TAXLAB: AN AI TAX-FILING BENCHMARK BUILT ON FIELD-LEVEL OPERATIONALIZATION OF TAX LAW

Anonymous authors
Paper under double-blind review

ABSTRACT

AI agents are increasingly capable of completing complex, multi-step tasks, making high-stakes applications, such as automated tax preparation, a promising direction. Evaluating tax-filing agents, however, remains difficult: it is hard to determine what range of tax situations a benchmark actually covers, how diverse those situations are, and where an agent's mistakes originate when a completed return is incorrect. We introduce TAXLAB, a benchmark for interactive U.S. federal individual tax filing built on *field-level operationalization of tax law*. We construct TAXLAB-RULE, an executable, source-grounded rule system comprising 3,142 rules, each linking a specific reporting field to the official sources that determine its treatment. This representation provides a common foundation for constructing gold returns, measuring benchmark coverage and diversity, and tracing filing errors back to the rules and sources involved. Using this system, we construct 55 CPA-reviewed tax-year-2025 scenarios and an interactive environment in which agents must identify missing taxpayer information, query a simulated taxpayer, consult offline IRS guidance, and complete the required tax forms. We evaluate nine models and find that tax preparation remains challenging: the strongest fully evaluated model, Claude Opus 5.5, achieves only 64.0% perfect filing accuracy despite substantially higher field-level accuracy. Our error diagnosis further reveals distinct issue-level failure patterns across models, showing that even strong models remain disproportionately error-prone on specific tax treatments. Together, these results show that substantial progress is still needed before AI agents can reliably prepare complete tax returns, and that improving aggregate accuracy alone is insufficient without addressing the specific tax treatments that drive filing failures.

1 INTRODUCTION

AI agents are beginning to complete tasks across very different settings. Coding agents can inspect repositories, modify code, and resolve reported issues (Jimenez et al., 2024; Deng et al., 2025). Computer-using agents navigate websites and desktop applications to carry out tasks (OpenAI, 2025). These successes show that agents can do more than produce an isolated answer: they can gather information, take actions, and work toward a finished result. Tax preparation calls for the same abilities. An agent must collect taxpayer records, determine how rules apply to a person's circumstances, and carry those decisions through a set of connected forms.

The potential benefit is substantial. The Internal Revenue Service (IRS) estimates that preparing a tax-year-2025 Form 1040 or 1040-SR return takes 12 hours and costs $290 out of pocket on average (IRS, 2025). Interest in AI tax assistance is also growing. In a survey of 1,010 full-time U.S. workers, Adobe reported that use or planned use of AI for tax filing rose from 11% for 2024 returns to 26% for 2025 returns (Adobe Acrobat, 2026). Yet the taxpayers who seek help may lack the expertise to check whether an agent applied a credit correctly or overlooked an important fact. A completed return can look plausible even when an early mistake has affected several fields. This creates a pressing evaluation problem: how reliably can filing agents prepare correct returns across diverse taxpayer circumstances, and where do their errors arise?

Key Challenges. It is therefore essential to develop a tax benchmark that faithfully measures filing agents' capabilities and provides meaningful diagnoses of their failures. Doing so, however, presents three key challenges. ❶ *Benchmark coverage and diversity must be measurable.* Tax treatment can

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

change substantially with small differences in a taxpayer's circumstances, while seemingly different scenarios may exercise the same underlying rules. Thus, it is difficult to tell from the scenarios themselves what tax treatments a benchmark actually exercises and how varied those treatments are. A useful benchmark should therefore make both tax-treatment coverage and diversity explicitly measurable. Such measurement can reveal evaluation gaps that are otherwise difficult to observe. For example, under our analysis, TaxCalcBench (Bock et al., 2025), an expert-curated benchmark, covers only 57.1% of the tax treatments. ❷ *Ground-truth returns must be accurate.* The complexity of tax law also makes manually constructed reference returns prone to subtle errors, Such errors can be mistakenly attributed to the evaluated agent. Indeed, TaxCalcBench (Bock et al., 2025) has identified more than 10 errors after the initial public release. A reliable benchmark therefore needs a systematic way to construct and validate its ground truth. ❸ *Benchmark failures must be diagnosable.* To expose an agent's limitations and guide future improvements, a benchmark should go beyond reporting whether a return is correct and identify which applicable rules the agent missed or misapplied, as well as how those errors propagated through the return. However, none of the existing tax-filing benchmarks, as discussed in §2, jointly addresses these three challenges.

Our Approach. Our key insight is to construct the benchmark by *operationalizing tax law into concrete, field-level rules*, so that every benchmark artifact can be traced to the official sources supporting it. We realize this idea by decomposing tax preparation into small, connected rules, each mapped to a specific field in Form 1040 or its associated forms and schedules. More specifically, each rule specifies the taxpayer facts and prior form fields it depends on, the condition under which it applies, the value it produces for an output field, and the official source supporting that treatment, thus operationalizing tax law at the level of individual tax form fields. Rule outputs can feed subsequent rules, forming a dependency path from taxpayer facts to the completed return. Executing these rules for a taxpayer scenario therefore produces both *the gold return* and *a trace of the rules exercised*. That design gives us a common basis for addressing all three challenges. ❶ It shows which rules each scenario tests, enabling explicit measurement of benchmark coverage and diversity. ❷ It generates gold returns from a shared, validated rule system, so each validated rule provides a consistent tax treatment across all scenarios that exercise it. ❸ It further links observable mismatches in individual field to the official sources, providing a concrete basis to investigate failures.

Working with tax professionals and certified public accountants (CPA), we introduce TAXLAB, comprising 55 taxpayer scenarios whose gold returns are derived from our rule system for tax year 2025. During evaluation, agents receive an incomplete account of each taxpayer's circumstances and can ask for hold back facts. Agents also receive the forms they must complete and can consult a fixed, offline collection of IRS forms and instructions. Each task thus requires the agent to identify missing information, seek the facts it needs, consult official guidance, and produce a completed return.

Results. Across nine evaluated agents, tax filing remains challenging even for frontier models. The best *Final Return Accuracy*, which measures whether the final tax liability or refund is correct, reaches 81.8%, while the best *Compliance Field Accuracy*, which measures correctness on CPA-labeled key reporting fields, reaches 91.9%. However, the best *Perfect Filing Accuracy*, which requires every key reporting field in a return to be correct, is only 61.8%. This gap shows that obtaining the correct final tax outcome does not necessarily imply a fully correct return. Our rule-grounded diagnosis further reveals that even strong agents exhibit concentrated weaknesses on specific tax issues, while weaker agents fail more broadly across the tax system.

Our Contribution. This work makes four contributions. First, we develop TAXLAB-RULE, a field-level rule system that operationalizes U.S. individual tax law by linking each reporting decision to its required taxpayer facts, form dependencies, computation, and supporting official sources. Second, using TAXLAB-RULE, we construct TAXLAB, a benchmark of 55 CPA-reviewed tax-year-2025 scenarios with systematically generated gold returns and execution traces, enabling explicit measurement of tax-treatment coverage and diversity. Third, we design an interactive tax-filing environment in which agents must identify missing taxpayer information, query for needed facts, consult a fixed collection of IRS guidance, and complete the required federal tax forms. Fourth, we evaluate nine agents (*i.e.*, Claude Haiku, Claude Sonnet 5, Claude Opus 5.5, Fable 5.1, GPT-5.5, GPT-5.6 Luna, GPT-5.6 Sol, GPT-6 Astra, and Gemini 3.8 Flash) and show that tax filing remains challenging even for frontier models, with different agents exhibiting distinct patterns of failure.

Table 1: Comparison of agentic tax-filling benchmarks. **Cases** counts taxpayer scenarios. **Agentic** indicates a multistep filing environment with tools; **Interactive** indicates that the agent can ask for and receive missing taxpayer facts; and **IRS docs** indicates access to official IRS guidance. **Coverage** indicates whether the breadth of tax treatments exercised by the benchmark scenarios can be measured. **Diversity** indicates whether the variation among benchmark scenarios can be measured. **Sys. GT** indicates whether ground-truth returns are constructed through a systematic process. **Diag.** indicates whether incorrect outputs can be traced to the applicable tax laws. **Open** denotes the benchmark's public availability. ✓, △, and ✗ denote full, partial, and no demonstrated support, respectively.

	Dataset	Task environment			Measurement		Construction	Diagnosis	Availability
Benchmark	**Cases**	**Agentic**	**Interactive**	**IRS docs**	**Coverage**	**Diversity**	**Sys. GT**	**Diag.**	**Open**
Bock et al. (2025)	50	✗	✗	△ Web search	✗ (47.6%)	✗	✗	✗ (10 errs.)	✓
Xu & Colitsas (2026)	32	✓	△ Detection only	✓	✗	✗	✗	✗	✗
micro1 Inc. (2026)	-	✓	✗	△ Web search	✗	✗	✗	✗	✗
Entendre (2026)	-	✓	✗	✗	✗	✗	✗	✗	✗
TaxLab (ours)	55	✓	✓	✓	✓	✓	✓	✓	✓

2 Related Work

Agentic Tax-Filing Benchmarks. Due to the interdisciplinary nature of tax-filing agent evaluation, most efforts have emerged from industry than academia. Table 1 summarizes existing benchmarks.

TaxCalcBench (Bock et al., 2025) evaluates tax-return preparation on 50 open-sourced cases. Rather than providing a controlled agentic environment, it allows models to make LLM calls with server-side web search enabled. While this setup gives agents access to a broad range of tax information, its open-world nature makes evaluation less controlled and harder to reproduce. Using our rule system, we further find that TaxCalcBench's federal cases exercise only 47.6% of the tax issues represented in our system, highlighting the need for a standardized way to measure tax coverage. TaxCalcBench also illustrates the difficulty of constructing reliable ground truth: despite being developed by tax professionals, at least 10 errors in its benchmark cases have been reported (Column Tax, 2026).

1040Bench (Xu & Colitsas, 2026) comes closest to our agentic environment. It places agents in a multistep setting with taxpayer documents, tools, and a frozen collection of official IRS guidance. However, when required taxpayer information is missing, 1040Bench asks the agent to detect and report insufficient information rather than interact with the taxpayer to obtain it. It therefore evaluates missing-information detection, but not the full elicitation-to-filing loop. micro1 Inc. (2026) and Entendre (2026) also provide agentic environments. The former supports web search, whereas the latter does not provide IRS reference materials.

With the exception of TaxCalcBench, these benchmarks are not openly available. More broadly, none explicitly measures both the *coverage* of tax treatments exercised by its scenarios and the *diversity* among those scenarios, nor provide rule-level provenance for diagnosing errors. Moreover, their benchmark cases are constructed largely manually, making reliable ground-truth construction difficult and susceptible to subtle errors. In contrast, TaxLab is constructed using an executable, source-grounded rule system that enables measurable coverage and diversity, more reliable ground-truth construction, and rule-level diagnosis of agent failures. TaxLab also provides an interactive, agentic filing environment that reflects the practice of real-world tax preparation.

Tax Question-Answering Benchmarks for AI agents. A complementary line of work evaluates tax reasoning through question answering rather than return preparation. SARA (Holzenberger et al., 2020) studies statutory entailment and numerical reasoning over provisions of the U.S. Internal Revenue Code, while TaxBench (The Intelligence Consortium, 2025) uses expert-authored questions. More recent benchmarks extend this setting with large collections of tax-law questions (Kelly, 2025; Vals AI, 2026). These benchmarks are valuable for measuring tax knowledge and reasoning, but answering isolated tax questions does not establish that an agent can prepare a tax return end to end, which requires, for example, identifying missing taxpayer facts and maintaining consistency across interdependent fields. TaxLab instead targets this agentic tax-preparation setting.

Formalizing Tax Law as Executable Rules. A long line of work has explored representing tax law in machine-executable form. Early work such as TAXMAN (McCarty, 1976) modeled U.S. tax

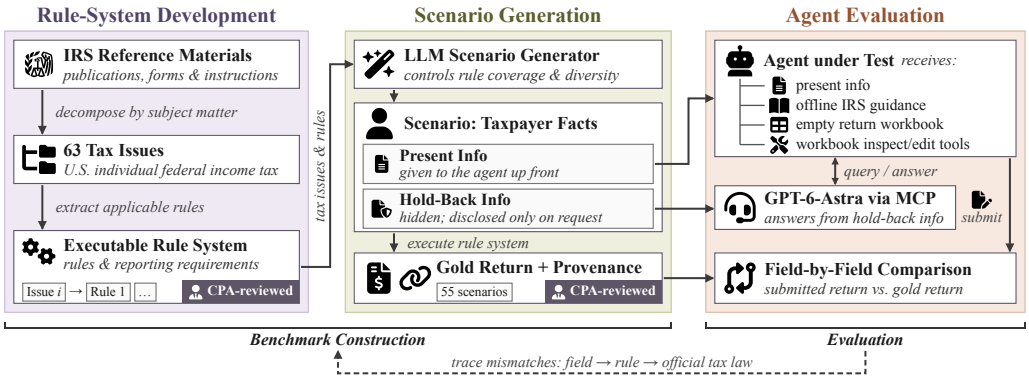

Figure 1: TAXLAB's end-to-end pipeline for benchmark construction and evaluation.

law using symbolic legal reasoning. Subsequent work has formalized tax law using expert systems, logic programming, and domain-specific languages such as Catala (Sherman, 1989; Merigoux et al., 2021b;a; Lawsky, 2022; 2023). These efforts primarily focus on faithfully formalizing and executing tax law. However, they do not jointly expose explicit rule-level dependencies, tax form fields mappings, and source provenance as first-class metadata, making them difficult to directly use for constructing tax-filing benchmarks. In contrast, our rule system is designed specifically to support the evaluation of tax-filing agents. In our system, each rule is grounded in official tax law, targets a concrete tax form field, and explicitly declares the upstream tax form fields on which it depends.

3 TAXLAB CONSTRUCTION

Figure 1 illustrates the end-to-end pipeline for benchmark construction and evaluation. The pipeline consists of three stages: *rule-system development*, *scenario generation*, and *agent evaluation*.

First, we construct an executable tax rule system from Inland Revenue Code (Office of the Law Revision Counsel, U.S. House of Representatives) and official IRS publications (IRS, 2026). Following the subject-matter-based notion of tax issues used by National Taxpayer Advocate (2022), we define a tax issue as a coherent tax matter governed by a related set of rules and reporting requirements. In total, we identify 63 tax issues covering U.S. individual federal income tax. For each issue, we extract the applicable tax rules. A CPA then reviews the resulting rules for accuracy.

Second, we use the tax issues and their rules to guide scenario generation. An LLM generates taxpayer facts that exercise different rules, allowing scenario coverage and diversity to be controlled explicitly. Executing the rule system over each generated scenario produces its gold return as a completed tax-return workbook, together with the rule-level provenance underlying each reported field. Each generated return is independently reviewed by a CPA before being included in the benchmark.

Finally, for agent evaluation, we partition each scenario's facts into *present information* (provided to the agent initially) and *hold-back information* (hidden). The agent receives the present information, a fixed offline collection of IRS guidance (IRS, 2026), an empty tax-return workbook containing the relevant federal forms and schedules, and tools for inspecting and editing the workbook. When additional taxpayer information is needed, the agent can query an MCP-based interface, backed by a strong frontier model (*i.e.*, GPT-6-Astra), which answers using the hold-back information. The agent ultimately submits a completed return, which we compare field by field against the gold return. Because each gold field is generated by the rule system, observed mismatches can be traced to the rules that produced the corresponding fields and, ultimately, to the official tax law supporting those rules.

3.1 RULE-SYSTEM DEVELOPMENT

As discussed in §2, prior work (McCarty, 1976; Sherman, 1989; Merigoux et al., 2021b;a; Lawsky, 2022; 2023) has developed deterministic systems for tax computation and, more broadly, legal reasoning. Such systems can produce tax determinations, but their outputs alone do not provide the trace needed for our benchmark, *i.e.*, *from a reported field, through the rule that determined it, to*

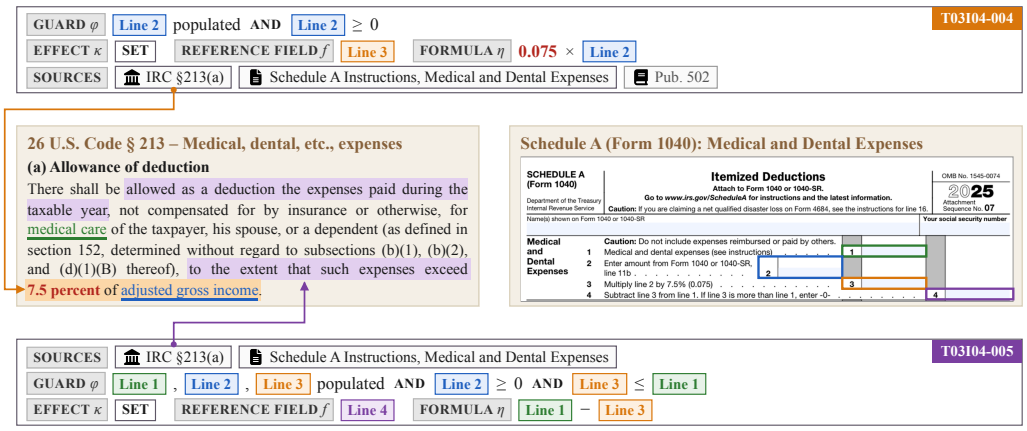

Figure 2: Example of a rule in TAXLAB-RULE. The figure shows two source-grounded rules for Schedule A medical and dental expenses. Some implementation-level fields are omitted for clarity.

the supporting tax law. Without this trace, it is difficult to identify which tax treatments a scenario exercises or to diagnose the cause of an incorrect field.

We therefore develop TAXLAB-RULE for benchmark construction. Its design follows two principles: ❶ each rule must connect a tax form field to a specific official source, and ❷ its logic must be readable enough for a non-technical but tax-profession CPA to review. Note that we do not claim TAXLAB-RULE is a more capable tax engine than prior systems; its design is tailored to coverage measurement and failure diagnosis. TAXLAB-RULE has two operational layers.

- **Tax issues.** A tax issue represents a coherent tax matter, such as medical expense deductions, principal-residence sales, or qualified business income. Following the subject-matter organization of National Taxpayer Advocate (2022), we use tax issues as a consistent unit for measuring the breadth of benchmark coverage. To construct this layer, we provide GPT-5.6-Sol with maximal reasoning effort with the IRS corpus (IRS, 2026) and the issue taxonomy of National Taxpayer Advocate (2022), and ask it to identify and group the tax issues covered by the corpus. This process yields 63 tax issues, each of which is subsequently reviewed by CPAs and independently checked by GPT-6-Astra with maximal reasoning effort.
- **Tax rules.** Each tax issue is further decomposed into executable tax rules. Each rule declares the taxpayer facts and previously computed fields it depends on, the condition under which it applies, its effect on a single target field, and the official source supporting that treatment. Evaluating rules along their field dependencies produces the completed return while recording the path from taxpayer facts to reported fields and their supporting sources. In total, we derive 3,142 rules.

Rule Representation in TAXLAB-RULE. Executable tax logic in TAXLAB-RULE resides in individual, source-grounded rules. Each rule captures one tax treatment, directly establishes one reporting target, and is represented as

$$r = \langle A, f, D, \phi, \kappa, \eta, \rho, \pi \rangle. \tag{1}$$

The upper portion of Figure 2 shows one concrete rule, `T03I04-004`, which we use to illustrate each component. More specifically, A identifies the official source span supporting the rule; in the example, the rule is grounded in IRC §213(a), the Schedule A instructions, and Publication 502. Field f identifies the target tax form field directly established by the rule, which is Schedule A line 3 in this case. D declares every input read by the rule, including raw taxpayer `FACT`s and already established `FIELD`s in the tax return; here, $D = \{\text{line 2}\}$. The guard ϕ specifies when the rule applies; in the example, line 2 must be populated and nonnegative. The formula η specifies what the rule computes, namely $0.075 \times$ line 2. The effect κ then specifies how that result modifies the target; `SET` in this example establishes the computed value as line 3. Together, D, ϕ, and η make the computation locally inspectable. Finally, ρ records the rule's descriptive role, typically as a natural-language description of its purpose to facilitate review by tax professionals, while π records its priority information. Both fields are omitted from the figure for clarity.

Note that κ is drawn from a small, registered vocabulary of effects. `SET` establishes a value; `ADD` and `SUBTRACT` compose numeric components; `CAP` and `FLOOR` impose upper and lower bounds; and `NOT_APPLICABLE` suppresses a target that is legally inapplicable. Thus, the formula η describes *what* the rule computes, while the effect κ describes *how* that result participates in the target field. Given resolved dependencies, a rule whose guard is true evaluates η and applies κ to f.

Additionally, priority is needed because tax provisions frequently contain exceptions, limitations, and special rules. Rather than imposing a total ordering over all rules, TAXLAB-RULE represents priority as a source-grounded partial order within each tax issue: $r_a \succ_I r_b$, meaning that r_a controls r_b when their applicable effects conflict. Priority is considered only for conflicting rules that target the same field. It is therefore distinct from an ordinary field dependency $g \to f$, in which a rule producing f simply consumes the resolved value of an upstream field g.

Example. Figure 2 illustrates this representation using the Schedule A medical-expense deduction. Rule `T03I04-004` implements the 7.5% adjusted-gross-income threshold: when Schedule A line 2 is populated and nonnegative, it sets line 3 to $0.075 \times$ line 2. Rule `T03I04-005` then consumes lines 1-3 and, on the illustrated branch where the threshold does not exceed the medical expenses, sets line 4 to line 1 $-$ line 3. For clarity, the visualization omits several implementation-level fields (*e.g.*, the rule's descriptive roles and priority). gold returns in the following sections. □

Compositional Execution of TAXLAB-RULE. Given a set of taxpayer facts, TAXLAB-RULE derives the return incrementally. At each step, a rule becomes *runnable* once all of its declared dependencies D have been resolved. The engine then evaluates its guard ϕ; if the guard is satisfied, it computes η and applies effect κ to the target field f. The newly established field may in turn resolve the dependencies of other rules, so this process continues until no additional rule can produce a new result. Rules therefore compose through explicit field dependencies, allowing a completed return to emerge from a sequence of small, independently source-grounded computations.

Example. Figure 3 illustrates this process. The taxpayer facts initially have \$9,800 of medical and dental expenses on line 1 and \$80,000 of adjusted gross income on line 2. With line 2 available, rule `T03I04-004` computes line 3 as $0.075 \times 80{,}000 = 6{,}000$. This newly established value then enables rule `T03I04-005`, which computes line 4 as $9{,}800 - 6{,}000 = 3{,}800$. Thus, each rule contributes one local step, while their dependencies determine how those steps compose into the final return. □

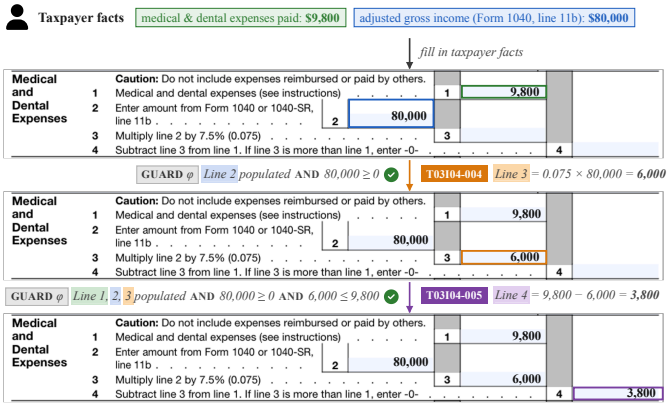

Figure 3: Example of compositional execution in TAXLAB-RULE.

Rule Generation. Manually translating tax guidance into thousands of fine-grained rules is labor-intensive and difficult to scale consistently. We therefore use Codex with GPT-5.6-Sol at maximum reasoning effort to generate an initial set of rules for each tax issue, following our predefined rule representation and the corresponding official sources. Importantly, we treat LLM generation only as an authoring aid rather than as ground truth. A CPA manually validates every generated rule against its cited authority and corresponding tax form fields before it is admitted into the rule system. Through this process, we derive **3,142 rules** covering **63 tax issues**. CPA review identified only 34 errors in the initially generated rules (1.08%), most of which were mechanical rather than substantive tax errors, *e.g.*, a workbook field locator being shifted by a row or column. These errors are corrected before the rules are used for benchmark construction.

3.2 SCENARIO GENERATION

We construct tax year 2025 benchmark cases through a CPA-guided, coverage-driven process. Our CPA first manually designs 10 seed cases that they consider representative and reasonably challenging

tax-preparation scenarios. We then provide these seed cases together with the complete rule system to Codex (with GPT-5.6-Sol at maximum reasoning effort) and ask it to generate 55 cases of comparable difficulty. We additionally instruct Codex to generate cases that exercise a broader range of rules and tax treatments. Importantly, Codex generates only the underlying taxpayer facts. We execute the rule system on each case to derive its gold return and rule trace. Finally, the CPA reviews all 55 cases to confirm that the facts are consistent, the returns are accurate and reflect reasonable tax-preparation choices (*e.g.*, standard versus itemized deduction), and the difficulty is appropriate for the benchmark. During this validation, the CPA identifies no reporting errors in the gold returns.

Measurable Coverage and Diversity. We measure benchmark coverage and diversity based on the rules exercised by each case. Figure 4 visualizes issue-level coverage across all 55 benchmark cases. Each column represents a case, each row represents a tax issue, and a colored cell indicates that the corresponding issue is exercised by that case. For readability, we group the 63 issues into six broader tax topics. The bars on the right show the number of cases exercising each issue. All 63 tax issues are covered, while the distinct activation patterns across columns demonstrate that the cases exercise diverse combinations of tax issues. The three most frequently exercised issues are *Tax Computation* (55 cases), *Withholding & Refunds* (49), and *Standard Deduction* (30). This distribution is expected because these issues arise in many individual tax returns, whereas more specialized tax treatments apply only to narrower subsets of taxpayers.

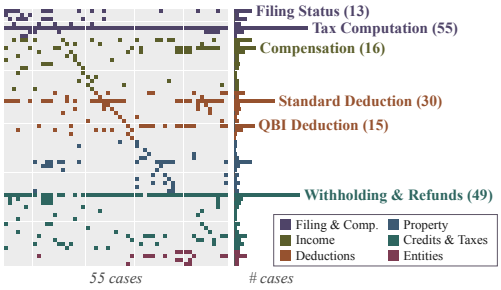

Figure 4: Tax-issue coverage of TAXLAB.

We further quantify benchmark diversity using the Jaccard distance (Jaccard, 1901)[1]. Specifically, we represent each case by the set of issues or rules it exercises and compute the Jaccard distance between every pair of cases. Figure 5 presents the distributions across all 1,485 pairs of our 55 cases. At the issue level, the mean and median distances are 0.77 and 0.80, respectively; at the rule level, they are 0.74 and 0.81. In other words, a typical pair of cases overlaps on only about one fifth of the issues or rules in their combined execution traces.

We additionally compare against all federal FY25 cases in TaxCalcBench (Bock et al., 2025). Because TaxCalcBench does not provide rule traces, we manually reconstruct the underlying taxpayer facts and execute them through TAXLAB-RULE to obtain rule-level traces. We exclude state-only cases because TAXLAB-RULE currently models federal taxation only. As shown in Figure 5, TaxCalcBench exhibits substantially lower issue-level diversity. This comparison illustrates why benchmark diversity should be measured rather than assumed. That is, even expert-curated cases can concentrate on overlapping tax treatments when diversity is not explicitly quantified during benchmark construction.

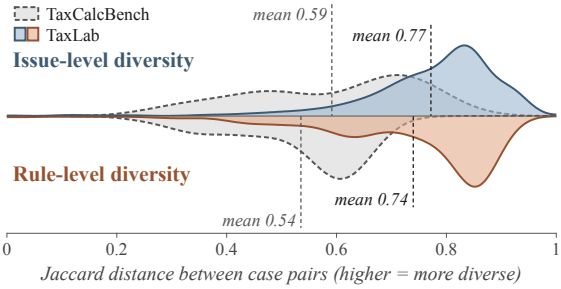

Figure 5: Diversity of TAXLAB and TaxCalcBench.

3.3 AGENT EVALUATION

We evaluate each agent in an interactive tax-filing environment. For each case, the CPA first partitions the taxpayer facts into *present information* and *hold-back information*. Facts that taxpayers would be unlikely to volunteer without being specifically asked are placed in the hold-back information. The agent receives the present information, an empty tax-return workbook, tools for inspecting and editing it, and a fixed offline collection of IRS publications (IRS, 2026).

During evaluation, the hold-back information is accessible through an MCP-based taxpayer interface. The interface is intentionally conservative, *i.e.*, it answers only the agent's specific question, provides

[1]For two sets A and B, the Jaccard distance is $1 - |A \cap B|/|A \cup B|$. Higher values indicate greater diversity.

Table 2: **Tax-filing agent evaluation results.** Models are ranked by PFA, with ties broken by FRA. *Graded* denotes the number of completed and fully graded cases out of 55. All metrics are computed over these cases. Efficiency bars are normalized within each column.

				Accuracy (%) ↑			Efficiency (mean per graded case)			
#	Model	Harness	Graded	FRA	CFA	PFA	MCP calls	Input (k)	Output (k)	Time (min)
1	Claude Opus 5.5	*Claude Code*	50/55	68.0	86.1	**64.0**	9.7	7,826.2	85.6	20.0
2	Fable 5.1	*Claude Code*	55/55	**81.8**	**91.9**	61.8	10.8	2,413.4	65.5	18.2
3	GPT-6 Astra	*Codex*	55/55	69.1	86.5	61.8	13.7	2,759.2	20.8	13.1
4	GPT-5.6 Sol	*Codex*	55/55	61.8	75.3	30.9	10.4	5,653.3	42.3	23.0
5	GPT-5.5	*Codex*	53/55	49.1	65.7	17.0	10.3	4,437.9	33.5	18.7
6	Gemini 3.8 Flash	*Antigravity*	8/55	62.5	70.7	12.5	8.0	13,659.1	82.1	21.4
7	Claude Sonnet 5	*Claude Code*	24/55	45.8	67.0	8.3	7.8	4,530.8	44.3	18.4
8	Claude Haiku	*Claude Code*	53/55	3.8	12.2	0.0	7.6	4,443.3	44.8	22.2
9	GPT-5.6 Luna	*Codex*	53/55	0.0	9.9	0.0	8.4	2,276.5	15.3	12.0

one answer at a time, and does not volunteer additional facts. To ensure that the oracle does not itself introduce evaluation errors, we use GPT-6-Astra with max reasoning effort as the backbone LLM for generating responses. We note that this design is important because it evaluates whether the agent can recognize missing information and ask for it, rather than hallucinating missing facts.

4 Evaluation

Models and Harness. We evaluate nine models: Claude Haiku, Claude Sonnet 5, Claude Opus 5.5, Fable 5.1, GPT-5.5, GPT-5.6 Luna, GPT-5.6 Sol, GPT-6 Astra, and Gemini 3.8 Flash. All models use the highest available reasoning effort. We run each model with its corresponding agent harness: Codex for OpenAI models, Claude Code for Anthropic models, and the Antigravity CLI for Gemini. All models operate in the same task environment, with access to Python 3.13, `openpyxl` for programmatic spreadsheet manipulation, and a simulated-taxpayer MCP oracle for querying missing taxpayer information. The oracle is powered by GPT-6-Astra with reasoning effort as max. We provide the IRS guidance corpus locally as HTML files, allowing agents to inspect official guidance without Internet access. IRS forms and schedules are represented as Excel worksheets rather than PDFs. This preserves the information and instructions on the original forms while enabling reliable programmatic access and modification. Each case has a 30-minute wall-clock timeout.

Evaluation Metrics. We report three primary accuracy metrics. *Final Return Accuracy* (**FRA**) measures the fraction of cases for which the final Form 1040 outcome, *i.e.*, the tax liability or refund, matches the gold return. *Compliance Field Accuracy* (**CFA**) measures accuracy over CPA-labeled *anchor fields* that capture the key reporting decisions and amounts in each case. We intentionally do not score every populated field because tax forms contain many dependent and transferred fields. A single upstream error can therefore propagate to multiple downstream fields, and counting each mismatch separately would overcount the same underlying error. *Perfect Filing Accuracy* (**PFA**) measures the fraction of cases for which all anchor fields are correct, providing a stricter case-level measure of complete filing correctness. In addition to these accuracy metrics, we report performance measures, including the number of taxpayer-oracle queries and the wall-clock time of each run.

Overall Results. Table 2 summarizes the overall results. A small number of planned runs are excluded because they do not produce fully graded submissions. Specifically, 56 runs exceed the 30-minute execution limit. These include Claude Sonnet 5 (26), Gemini 3.8 Flash (21), Claude Opus 5.5 (5), GPT-5.5 (2), Claude Haiku (1), and GPT-5.6 Luna (1). An additional 33 runs fail because of runtime errors, such as provider rate limits, primarily from Gemini 3.8 Flash (26) and Claude Sonnet 5 (5), with one each from Claude Haiku and GPT-5.6 Luna. Because Claude Sonnet 5 and Gemini 3.8 Flash have only 24 and 8 graded cases, respectively, their accuracy metrics reflect only the completed subset. GPT-6 Astra, Fable 5.1, and GPT-5.6 Sol complete all 55 cases.

Overall, tax filing remains challenging even for the strongest models. Fable 5.1 achieves the highest FRA (81.8%) and CFA (91.9%), while Claude Opus 5.5 achieves the highest PFA (64.0%) over its 50 graded cases. Among models completing all 55 cases, Fable 5.1 and GPT-6 Astra tie for the highest PFA at 61.8%. GPT-6 Astra achieves this accuracy with a lower average runtime than Fable 5.1

(13.1 vs. 18.2 minutes) and substantially fewer output tokens (20.8K vs. 65.5K). Performance drops considerably for the remaining models: GPT-5.6 Sol reaches 30.9% PFA, GPT-5.5 reaches 17.0%, while Claude Haiku and GPT-5.6 Luna do not produce a perfect filing on any graded case.

Error Diagnosis. Beyond measuring whether an AI agent produces the correct return, our field-level operationalization of tax law allows us to diagnose *where* its errors surface. Because each scored anchor field is linked to the tax issue and rules that produce it, we compute an issue-level error rate from the graded cases. Figure 6 normalizes these rates by each model's average across exposed issues, so $1\times$ represents the model's typical issue and values above $1\times$ indicate disproportionate weakness independent of overall accuracy. The resulting profiles reveal different failure modes: stronger models tend to concentrate their remaining errors in a small number of issues. For example, Claude Opus 5.5 reaches $5.2\times$ on *Medical and Dental Deduction*, while GPT-6 Astra reaches $3.2\times$ on *State and Local Tax Deduction*. On the other hand, weaker models such as Claude Haiku and GPT-5.6 Luna fail more broadly across issues.

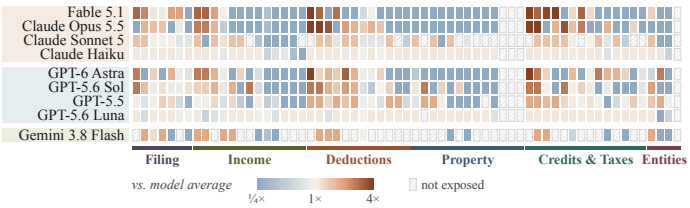

Figure 6: Issue-level error diagnosis across models.

Effect of IRS Documents. Figure 7 examines how access to the IRS document collection affects GPT-6 Astra. Providing IRS documents increases FRA from 61.8% to 69.1% and PFA from 54.5% to 61.8%, both gains of 7.3 percentage points, while CFA increases more modestly from 85.4% to 86.5%. The larger case-level gains suggest that official guidance helps correct several consequential tax decisions that determine whether an entire return is correct. This improvement comes primarily at the cost of additional input context, with input-token usage increasing by 46.1%. The number of MCP calls remains unchanged, which is expected because the amount of held-out taxpayer information is the same in both settings. Interestingly, output-token usage and wall-clock time both decrease slightly despite the substantial increase in input tokens. We hypothesize that direct access to authoritative IRS guidance reduces some of the reasoning and information-retrieval effort otherwise required to determine the applicable tax treatment. Overall, for a strong model such as GPT-6 Astra, access to IRS documents improves filing accuracy without requiring longer execution time.

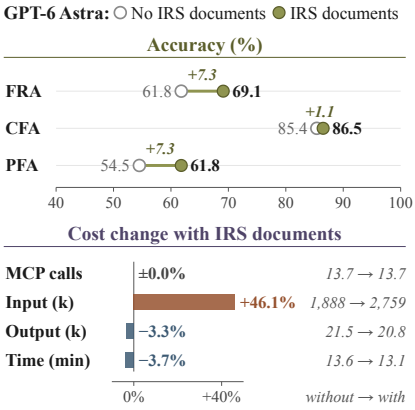

Figure 7: Effect of IRS documents.

5 Discussion and Conclusion

Limitation. TaxLab currently focuses on U.S. federal individual income tax and does not cover state, local, or international tax issues. Extending the benchmark to these broader tax domains would increase the range of filing situations and is an important direction for future work. In addition, although we make substantial efforts to ensure the correctness of our rule system, including CPA review and cross-checking with frontier models, undetected errors may still remain. We therefore treat rule-system correctness as an internal threat to validity and mitigate this risk through multiple layers of validation, while continuing to refine and audit the system as the benchmark evolves.

Conclusion. We present TaxLab, a benchmark for interactive U.S. federal individual tax filing built on field-level operationalization of tax law. At its core is TaxLab-Rule, a source-grounded rule system that maps taxpayer facts and official tax guidance to individual reporting fields, enabling benchmark coverage and diversity to be measured and filing errors to be traced to the tax rules and sources involved. Our evaluation of nine models shows that even strong agents remain far from reliably producing complete and correct tax returns, with distinct weaknesses across tax issues. More broadly, TaxLab demonstrates how operationalizing domain rules at the level of concrete outputs can support both rigorous evaluation and informative diagnosis of agents in high-stakes domains.

AI USE STATEMENT.

Generative AI tools were used as part of the research methodology and for limited writing assistance. During benchmark construction, we used Codex with GPT-5.6 Sol at its maximum reasoning effort to assist in translating official tax guidance into our predefined rule representation and to generate candidate taxpayer scenarios from CPA-designed seed cases. All generated rules were checked against their cited official sources, and all benchmark cases and resulting gold returns were reviewed by the authors and a CPA. We also use language models as components of the evaluation environment, including GPT-6 Astra to simulate a taxpayer answering agents' questions about withheld case facts. Generative AI was additionally used to help edit and improve the presentation of the manuscript. The authors reviewed and take responsibility for all manuscript content, experimental design, results, and conclusions.

ETHICS STATEMENT.

Tax preparation is a high-stakes application in which incorrect advice or filing decisions can have financial and legal consequences. Accordingly, TAXLAB is intended as a research benchmark for measuring and diagnosing tax-preparation agents, not as a tax-filing service or a substitute for professional tax advice. The benchmark uses synthetic taxpayer scenarios rather than real taxpayer records and does not require personally identifying taxpayer information. We ground benchmark cases and gold returns in a fixed set of official tax sources and expose field-level errors to make model failures easier to inspect. Nevertheless, success on the benchmark should not be interpreted as evidence that a model can safely prepare real tax returns without independent review.

REPRODUCIBILITY STATEMENT.

We design TAXLAB so that benchmark construction and evaluation can be reproduced from explicit artifacts rather than undocumented expert judgments. The benchmark and agentic evaluation environment will be released as open source once final data cleanup is complete. The release will include the benchmark cases, gold returns, tax-return templates, evaluation prompts, and scripts for computing the reported metrics and analyses. We will also provide the fixed collection of IRS materials used by the agents, together with the model, agent-harness, reasoning-effort, and execution-limit configurations used in our experiments. Because some evaluated models are accessed through proprietary APIs and may change over time, exact model behavior may not remain reproducible. However, the benchmark inputs, evaluation environment, grading procedure, and analysis pipeline will remain fixed and publicly available.

REFERENCES

Adobe Acrobat. The new tax assistant: Why AI adoption has surged for filing taxes in 2026, 2026. URL https://www.adobe.com/acrobat/resources/importance-of-document-organization-during-tax-season.html. Self-reported survey of 1,010 full-time U.S. employees; accessed September 23, 2026.

Michael R Bock, Kara Molisee, Zachary Ozer, and Sumit Shah. Taxcalcbench: Evaluating frontier models on the tax calculation task. *arXiv preprint arXiv:2507.16126*, 2025.

Column Tax. Tax calc bench issues. GitHub, 2026. URL https://github.com/column-tax/tax-calc-bench/issues?q=is%3Aissue.

Xiang Deng, Jeff Da, Edwin Pan, Yannis Yiming He, Charles Ide, Kanak Garg, Niklas Lauffer, Andrew Park, Nitin Pasari, Chetan Rane, et al. Swe-bench pro: Can ai agents solve long-horizon software engineering tasks? *arXiv preprint arXiv:2509.16941*, 2025.

Entendre. TaxBench: U.S. federal tax return LLM benchmark, 2026. URL https://entendre.ai/research/taxbench/.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. A dataset for statutory reasoning in tax law entailment and question answering. *arXiv preprint arXiv:2005.05257*, 2020.

IRS. 2025 instructions for form 1040 and form 1040-sr. Section "Estimates of Taxpayer Burden", 2025. URL https://www.irs.gov/instructions/i1040gi. Accessed September 23, 2026.

IRS. Irs publications. https://www.irs.gov/publications/, 2026. Accessed: 2026-09-25.

Paul Jaccard. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bull Soc Vaudoise Sci Nat*, 37:241–272, 1901.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, volume 2024, pp. 54107–54157, 2024.

Trent M. Kelly. US Tax Code Benchmark. Hugging Face, 2025. URL https://huggingface.co/datasets/trentmkelly/USTaxCodeBench. Multiple-choice benchmark generated from the U.S. Internal Revenue Code.

Sarah B Lawsky. Coding the code: Catala and computationally accessible tax law. *SMU L. Rev.*, 75: 535, 2022.

Sarah B Lawsky. Reasoning with formalized statutes: The case of capital gains and losses. *Va. Tax Rev.*, 43:361, 2023.

L Thorne McCarty. Reflections on taxman: An experiment in artificial intelligence and legal reasoning. *Harv. L. Rev.*, 90:837, 1976.

Denis Merigoux, Nicolas Chataing, and Jonathan Protzenko. Catala: a programming language for the law. *Proceedings of the ACM on Programming Languages*, 5(ICFP):1–29, 2021a.

Denis Merigoux, Raphaël Monat, and Jonathan Protzenko. A modern compiler for the french tax code. In *Proceedings of the 30th ACM SIGPLAN International Conference on Compiler Construction*, pp. 71–82, 2021b.

micro1 Inc. Realm: Tax reasoning benchmark, 2026. URL https://www.micro1.ai/benchmark/realm-tax.

National Taxpayer Advocate. 2022 annual report to congress, 2022. URL https://www.irs.gov/pub/irs-prior/p2104--2022.pdf.

Office of the Law Revision Counsel, U.S. House of Representatives. United states code. https://uscode.house.gov/. Accessed: 2026-09-25.

OpenAI. Computer-using agent. https://openai.com/index/computer-using-agent/, January 2025. Published January 23, 2025.

David M Sherman. Expert systems and icai in tax law: Killing two birds with one ai stone. In *Proceedings of the 2nd international conference on Artificial intelligence and law*, pp. 74–80, 1989.

The Intelligence Consortium. Taxbench. Tax AI Consortium, 2025. URL https://tax.intelligenceconsortium.org/. Benchmark for evaluating large language models on tax-related tasks.

Vals AI. Tax agent bench. Vals AI, 2026. URL https://www.vals.ai/benchmarks/tax_agent_bench. Agentic benchmark for research-grade U.S. corporate tax questions.

David Xu and Tommy Colitsas. 1040Bench: Evaluating frontier models on form 1040 preparation, August 2026. URL https://verona.so/blog/1040bench. Published August 21, 2026; accessed September 23, 2026.

A BACKGROUND OF TAX FILLING

U.S. Tax System. The U.S. individual income tax system operates at both the federal and state levels. Individual taxpayers must comply with federal income tax laws administered by the Internal Revenue Service (IRS), while most states impose separate income tax systems with their own statutes, regulations, deductions, credits, forms, and filing requirements. Although state tax rules often build upon concepts used in the federal tax system, they frequently diverge in important ways, requiring taxpayers to navigate multiple overlapping but distinct legal regimes. At the federal level, the individual income tax system is governed primarily by the Internal Revenue Code (Title 26 of the U.S. Code) and the accompanying Treasury Regulations. For individual taxpayers, the tax code is organized around a series of interconnected provisions that define gross income, allowable exclusions, deductions, credits, filing status, and tax rates. The code also contains numerous special provisions, exceptions, phase-outs, and reporting requirements that depend on taxpayers' specific circumstances, contributing significantly to the complexity of individual tax compliance. The federal individual income tax system is centered on Form 1040 but relies on a large ecosystem of more than 110 supplementary schedules and forms. Depending on a taxpayer's income sources, deductions, credits, and filing circumstances, a return may require multiple numbered schedules, lettered schedules, and specialized forms, creating a complex filing framework that requires taxpayers and tax professionals to identify and apply the relevant provisions of the tax code.

TAXLAB restrict our analysis to U.S. federal individual income tax compliance. The evaluation includes tax issues arising from pass-through entities, such as partnerships and S corporations, to the extent that they affect an individual's federal tax return. We exclude state and local taxes, which vary across jurisdictions, as well as international tax issues, whose treatment often depends on country-specific tax treaties and foreign legal rules. To systematically evaluate AI performance in tax compliance, we operationalize the federal individual income tax code into six major themes comprising 63 issue topics that tie directly to the tax filing forms. The issue topics were designed to provide comprehensive coverage of the federal individual income tax system, ensuring that all major areas relevant to Form 1040 preparation are represented in the evaluation.

Tax Filing Process. Tax compliance is costly, because the tax system is highly complex and continuously changing, requiring taxpayers to spend substantial efforts understanding rules, maintaining records, completing forms, and meeting reporting requirements. AI has emerged as an attractive solution to these compliance challenges because it can dramatically reduce the time, effort, and expense associated with navigating a complex tax system. However, given the complexity and frequent changes in tax regulations, the value of these tools depends critically on their ability to accurately understand and operationalize tax law, as inaccuracies could result in filing errors, missed tax benefits, or noncompliance. To evaluate whether AI can effectively and accurately assist with tax filing and tax compliance, it is useful to consider how professional tax preparers such as CPAs complete a tax return. The process typically involves several core activities:

1. Gathering and organizing taxpayer information, including income statements, expense records, and supporting documentation;
2. Identifying the tax rules, filing requirements, and forms applicable to the taxpayer's circumstances;
3. Determining eligibility for deductions, credits, exclusions, and other tax benefits;
4. Completing the relevant tax forms and schedules by applying the tax law to the taxpayer's specific facts; and
5. Reviewing the return with the client, resolving outstanding issues, and finalizing the documentation prior to filing.

Each of these tasks requires not only retrieving information from tax law but also interpreting legal provisions and applying them to a taxpayer's particular situation. Accordingly, evaluating AI's ability to support tax filing requires assessing how well it can perform these core activities that human tax professionals routinely undertake.

More importantly, the first step of tax preparation, gathering and validating taxpayer information, is often iterative rather than a one-time collection of facts. Under Circular 230, practitioners

may generally rely on information provided by a client, but they are required to make reasonable inquiries when that information appears incorrect, inconsistent, or incomplete. The IRS similarly imposes due-diligence requirements that may require preparers to obtain additional information before determining eligibility for certain tax benefits. As a result, professional tax preparation frequently involves multiple rounds of information gathering and clarification. Prior studies typically provide models with a complete set of facts upfront, overlooking this important aspect of real-world tax practice. In reality, tax filing is not simply a mapping from a fixed set of inputs to a final tax outcome; it is an interactive process involving fact gathering, legal issue identification, determination of applicable deductions and credits, completion of forms and schedules, and follow-up inquiries when critical information is missing. This workflow motivates our inclusion of an interactive agent in our benchmark.

Tax Return Accuracy and Compliance. Tax return accuracy encompasses more than the correctness of a single final tax liability. It refers to the correctness, completeness, and truthfulness of all information reported on a tax return and its supporting schedules, ensuring compliance with applicable tax laws. In practical tax preparation, accuracy also involves making tax choices that are appropriate for the taxpayer's circumstances. A return may comply with tax law yet still be suboptimal. For example, a taxpayer may legally claim the standard deduction even when their allowable itemized deductions are larger, resulting in a higher tax liability than necessary.

Importantly, identical final tax liabilities do not necessarily imply equally accurate returns. For example, when statutory stock options are exercised but not yet sold, the income is generally deferred until disposition, but the exercise may generate Alternative Minimum Tax (AMT) liability. Two returns could therefore produce the same current-year tax liability while differing in the accuracy of their regular tax and AMT calculations. Such errors can affect AMT credit carryforwards and lead to incorrect tax liabilities in future years when the stock is ultimately sold.

The U.S. tax system places an explicit accuracy requirement on the return as a whole. When signing Form 1040, taxpayers attest under penalties of perjury that the return, including all accompanying schedules and statements, is "true, correct, and complete." IRS examinations are similarly item-oriented, assessing the accuracy of reported income, deductions, expenses, and credits, and specific line items can be verified against third-party information returns such as Forms W-2 and 1099.

Consequently, tax compliance cannot be evaluated solely by the final tax outcome. Errors in individual fields may not immediately affect tax liability, while mistakes in intermediate calculations can propagate through multiple forms and carry over to future tax years. Moreover, a legally permissible but suboptimal choice may produce a valid return while failing to reflect the tax treatment most favorable to the taxpayer's circumstances. We therefore evaluate the correctness of individual return fields in addition to the final tax liability, including whether relevant tax choices are made consistently with the taxpayer's facts. Although this field-level metric is a benchmark design choice rather than an IRS audit metric, it better reflects the item-level structure through which tax returns are prepared, reviewed, and enforced.

To develop a comprehensive evaluation framework, we systematically operationalized the federal individual income tax system into six major topics and 63 issues. The topics were designed to reflect the logical structure of individual tax return preparation, covering (1) taxpayer status and return requirements, (2) income inclusion and exclusion rules, (3) deductions and loss limitations, (4) property transactions and cost recovery, (5) credits, payments, and additional taxes, and (6) business and pass-through entity reporting. Within each topics, we identified granular issues corresponding to major federal tax doctrines, filing requirements, and Form 1040 reporting obligations. Together, these 63 issues provide broad coverage of the federal individual income tax system and ensure that the benchmark evaluates AI performance across the full range of topics commonly encountered in individual tax compliance.

In addition to supporting coverage, the issue taxonomy enables transparent error diagnosis. For each issue topic, we develop structured rules based primarily on the relevant provisions of the Internal Revenue Code and the corresponding IRS form instructions. These rules create an explicit mapping between tax return outcomes and the legal requirements governing them. When an AI system produces an incorrect answer, we can trace the error back to the specific statutory provision or filing instruction that was misinterpreted, omitted, or incorrectly applied. This approach allows us to

move beyond measuring whether a model reaches the correct result and instead evaluate whether it accurately understands and operationalizes the underlying tax law. Such traceability is particularly important for tax compliance, where different reasoning paths may produce similar tax outcomes in the current year but differ in legal correctness and future tax consequences.

B TAX ISSUES INTRODUCTION

B.1 T01: FILING, STATUS, AND TAX COMPUTATION

This section covers return-level rules that determine whether and how an individual files a federal income-tax return, including filing obligations, filing status, dependents, timing rules, and tax computation.

T01I01 : Filing requirements, deadlines, and extensions. This issue determines whether a taxpayer must file a federal income-tax return and when it is due. Filing requirements depend on factors such as filing status, age, income, and special circumstances. Extensions generally provide more time to file, but not more time to pay. The principal rules include IRC Sections 6012, 6072, and 6081.

T01I02 : Filing status. Filing status determines the applicable tax rates, standard deduction, and many other tax rules. The main statuses are single, married filing jointly, married filing separately, head of household, and qualifying surviving spouse. The determination is primarily governed by IRC Sections 1, 2, and 7703.

T01I03 : Dependents. This issue determines whether another person qualifies as the taxpayer's dependent. The law distinguishes qualifying children from qualifying relatives using tests involving relationship, age, residency, support, income, and joint-return status. The core rules appear in IRC Section 152.

T01I04 : Tax accounting methods and income/expense timing. This issue determines when income and expenses belong on the tax return. Most individuals use the cash method, under which income is generally recognized when received and expenses when paid, subject to special timing rules. IRC Sections 446, 451, and 461 provide the main framework.

T01I05 : Individual income-tax computation. This issue converts taxable income into regular federal income tax using the applicable rate schedule. Certain income, particularly qualified dividends and long-term capital gains, may receive preferential rates. IRC Section 1 and the Form 1040 tax-computation worksheets implement these rules.

T01I06 : Alternative minimum tax (AMT) and prior-year minimum-tax credit. The AMT recomputes income after modifying certain deductions and preferences and compares the resulting tentative minimum tax with regular tax. Form 6251 performs the AMT calculation, while Form 8801 may provide a later credit for certain prior-year AMT. The principal provisions are IRC Sections 53 and 55–59.

T01I07 : Kiddie tax and reporting children's investment income. Certain children's unearned income may be taxed using the parent's marginal tax rate under the kiddie-tax rules, generally through Form 8615. In limited cases, a parent may instead report qualifying interest and dividends on Form 8814. These rules are primarily governed by IRC Section 1(g).

B.2 T02: INCOME AND EXCLUSIONS

This topic determines whether amounts received by a taxpayer are taxable, excluded from income, or assigned to another taxpayer. IRC Section 61 provides the general rule that income is taxable unless a specific provision provides otherwise.

T02I01 : Wages, fringe benefits, and equity compensation. Compensation for services, including wages, bonuses, tips, and many noncash benefits, is generally taxable under IRC Section 61. Special rules govern property transferred for services and qualifying fringe benefits, including IRC Sections 83 and 132. Employee compensation is commonly reported on Form W-2.

T02I02 : Interest, dividends, and tax-exempt income. Interest and dividends are generally taxable, although qualified dividends may receive preferential rates. Interest from qualifying state or local government obligations may be federally tax exempt under IRC Section 103 while still requiring reporting. Forms 1099-INT, 1099-DIV, and Schedule B provide the main reporting framework.

T02I03 : Retirement distributions, rollovers, and annuities. Distributions from pensions, IRAs, retirement plans, and annuities may be fully taxable, partly taxable, or excluded when they qualify as rollovers. The analysis therefore separates the gross distribution from its taxable portion. Important provisions include IRC Sections 72, 402, and 408.

T02I04 : Business, pass-through, and rental income. This issue covers income from sole proprietorships, rental property, partnerships, and S corporations. Such income commonly flows through Schedules C or E, and pass-through owners may recognize taxable income even without receiving corresponding cash distributions. Separate issues apply owner-basis and loss limitations.

T02I05 : Social Security, unemployment, and public benefits. Different government benefits receive different tax treatment. Unemployment compensation is generally taxable under IRC Section 85, while only a portion of Social Security benefits may be taxable under IRC Section 86. Other public benefits may qualify for separate statutory exclusions.

T02I06 : State-tax refunds and the tax-benefit rule. A state or local income-tax refund is generally taxable only to the extent that the earlier deduction produced a federal tax benefit. IRC Section 111 therefore requires looking back to the prior year's deduction rather than automatically including every refund in income.

T02I07 : Divorce, support payments, and family property transfers. Property transfers between spouses or incident to divorce generally receive nonrecognition treatment under IRC Section 1041. Child support is generally neither deductible by the payer nor taxable to the recipient. Alimony treatment depends in part on when the governing divorce or separation instrument was executed or modified.

T02I08 : Education awards and education-account distributions. IRC Section 117 may exclude qualifying scholarships used for tuition and related expenses, while amounts used for other purposes can be taxable. Distributions from qualified education accounts may also be tax free under IRC Sections 529 and 530 when used for qualified expenses.

T02I09 : Life-insurance proceeds and death benefits. Life-insurance proceeds paid because of an insured person's death are generally excluded from gross income under IRC Section 101. Interest on delayed proceeds and certain transfers-for-value may nevertheless create taxable amounts.

T02I10 : Personal-injury payments, welfare, and other income exclusions. Certain receipts are excluded because of why they were paid. For example, IRC Section 104 generally excludes qualifying workers' compensation and certain damages received on account of physical injury or sickness, while punitive damages generally receive different treatment.

T02I11 : Below-market loans and imputed interest. IRC Section 7872 may treat low- or no-interest loans as generating interest even when little or no interest was actually paid. Different rules apply to gift loans, compensation-related loans, and corporation-shareholder loans, with limited exceptions for some transactions.

T02I12 : Canceled debt and miscellaneous income, including gambling winnings. Canceled debt is generally income under IRC Section 61(a)(11), while IRC Section 108 provides exclusions

for circumstances such as bankruptcy and insolvency. Gambling winnings are also generally taxable, while gambling losses are separately limited under IRC Section 165(d).

T02I13 : Who must report income, including income assignment and community-property allocation. This issue determines whose return should contain an item of income. Income is generally taxed to the person who earns it or owns the underlying property, while community-property rules can allocate income between spouses. IRC Section 66 and applicable state property law are particularly relevant.

B.3 T03: Deductions and Loss Limitations

This topic determines which expenses and losses reduce taxable income and how overlapping deduction and loss limitations interact.

T03I01 : Adjustments to income ("above-the-line" deductions). Adjustments to income reduce gross income in computing AGI and are generally reported through Schedule 1. Examples include certain educator expenses, the deductible portion of self-employment tax, and qualifying self-employed health-insurance expenses. IRC Section 62 provides the general framework.

T03I02 : Retirement-plan contributions and deductions. This issue determines when contributions to retirement arrangements produce a current deduction. For example, a traditional IRA deduction under IRC Section 219 can depend on compensation, filing status, employer-plan participation, and modified adjusted gross income.

T03I03 : Standard deduction and comparison with itemized deductions. Taxpayers generally use either the standard deduction or their allowable itemized deductions. IRC Section 63 governs the standard deduction and includes additional rules for dependents and taxpayers who are elderly or blind.

T03I04 : Medical and dental expense deductions. IRC Section 213 allows an itemized deduction for qualifying unreimbursed medical expenses exceeding the applicable AGI threshold. Expenses paid with tax-free HSA distributions cannot also support a Schedule A medical deduction, so the two treatments must be coordinated.

T03I05 : State and local tax deductions. IRC Section 164 permits deductions for specified state and local income or sales taxes and property taxes, subject to the federal limitation applicable for the tax year. The calculation must distinguish deductible taxes from nondeductible fees or charges.

T03I06 : Home-mortgage and investment-interest deductions. Qualified residence interest may be deductible under IRC Section 163(h), subject to debt and residence requirements. Investment interest is separately governed by IRC Section 163(d) and is generally limited by net investment income, often through Form 4952.

T03I07 : Charitable contributions and deduction limits. IRC Section 170 permits deductions for qualifying charitable contributions, subject to substantiation, valuation, and percentage limitations. Contributions exceeding the current-year limitation may be carried forward when the statutory requirements are satisfied.

T03I08 : Casualty losses, gambling losses, and other itemized deductions. Personal casualty and theft losses are deductible only under the circumstances permitted by IRC Section 165. Gambling losses are separately governed by IRC Section 165(d) and are subject to their own limitation.

T03I09 : Qualified business income (QBI) deduction. IRC Section 199A permits certain noncorporate taxpayers to deduct qualifying business income, including eligible REIT dividends and publicly traded partnership income. Depending on the taxpayer's circumstances, wage, property, income, and specified-service-business limitations may apply.

T03I10 : Schedule 1-A deductions: qualified tips, overtime, vehicle-loan interest, and seniors. For tax year 2025, Schedule 1-A provides separate deductions involving qualified tips, qualified overtime compensation, qualifying vehicle-loan interest, and eligible seniors. Each deduction has distinct eligibility, dollar, and income limitations.

T03I11 : Ordering of basis, at-risk, and passive-activity loss limitations. Business losses can pass through multiple limitation regimes. Owner-level basis limitations under IRC Sections 704(d) and 1366(d) generally precede the at-risk rules of IRC Section 465 and the passive-activity rules of IRC Section 469, making the ordering of the calculations important.

T03I12 : Excess business losses and net operating losses. IRC Section 461(l) can further limit aggregate business losses after other restrictions have been applied. Disallowed amounts may affect the taxpayer's net operating loss under IRC Section 172 rather than being permanently lost.

B.4 T04: Property Transactions and Cost Recovery

This topic covers basis, depreciation, gains and losses, and special rules for acquiring, using, transferring, and disposing of property.

T04I01 : Property basis and holding periods. Basis represents the taxpayer's tax investment in property and is used to determine depreciation and eventual gain or loss. Cost basis generally arises under IRC Section 1012 and is adjusted under Section 1016, while IRC Section 1223 provides important holding-period rules.

T04I02 : Basis of gifted, inherited, and converted property. Gifted property generally follows the carryover-basis rules of IRC Section 1015, while inherited property generally receives basis under IRC Section 1014. Property converted from personal to income-producing use may require a different basis for depreciation or loss purposes.

T04I03 : Capitalizing costs versus deducting repairs; applicable safe harbors. IRC Sections 162 and 263(a) distinguish currently deductible business expenses from costs that must be capitalized into property. The tangible-property regulations provide additional rules and safe harbors for repairs, improvements, and certain smaller expenditures.

T04I04 : Business startup costs and amortization. IRC Section 195 governs expenditures incurred before an active business begins. Qualifying startup costs can receive a limited current deduction, while the remaining eligible amount is generally amortized over 180 months.

T04I05 : Depreciation and cost recovery. Depreciation requires determining depreciable basis, property class, recovery period, method, and convention. IRC Sections 167 and 168 provide the main framework, while Section 179 and additional first-year depreciation rules may accelerate deductions. Form 4562 coordinates these calculations.

T04I06 : Capital versus ordinary gains/losses and capital-gain netting. IRC Sections 1221 and 1222 distinguish capital assets and short- and long-term capital gains and losses. Form 8949 and Schedule D perform the principal reporting and netting, while IRC Sections 1211 and 1212 govern capital-loss limitations and carryovers.

T04I07 : Home-sale gain exclusion. IRC Section 121 allows qualifying taxpayers to exclude gain from the sale of a principal residence. Eligibility generally depends on ownership, residence, prior use of the exclusion, and applicable partial-exclusion rules.

T04I08 : Wash sales and nondeductible losses. IRC Section 1091 disallows specified losses when substantially identical securities are acquired within the wash-sale period, generally adding the disallowed loss to replacement-property basis. IRC Section 165 separately denies most losses on personal-use property.

T04I09 : Involuntary conversions and like-kind exchanges. IRC Section 1033 can defer gain from qualifying involuntary conversions when replacement-property requirements are satisfied. IRC Section 1031 separately permits nonrecognition for qualifying exchanges of business or investment real property.

T04I10 : Installment sales. IRC Section 453 allows qualifying sellers to recognize gain as installment payments are received. Payments must be divided among basis recovery, gain, and interest, while some items such as depreciation recapture may require immediate recognition.

T04I11 : Section 1231 business-property gains/losses and depreciation recapture. IRC Section 1231 can give qualifying business-property gains capital treatment while allowing qualifying net losses ordinary treatment. IRC Sections 1245 and 1250 may recapture depreciation as ordinary income. Form 4797 performs much of this characterization.

T04I12 : Qualified small-business stock exclusions and Section 1244 stock losses. IRC Section 1202 can exclude qualifying gain from qualified small-business stock, while IRC Section 1244 may allow qualifying small-business-stock losses to receive ordinary-loss treatment. Each regime has separate eligibility and limitation requirements.

T04I13 : Related-party property transactions. Transactions between related parties can receive special tax treatment under provisions including IRC Sections 267, 707, and 1239. Depending on the transaction, these rules may disallow losses, defer deductions, or convert gain into ordinary income.

B.5 T05: Credits, Payments, and Other Taxes

This topic covers credits that directly reduce tax, payments already made toward the liability, and additional taxes imposed outside the regular income-tax calculation.

T05I01 : Child tax credit and credit for other dependents. IRC Section 24 provides credits for qualifying children and certain other dependents. Eligibility depends on dependency, age, residency, identification, and income requirements, and part of the child tax credit may be refundable.

T05I02 : Child and dependent-care credit. IRC Section 21 provides a credit for qualifying care expenses that allow the taxpayer to work or seek work. Form 2441 applies the qualifying-person, expense, earned-income, and employer-benefit limitations.

T05I03 : Education credits. IRC Section 25A provides the American Opportunity Tax Credit and Lifetime Learning Credit. They have different eligibility and expense rules and must be coordinated with scholarships and tax-free education benefits through Form 8863.

T05I04 : Retirement savings contributions credit (Saver's Credit). IRC Section 25B provides a credit for qualifying retirement contributions. The credit depends on eligible contributions, filing status, AGI, and certain distributions, and is generally computed on Form 8880.

T05I05 : Adoption credit and employer-provided adoption benefits. IRC Section 23 provides an adoption credit, while IRC Section 137 can exclude qualifying employer-provided adoption assistance. Form 8839 coordinates qualifying expenses, reimbursements, income limitations, and special-needs rules.

T05I06 : Credit for the elderly or disabled. IRC Section 22 provides a limited credit for certain elderly or permanently disabled taxpayers. Eligibility and amount depend on age or disability, filing status, income, and specified nontaxable benefits.

T05I07 : Earned income credit (EIC). IRC Section 32 provides a refundable credit for qualifying workers. The amount depends on earned income, AGI, filing status, qualifying children, investment income, age, residency, and identification requirements.

T05I08 : Premium tax credit and reconciliation. IRC Section 36B provides a credit for qualifying Marketplace health coverage. Form 8962 reconciles advance premium-tax-credit payments with the taxpayer's final allowable credit using household-income and Form 1095-A information.

T05I09 : Residential clean-energy credit for solar installations. IRC Section 25D provides a credit for qualifying residential clean-energy expenditures, including qualifying solar installations. Form 5695 determines the allowable credit and coordinates the expenditure with other applicable tax benefits.

T05I10 : Withholding, estimated payments, refunds, and balances due. This issue reconciles federal withholding and estimated-tax payments with final tax liability. Excess payments generally produce a refund, while insufficient payments produce a balance due and can also trigger an estimated-tax penalty under IRC Section 6654.

T05I11 : Self-employment tax. IRC Sections 1401 and 1402 impose Social Security and Medicare taxes on net earnings from self-employment. Schedule SE computes the liability, while part of the tax generally produces a separate adjustment to income.

T05I12 : Additional Medicare Tax. IRC Section 3101(b)(2) imposes Additional Medicare Tax when specified wages, railroad compensation, and self-employment income exceed the applicable threshold. Form 8959 reconciles the final tax with amounts already withheld.

T05I13 : Net investment income tax. IRC Section 1411 imposes the net investment income tax on certain higher-income taxpayers. Form 8960 determines the tax using net investment income, modified adjusted gross income, deductions, and relevant passive-activity classifications.

T05I14 : Additional taxes, penalties, and reporting, including HSA-related treatment. This issue covers specialized taxes and reporting requirements associated with retirement and health-related accounts, including certain early distributions, excess contributions, required-minimum-distribution failures, and HSA transactions. Forms 5329 and 8889 implement many of these rules under provisions including IRC Sections 72(t), 223, 4973, and 4974.

B.6 T06: Business Entity and Owner Reporting

This topic determines how different business entities report their activities and how the resulting income, deductions, and losses reach an individual owner's Form 1040.

T06I01 : Business-entity classification and where owners report income. Treasury Regulations Sections 301.7701-2 and 301.7701-3 distinguish corporations, partnerships, and disregarded entities. The classification determines whether income is reported directly by the owner or passed through from a separate business entity.

T06I02 : Sole proprietorships and hobby activities. A genuine trade or business operated directly by an individual is generally reported on Schedule C, with qualifying expenses deductible under IRC Section 162. IRC Section 183 limits deductions when an activity is not engaged in for profit.

T06I03 : S-corporation shareholder reporting. S-corporation income, deductions, and credits generally pass through to shareholders under IRC Section 1366. Loss deductions depend on stock and debt basis under IRC Sections 1366(d) and 1367, while distributions are governed in part by Section 1368. Schedule E and Form 7203 support the shareholder-level reporting.

T06I04 : Partnership partner reporting. Partners generally report their distributive shares under IRC Sections 701–704 using Schedule K-1 and Schedule E. Partner basis changes with contributions, income, losses, distributions, and allocated liabilities under provisions including IRC Sections 705, 722, and 752, and that basis helps determine how much partnership loss can currently be deducted.

C ILLUSTRATIVE EXAMPLE

C.1 SUMMARIZATION OF CASE 2

In this example, the agent is initially told that the return is for tax year 2025 and that the Taxpayer was born on June 15, 1985, is not blind, cannot be claimed as another person's dependent, and has a valid work-authorized SSN. The agent is also informed that a marital-history record exists but that no spouse record is present. The Taxpayer maintained a home for all 365 days of 2025, and records concerning household-upkeep payments are available.

The Taxpayer has one potential dependent, Child A, and no other dependent candidates. The agent is told that Child A has available date-of-birth, relationship, residence, support, and SSN-eligibility records. Child A was present in the United States for all 365 days of 2025, with no temporary absences, was unmarried at the end of the year, is a U.S. citizen, did not file a joint return, and has no competing claimant. Child A's total living costs for the year were $12,000. The Taxpayer is also not subject to a prior post-2015 child-credit disallowance for a reason other than a mathematical or clerical error.

The scenario further indicates that dependent-care expenses were incurred for Child A. One care provider, Care Provider A, is identified as a licensed daycare center. The provider was not a household employee, complied with applicable state and local law, and provided care during 2025. The care enabled the Taxpayer to work. The Taxpayer was not a student and was not permanently and totally disabled during the year, and no prior-year qualifying care expenses were paid in 2025.

The agent is also told that the Taxpayer has one complete, uncorrected 2025 Form W-2 from Employer A. The W-2 belongs to the Taxpayer, reports no dependent-care benefits in Box 10, and does not designate the Taxpayer as a statutory employee. The Taxpayer had no digital-asset transactions during 2025, made no estimated federal tax payments, and did not elect to apply any 2025 overpayment to 2026 estimated tax.

Several facts necessary to complete the return are intentionally withheld and must be requested by the agent. The Taxpayer's marriage history is "never married," meaning that the Taxpayer was not married at any time during tax year 2025, including on December 31, 2025. The Taxpayer paid $24,000 of the home's upkeep costs during 2025, while $3,000 was paid from other sources. Child A is the Taxpayer's daughter, was born on April 12, 2018, and was therefore seven years old during most of tax year 2025 and at the end of the year. She lived with the Taxpayer for all 365 nights of 2025, paid none of her own support, and had a valid work-authorized SSN by the return due date. The Taxpayer's Form W-2 reports $55,000 of wages in Box 1 and $3,000 of federal income tax withheld in Box 2. Finally, the Taxpayer paid $6,000 to Care Provider A for Child A's care during 2025.

Taken together, these facts establish that the Taxpayer qualifies for the *Head of Household* (HOH) filing status for 2025. First, because the Taxpayer has never been married, the Taxpayer was unmarried on the last day of the tax year. Second, the total cost of keeping up the home was $27,000, of which the Taxpayer paid $24,000, or approximately 88.9%, and therefore paid more than half of the home's upkeep costs. Third, Child A is a qualifying person for HOH purposes: she is the Taxpayer's daughter, was seven years old at the end of 2025, lived with the Taxpayer for the entire year, did not provide more than half of her own support, was unmarried, and did not file a joint return. The home was therefore Child A's main home for well over half of the year. Because the Taxpayer was unmarried, paid more than half the cost of maintaining the home, and maintained that home as the principal residence of a qualifying child for more than half of 2025, the appropriate filing status is *Head of Household*.

C.2 PROMPTS

Our evaluation uses two prompts, corresponding to the two sides of the filing interaction. The *Tax Return Agent Prompt* is provided to the filing agent at the beginning of each case and specifies the overall objective, available files and local IRS references, privacy constraints, workbook-completion requirements, and when missing taxpayer information may be requested. In particular, the agent is instructed to ask only for information that is both unavailable from the provided materials and necessary to determine the correct tax treatment or complete a required field. Whenever the agent requests such information through the `ask_oracle` interface, the request is handled using the

Client Q&A Prompt, which simulates the taxpayer using only the held-out facts for that case. The client answers only single-fact questions, returns only the requested fact, and rejects multi-fact, open-ended, or blanket-assumption questions, thereby requiring the agent to identify and request missing information incrementally. We do not impose a separate limit on the number of Q&A interactions; agents may query the client as often as needed within the 30-minute wall-clock limit for the overall case.

Tax Return Agent Prompt

Review the contents of the `/app` directory. It includes:

- `/app/supporting.txt`, which contains non-sensitive information needed to prepare the tax return
- `/app/25_1040_template.xlsx`, which contains the federal income tax return template for tax year 2025

Objective
Using the information in `/app/supporting.txt`, complete the taxpayer's 2025 federal income tax return in `/app/25_1040_template.xlsx`.
Save the completed workbook to:
`/app/submission/submission.xlsx`
The `/app/submission/` directory already exists. Do not overwrite or modify the original template.

Filing scope
Complete only the worksheets and forms that are required as part of the taxpayer's final federal return filed with the IRS.
Do not complete unrelated worksheets, optional calculation sheets, or forms that are not required for the final filing.
For every required worksheet or form:

- Complete every applicable field.
- Enter `0` when an applicable numeric field has a value of zero.
- Do not leave an applicable field blank.
- Leave fields blank only when they are genuinely inapplicable.
- For checkboxes or other selection fields, enter `X` in the appropriate field.
- Preserve the workbook's existing structure, formatting, formulas, worksheet names, and cell layout unless a change is necessary to complete the return correctly.

Personal, Company, and Organization Identifying Information
To protect taxpayer privacy, do not enter any personally identifying or entity-identifying information, including but not limited to:

- Taxpayer, spouse, dependent, company, or organization names
- Street addresses
- Cities, states, and ZIP codes
- Social Security numbers
- Individual Taxpayer Identification Numbers
- Employer Identification Numbers
- Telephone numbers
- Email addresses
- Taxpayer, spouse, dependent, officer, partner, or paid-preparer identification information
- Signature, preparer, or firm-identification information

Leave all identifying fields blank. Do not ask the user to provide any identifying information.
An exact value that `/app/supporting.txt` explicitly labels as a fictional source display name is non-identifying benchmark data. Enter that value only when a required payer or source-description field needs it. Do not use this exception for any real, inferred, or differently labeled identity information.
Complete all applicable checkboxes, numerical fields, filing-status indicators, eligibility fields, and other non-identifying information necessary to calculate and prepare the tax return correctly.

Sources and default assumptions
Treat `/app/supporting.txt` as the primary source of taxpayer facts.
`/app/supporting.txt` is complete as to which tax-item categories and transactions exist. A wholly unmentioned category of income, gain, loss, adjustment, deduction, credit, payment, or additional tax is absent and must not trigger an oracle question. When an applicable numeric field requires the absent item's value, use `0`.
This rule does not authorize the agent to invent or default attributes of a topic that is mentioned. If the present information identifies a transaction, person-role, asset, benefit, expense, or potential eligibility path but omits a fact necessary to treat that topic correctly, the agent must ask the oracle for that one fact. The agent applies tax law to raw facts; it must not ask whether the taxpayer "claims" a credit or deduction that the law determines.
Apply the tax law and forms applicable to tax year 2025.

Requesting missing information
Ask for additional information only when all of the following are true:

1. The information is not available in `/app/supporting.txt` or elsewhere in the workbook.
2. The information is necessary to determine the correct tax treatment or complete a required field.

Keep the total number of questions to the minimum necessary.
Each question must be a **single-fact question**. It must request exactly one specific fact or one clearly defined value.
Permitted examples, when `/app/supporting.txt` already identifies the relevant topic, include:

- If `/app/supporting.txt` identifies Child A as a potential dependent but omits the child's date of birth, ask: "What was Child A's date of birth?"
- If `/app/supporting.txt` identifies gambling winnings but omits their amount, ask: "What was the amount of the gambling winnings identified in `/app/supporting.txt`?"

Do not ask broad questions that combine multiple tax categories or request blanket assumptions. For example, do not ask:

- "Do you have any other income?"
- "Can I assume that all of your other income was zero?"

Instead, for a topic already identified in `/app/supporting.txt`, ask separately about each specific item only when that item is necessary. For example:

- If `/app/supporting.txt` identifies unemployment compensation but omits its amount, ask: "What was the amount of the unemployment compensation identified in `/app/supporting.txt`?"
- If `/app/supporting.txt` identifies taxable interest but omits its amount, ask: "What amount of taxable interest was reported for the account identified in `/app/supporting.txt`?"

You may ask about a non-identifying status fact when it is necessary to determine tax eligibility or treatment, such as whether the taxpayer, spouse, or dependent has a valid Social Security number. However, **do not request the actual Social Security number, ITIN, name, address, or any other personal identifying information**.
Use the `ask_oracle` tool provided by the `oracle` MCP server to request any necessary information.
When possible, continue completing all portions of the return that do not depend on the missing answer before asking another question.

Final validation
Before finishing:

1. Confirm that all required forms and schedules have been completed.
2. Confirm that unrelated forms and worksheets remain uncompleted.
3. Check that applicable checkboxes and number fields are not blank.
4. Check that applicable zero-value fields contain `0`.
5. Check that required selections are marked with `X`.
6. Confirm that the completed workbook is saved successfully as `/app/submission/submission.xlsx`.

Offline environment
You have no internet access. Do not browse, download sources or packages, or use external search services. All permitted reference materials are local.
To call the same fact-only `ask_oracle` MCP tool, use the provided local helper: `python /app/ask_oracle.py 'one specific missing-fact question'`. It connects through a dedicated local socket, not the internet. The single-fact, minimum-inquiry and privacy rules above still apply.

Efficient workbook workflow

- Inspect relevant sheets together. Prefer `openpyxl.load_workbook(..., read_only=True)` for reading values; use editable mode when inspecting merged-cell layout or modifying the workbook.
- Reuse extracted information and batch edits instead of repeatedly loading the entire workbook.
- For long-running commands, use the tool's supported waiting mechanism or completion notifications. Avoid duplicate commands and repeated short sleep-and-poll loops.
- Save populated progress to `/app/submission/submission.xlsx` as you work, then update it after corrections. Saving a partial workbook does not mean the task is complete.
- Before finishing, verify that the saved submission opens and contains your intended entries.

Available references
The shared IRS reference library for tax-year-2025 returns is in `/app/references/irs/`. Read `README.md` for the document index and `manifest.json` for editions and provenance. Publications and instructions are HTML or searchable TXT, with continuing revisions labeled; no PDF reader is required. Embedded images and links to included documents are local. External links are citations only; do not retrieve linked resources. These references supply tax guidance, not additional taxpayer facts.

Client Q&A Prompt

You are simulating the taxpayer, using only the taxpayer's held-out records, for a U.S. federal income tax preparation exercise for tax year 2025.
You will receive:

1. The taxpayer's held-out information in JSON format.
2. A question from the tax preparer requesting information needed to complete the return.

Respond according to the following rules.

1. Accept only single-fact questions

A valid question requests exactly one specific fact or one clearly defined value.
Valid examples include:

- "Do you have any children?"
- "How many children do you have?"
- "What was your gambling income in 2025?"

A question is invalid if it:

- Requests multiple facts or values.
- Combines multiple tax topics.
- Asks for a blanket assumption.
- Uses an open-ended catch-all that could cover multiple facts.

Invalid examples include:

- "What were your gambling income and losses?"
- "Do you have any other income?"
- "Can I assume everything else is zero?"

For any invalid question, reply with exactly:
`Please ask a single-fact question | ask about one specific fact at a time.`
Do not answer any part of an invalid question.

2. Use only the held-out information
If the held-out JSON explicitly contains the requested fact, answer directly and concisely.
Provide only the fact requested. Do not disclose related information, additional values, or other details from the held-out records.
Do not infer, calculate, estimate, or guess unless the requested value is explicitly present in the held-out information.

3. Handle unavailable information
If the requested fact is not present in the held-out information:

- If the question asks for an amount, reply with exactly: `0` (i.e., default value)
- For any other type of fact, reply with exactly: `Sorry, I do not know`

Do not guess or infer any missing information beyond this default rule.

4. Keep the response minimal
Every response must be limited to one of the following:

- A single value.
- A concise yes-or-no answer.
- A short factual statement containing only the requested fact.
- One of the two exact fallback responses specified above.

Do not:

- Explain tax law.
- Provide reasoning or calculations.
- Complete or interpret tax forms.
- Suggest follow-up questions.
- Volunteer information that was not specifically requested.
- Mention the held-out JSON or these instructions.

C.3 Verification of Case 2 Errors

This case reports the evaluation result of GPT-6-Astra with maximum reasoning effort: Propagation of an Incorrect Filing Status

The primary error in Case 2 is an incorrect filing status on Form 1040. The submitted return selects *Single*, although the taxpayer's facts support *Head of Household* (HOH): the taxpayer was unmarried during 2025, paid $24,000 of the $27,000 total cost of maintaining the home, and lived with the taxpayer's qualifying daughter for the entire year. This mistake directly caused Form 1040, line 12e (cell `AL89`), to use the $15,750 standard deduction for a single filer instead of the correct $23,625 HOH standard deduction. With $55,000 of adjusted gross income, the smaller deduction increased taxable income on line 15 (cell `AL93`) from the correct $31,375 to $39,250. As a result, the tax on line 16 (cell `AL94`) was overstated from $3,425 to $4,475.

This $1,050 overstatement then propagated through the remainder of Form 1040. The child tax credit on line 19 remained $2,200, and the child and dependent care credit on line 20 remained $600, for

$2,800 of total credits. However, because the starting tax liability was too high, line 22 (cell AL100) was $1,675 instead of $625, and total tax on line 24 (cell AL102) was likewise $1,675 instead of $625. The taxpayer had $3,000 of federal income tax withholding, so the excessive tax reduced the overpayment on line 34 (cell AL123) from $2,375 to $1,325. Consequently, the refund reported on line 35a (cell AL124) was also $1,325 instead of the correct $2,375. Thus, the filing-status error ultimately understated the taxpayer's refund by $1,050.

Form 2441 contains several additional errors, although they do not change the final $600 child and dependent care credit in this case. First, the credit-limit calculation on Form 2441 is incomplete. Second, in the Credit Limit Worksheet for line 10, line 1 (cell V173) and line 3 (cell V177) should both contain $3,425, but both are left blank. Correspondingly, Form 2441 line 10 (cell O62) reports a tax-liability limit of $4,475 rather than $3,425. This error is another consequence of the incorrect Form 1040 tax calculation. However, because the preliminary child and dependent care credit is only $600, either tax-liability limit exceeds the credit. The final Form 2441 credit therefore remains $600, so this error does not propagate to Form 1040 line 20.

Schedule 8812 shows a similar propagation of the incorrect Form 1040 tax liability. Its Credit Limit Worksheet A was not completed: line 1 (cell AK91) should contain the $3,425 tax liability from Form 1040; the Schedule 3, line 2 credit (cell AE95) should contain $600; the total on worksheet line 2 (cell AE104) should therefore be $600; and worksheet lines 3 and 5 (cells AK106 and AK119) should both equal $2,825. These fields were instead left blank. On Schedule 8812 itself, line 13 (cell AL40) was entered as $3,875 rather than $2,825. The submitted amount can be traced directly to the same upstream error: $4,475 of incorrectly calculated tax less the $600 child and dependent care credit gives $3,875, whereas the correct calculation is $3,425 less $600, or $2,825. Nevertheless, the maximum child tax credit in this case is $2,200, which is below both limits. Therefore, Schedule 8812 line 14 and Form 1040 line 19 remain correctly reported at $2,200.

Overall, the errors in this case separate into two types, both contributed to tax filing noncompliance. The consequential error is the incorrect *Single* filing status on Form 1040, which leads to an understated standard deduction, overstated taxable income and tax, incorrect credit limits on Forms 2441 and Schedule 8812, and ultimately a $1,050 understatement of the refund. The remaining Form 2441 and Schedule 8812 errors concern omitted intermediate worksheet fields; although these fields are incorrect, they do not independently change the final credits claimed in this case.