk profile of an AI tool depends on what the system is authorized to do and at what stage of return preparation it operates.

A tool authorized only to extract document data carries a fundamentally different risk profile than one authorized to draft tax positions. Framing the deployment precisely — before evaluating any risk — is what makes subsequent steps actionable.

Key distinction for preparers: the signing preparer bears full professional responsibility for every position on the return, regardless of whether it was AI-suggested.

DEPLOYMENT PROFILE — CASE APPLICATION

Dimension	Definition / Boundary for This Deployment
Core function	Ingests client financial documents; suggests deductions and credits; drafts initial tax positions and return sections; generates natural language explanations for preparer review.
User population	All preparers at the firm — including junior staff. No role-based access restrictions specified in the vendor proposal.
Client population	Individual and small business clients across all complexity levels. No complexity-based restrictions specified.
Deployment scope	Document extraction through position drafting — the highest-stakes stage of return preparation. AI outputs are formatted for direct filing incorporation.
Output type	Drafted tax positions formatted as complete, citation-supported suggestions — increasing the risk of uncritical acceptance.
Human oversight	Not specified beyond standard preparer sign-off. No independent verification requirement or escalation protocol defined in the vendor proposal.
Data handled	Full client financial documents — W-2s, 1099s, prior-year returns, supporting materials. Highly sensitive. Vendor data retention terms not reviewed.
Professional context	HIGH. Signing preparer bears full professional responsibility. No vendor liability shield exists for malpractice exposure from AI-generated errors.

The proposed deployment activates all 15 risks in the register. A tool authorized to ingest full client documents, suggest deductions, and draft tax positions — with no scope restriction, no human oversight requirement, and no phase boundary — presents a plausible failure pathway for every risk in the stack.

The next slide scores each risk individually using the Likelihood × Impact matrix. The slide after that maps how they connect.

THE STACK

Risk escalates upward: a Technical failure unchecked cascades through Practice to Legal.

Score Technical first — an uncontrolled HIGH at Technical is itself a push factor for Practice above it.

LIKELIHOOD × IMPACT MATRIX

IMPACT ↓ / LIKELIHOOD →	LOW	MEDIUM	HIGH
HIGH Impact	MEDIUM	**HIGH**	**HIGH**
MEDIUM Impact	LOW	MEDIUM	**HIGH**
LOW Impact	LOW	LOW	MEDIUM

CONSIDERATIONS WHEN SCORING IN TAX AI CONTEXTS

Factor	Preparer Application
Scale	AI across all clients multiplies any single error. Low-frequency hallucinations produce material numbers at volume.
Irreversibility	A filed return cannot be unfiled. A penalty assessed cannot be unassessed. A client relationship damaged may not be recovered.
Legal Exposure	Signing preparer bears full responsibility. No vendor liability shield for AI-generated errors.
Self-Reinforcing Failure	Automation bias that goes uncorrected becomes normalized practice — future failures become more likely.

SCORE	RISKS IN THIS BUCKET	SHARED RATIONALE
HIGH 8 of 15	Hallucinated Advice · Client Data Leakage · Incorrect Document Extraction · Failure of Professional Judgment · Misplaced Authority / Over-Trust · Control Dilution · Malpractice Exposure · Skill Erosion	All involve a direct failure pathway to a filed return — through incorrect AI output, absent human oversight, or missing documentation. Scale, irreversibility, and legal exposure apply to each. All three push factors elevate these above the matrix placement alone.
MEDIUM 7 of 15	Liability Ambiguity · Regulatory Noncompliance · Lack of Transparency · Client Trust Erosion · Reputational Harm · Vendor Dependence · Institutional Knowledge Loss	Significant but manageable with defined ongoing governance. Most are downstream consequences — if HIGH risks are controlled, several of these are substantially mitigated. Require baseline controls before deployment and active monitoring throughout.
LOW 0 of 15	*None*	*No risk in this deployment scores LOW. The breadth of scope proposed by the vendor leaves no plausible failure pathway dormant.*

The purpose of Step 4 is to identify where in that chain a single control — if enforced — stops the harm from reaching the next layer.

Risk	Layer	Classification	Cascade Note
Hallucinated Advice	Technical	**Must Control**	Origin of the primary cascade. Uncontrolled hallucination at Technical directly elevates Practice layer risks above it.
Client data leakage	Technical	**Must Control**	Independent of cascade — creates immediate legal liability. Pre-deployment contractual remediation required.
Incorrect Document Extraction	Technical	**Must Control**	Silent errors from document misreads flow into the return without flags. Independent verification of all AI-extracted figures against source documents is required before any filing.
Failure of professional judgment	Practice	**Must Control**	The critical cascade interrupt. If independent verification is enforced here, a Technical hallucination is caught before it becomes a filed position.
Misplaced Authority / Over-Trust	Practice	**Must Control**	Directly enables the cascade: a preparer who over-trusts AI output bypasses the only Practice-layer control on Technical errors.
Control Dilution	Practice	**Must Control**	Systematic control dilution at Practice is structurally equivalent to having no controls. Requires redesign before deployment.
Malpractice exposure	Legal & Regulatory	**Must Control**	Legal consequence of uncontrolled cascade. Controls at Legal layer are necessary but not sufficient — primary controls are Technical and Practice.
Skill Erosion	Workforce & Strategic	**Must Control**	Begins immediately upon deployment and is self-reinforcing. Requires affirmative controls from Day 1.

How to read this table: Risks classified Must Control are those where failure directly enables harm to a client or the firm, or where failure removes the ability to catch harm downstream. Risks classified Active Governance are significant but do not individually trigger the primary cascade — they require ongoing oversight but are not deployment blockers on their own.

The primary cascade in this deployment runs from bottom to top of the stack: a Technical failure (hallucinated output) that passes through Practice unchecked (preparer accepts without verification) produces a Legal consequence (malpractice exposure from a signed return). Controlling the Practice layer — specifically, enforcing independent verification as a genuine requirement rather than a formality — is the single most effective intervention point in the chain.

The purpose of Step 4 is to identify where in that chain a single control — if enforced — stops the harm from reaching the next layer.

Risk	Layer	Classification	Cascade Note
Liability Ambiguity	Legal & Regulatory	**Active Governance**	Contractual remediation required pre-deployment. Ongoing monitoring of vendor contract performance throughout deployment.
Regulatory noncompliance	Legal & Regulatory	**Active Governance**	Compliance review required pre-deployment. Ongoing monitoring as professional standards evolve.
Lack of Transparency	Business	**Active Governance**	Client disclosure policy must be established before deployment.
Vendor Dependence	Workforce & Strategic	**Active Governance**	Develop contingency plan within 12 months.
Client trust erosion	Business	**Active Governance**	Downstream monitoring. Escalate if Legal layer failures materialize.
Reputational Harm	Business	**Active Governance**	Market-level consequence of Legal layer failures accumulating. Quality control systems and incident response protocol required from deployment outset.
Institutional Knowledge Loss	Workforce & Strategic	**Monitor & Mitigate**	Slow-moving. Multi-year horizon. Reclassify if skill degradation metrics deteriorate.

How to read this table: Risks classified Must Control are those where failure directly enables harm to a client or the firm, or where failure removes the ability to catch harm downstream. Risks classified Active Governance are significant but do not individually trigger the primary cascade — they require ongoing oversight but are not deployment blockers on their own.

For risks classified Must Control Before Deployment, every item in the full-intensity column must be demonstrably in place before any client returns are processed with the tool.

Control Dimension	HIGH — Full Intensity	MEDIUM — Standard Intensity	LOW — Baseline Intensity
Human Oversight	Mandatory independent verification of every AI-suggested position — not approval, but genuine analysis. Signing preparer confirms every AI-drafted position before filing. No AI-only determination of any tax position.	Documented review policy with sign-off. Escalation for complex or novel positions. Senior review for high-value clients.	Firm policy that AI output is a starting point. Exception-based escalation. Standard sign-off process.
Review Frequency	Weekly quality review of AI-assisted returns. Real-time error flag monitoring during busy season. Monthly review of override rates by preparer.	Monthly quality check. Escalation pattern review quarterly.	Annual review of AI-assisted work quality. Periodic spot-check.
Coverage & Sampling	100% logging of every AI suggestion and preparer decision (accept, modify, reject). All audit findings on AI-assisted returns reviewed individually. Override rate tracked by preparer.	100% logging. Sample review of AI suggestions monthly. High-severity errors reviewed individually.	100% logging. Aggregate error review. Annual sample of AI-assisted returns.
Escalation Threshold	Immediate escalation for any AI-suggested position the preparer cannot independently verify. Novel or uncertain positions escalate to senior review regardless of AI confidence. Error rate threshold triggers tool suspension.	Escalation for complex or high-value positions. Notification if error pattern emerges. Review triggered by client dispute.	Escalation for significant errors. Management notification after confirmed pattern.

For risks classified Must Control Before Deployment, every item in the full-intensity column must be demonstrably in place before any client returns are processed with the tool.

Control Dimension	HIGH — Full Intensity	MEDIUM — Standard Intensity	LOW — Baseline Intensity
Documentation & Audit Trail	Full log of every AI suggestion, preparer decision, and override rationale on every return. Retained 7 years. Audit-ready at all times. Engagement letter documents AI use and responsibility allocation.	Full log retained minimum 5 years. Detailed records for disputed or complex positions. Engagement letter references AI use.	Standard record retention per professional standards. Basic documentation of AI-assisted positions.
Vendor Accountability	Contractual Accuracy SLA with financial penalties for systematic errors. Vendor must notify firm 60 days before any model update affecting tax position outputs. Vendor contract addresses liability allocation for AI-suggested positions.	SLA with accuracy metrics. Quarterly vendor performance review. Vendor liability clauses reviewed. Right to terminate for material error rate.	Standard vendor contract. Annual review. Right to terminate for breach.
Deployment Scope & Phasing	Phase 1 restricted to document summarization and research support only — no AI-drafted positions on client returns. Phase 2 gate conditions must be verified before AI positions are incorporated in any filed return. No exceptions.	Moderate Phase 1 with monitoring gates. Director-level sign-off before expansion to position drafting.	Standard phased rollout. Management sign-off before expansion.

Diligence operates in two modes: questions the firm must ask itself before authorizing use and questions that must be answered by the vendor before signing.

ASK YOUR ORGANIZATION

Relates To	Question
Human Oversight	Do we have a documented policy defining how AI-suggested positions must be independently verified — not just approved — and is it actively enforced?
Human Oversight	Have we trained all staff on this tool's specific failure modes — particularly how it produces confident, well-formatted outputs that are factually incorrect?
Human Oversight	Can our sign-off process demonstrate genuine independent analysis? How will we audit compliance with that requirement?
Documentation	Do we have an audit trail system that logs every AI suggestion and the preparer's decision — producible in an IRS examination or client dispute?
Legal	Has counsel reviewed our engagement letters to address AI use, responsibility allocation, and client disclosure before the tool is used on any client matter?
Workforce	Can our staff perform this work without AI right now — and do we have a plan to maintain that capability as tool use increases?
Governance	Who in the firm owns AI governance with authority to restrict or suspend the tool if an error pattern emerges?

ASK YOUR VENDOR

Relates To	Question
Hallucination	What is the documented error rate for tax-specific deduction and position suggestions on returns of the complexity our firm handles — not general benchmarks?
Hallucination	Does the tool produce source citations traceable to a specific IRC section or ruling for every suggested position, or does it present conclusions without verifiable attribution?
Hallucination	How does the tool behave on questions outside its training scope — does it flag uncertainty or generate a plausible but unsupported answer?
Model Updates	Will you notify us before any model update that could affect output accuracy, and provide a re-validation window before continued use on client returns?
Liability	What does your contract say about liability if an AI-suggested position is disallowed on audit? What indemnification, if any, does the vendor provide?
Data	Where is client data stored, who has access, how long is it retained, and is it used for model training or improvement?
Accuracy SLA	Will you commit to a contractual Accuracy SLA — and what financial remedy is available to the firm if systematic errors exceed the agreed threshold?
Context Retention	If a preparer is working through a complex client situation across multiple prompts, at what point does earlier context degrade or drop, and is the preparer alerted? (sometimes referred to as the “goldfish” challenge)

ILLUSTRATIVE CONTRACT TERMS

Requirement	Substance
Accuracy SLA	Vendor warrants that the system produces positions consistent with current tax law within a defined error tolerance. Material accuracy failure triggers financial remedy and right to immediate termination.
Liability Provision	Contract must specify responsibility allocation for AI-suggested positions disallowed on audit. Vendor liability for systematic errors must be explicit — not left to implication.
Data Protection	Vendor may not retain client data beyond the engagement period and may not use client data for model training without explicit written consent.
Model Update Notification	Vendor must notify the firm 60 days before any material change to the model, training data, or citation methodology, and provide a re-validation window before continued use.
Independent Audit Rights	Firm retains the right to commission independent accuracy testing on tax-specific content and to audit vendor systems with 30 days' notice.
Incident Notification	Vendor must notify the firm within 24 hours of discovering any systematic output error, accuracy threshold breach, or client data access incident.
Scope and Use Limitation	Contract must specify authorized use cases and prohibit vendor-side expansion of capabilities without written firm approval.
Context Retention Disclosure	Vendor must document how the system manages conversational context, including any degradation in retention as session length increases, and must notify the firm of any change to context handling in model updates.

Note on Professional Responsibility: A vendor contract cannot transfer the signing preparer's professional responsibility. These contractual requirements reduce exposure and create remedies — they do not change who bears responsibility for positions filed on a client's return.

The boundary between Phase 1 and Phase 2 — research support only vs. AI-drafted positions on client returns — is the primary structural control for the conclusions reached in Step 4.

PHASE 1 — Research & Summarization Only

AUTHORIZED SCOPE

Document summarization, data extraction, and research support only. The tool may identify relevant code sections and summarize document content. No AI-drafted tax positions. No AI-suggested deductions incorporated into any client return.

GATE CONDITIONS FOR EXPANSION

- Firm independent verification policy documented, trained, and tested
- 100% logging confirmed operational
- Engagement letters updated and approved by counsel
- Vendor Accuracy SLA and data protection terms in executed contract
- Hallucination rate on firm's client return types verified at ≤2% in independent testing
- Managing Partner sign-off

REQUIRED CONTROLS

100% logging of all AI outputs. Weekly quality review. All staff trained on tool failure modes before first use. Disclaimer on all AI-generated research summaries.

PHASE 2 — Deduction Suggestions

AUTHORIZED SCOPE

Add AI-suggested deductions and credits, subject to mandatory independent verification by the preparer before any suggestion is incorporated into a filed return. AI-drafted return sections not yet authorized.

GATE CONDITIONS FOR EXPANSION

- Phase 1 gate conditions met and documented
- Hallucination rate maintained at ≤2% for 60 consecutive days
- Override rate confirms genuine independent verification — rate below 5% per preparer triggers mandatory review
- Client disclosure language added to engagement letters
- Senior partner and counsel sign-off

REQUIRED CONTROLS

All Phase 1 controls. Every AI-suggested deduction logged with preparer decision. Monthly error pattern review. Novel or uncertain positions escalate to senior review regardless of AI confidence.

PHASE 3 — Full Position Drafting

AUTHORIZED SCOPE

AI-drafted initial tax positions and return sections authorized, subject to mandatory signing preparer review and independent verification of every drafted position before filing.

GATE CONDITIONS FOR EXPANSION

- Phase 2 error rate verified at ≤1% for 90 consecutive days
- No unresolved client disputes arising from AI-suggested positions
- Annual capability assessment confirms staff can perform AI-assisted tasks without the tool
- Full audit of Phase 2 operations completed
- Managing Partner and General Counsel sign-off

REQUIRED CONTROLS

All Phase 1 and Phase 2 controls. 100% of AI-drafted positions independently verified — documented in audit trail. Quarterly independent accuracy testing. Automatic tool suspension if error rate exceeds 1% in any 30-day window.

ELEMENT 1 — ACCOUNTABILITY

Role	Responsibility
Managing Partner	Designated AI governance owner. Accountable for overall risk profile and authorizing phase expansions based on gate condition evidence.
Tax Practice Director	Accountable for tax position accuracy and pre-deployment verification sign-off for each phase.
General Counsel / Outside Counsel	Accountable for malpractice exposure, engagement letter adequacy, and vendor contract compliance. Required sign-off before Phase 2 and Phase 3.
IT / Data Security Lead*	Accountable for vendor data controls, access monitoring, and breach notification protocols.
Quality Review Lead	Accountable for weekly quality review of AI-assisted returns, override rate monitoring, and escalation pattern reporting.

ELEMENT 2 — INDICATORS AND THRESHOLDS

Indicator	Threshold and Response
Hallucination / error rate	>2% in any 30-day window → immediate tool suspension. >1% for two consecutive weeks → scope restriction to Phase 1.
Override rate by preparer	Override rate below 5% for any preparer → mandatory review. Rate that low signals rubber-stamping, not independent verification.
Audit disallowances on AI-assisted returns	Any disallowance on an AI-suggested position → review of all similar positions filed in prior 12 months.
Client disputes citing AI output	Any dispute referencing an AI-generated position → Legal layer escalation and review of client disclosure compliance.
Annual skill assessment	Declining ability to perform AI-assisted tasks without the tool → reclassify Skill Degradation to Must Control and mandatory training.
Scope expansion requests	Any proposal to expand use cases → new relevance screening and scoring exercise before approval.
Vendor performance	Two consecutive months below SLA accuracy threshold → formal vendor notice and consideration of termination.

***Whoever holds the AI governance owner role must have direct access to expertise in data science and statistical methodology. AI applications are advanced statistical models; governance without quantitative competency is insufficient to detect bias, evaluate training data fitness, or assess model appropriateness.**

ELEMENT 3 — ESCALATION PROTOCOL

Level 1 Quality Review Lead	Trigger: Any single anomaly, indicator threshold breach, or preparer escalation. **Response: Response within 24 hours.**
Level 2 Tax Practice Director + Managing Partner	Trigger: Confirmed error pattern (3+ incidents) or Level 1 unresolved within 48 hours. **Response: Tool scope restriction considered.**
Level 3 General Counsel	Trigger: Level 2 unresolved, client dispute received, audit disallowance on AI-suggested position, or data security incident. **Response: Tool suspension considered.**
Level 4 Tool Suspension	Trigger: Must Control risk materializes and controls cannot be made adequate. **Response: This is the defined protocol — not a last resort.**

ELEMENT 4 — LIFECYCLE TRIGGERS

Trigger	What It Requires
Active Governance → Must Control	Triggered when an Active Governance risk develops a cascade connection to a Must Control risk — for example, Vendor Dependence escalating when a vendor outage reveals the firm has no manual workflow alternative.
Monitor → Active Governance	Triggered when a Monitor and Mitigate risk shows sustained threshold breaches across two review cycles.
Phase Expansion	Requires completion of all gate conditions for the current phase and formal documentation of evidence before the governance owner approves expansion. Gate conditions are not negotiable.
Scope Expansion	Any proposed expansion of the tool's authorized use cases — beyond the Deployment Profile in Step 1 — triggers a full restart of Steps 2 through 4 before approval.
Tool Suspension (Post-Launch)	Required if a Must Control risk materializes and controls cannot be made adequate. Determination made jointly by Managing Partner, Tax Practice Director, and General Counsel.