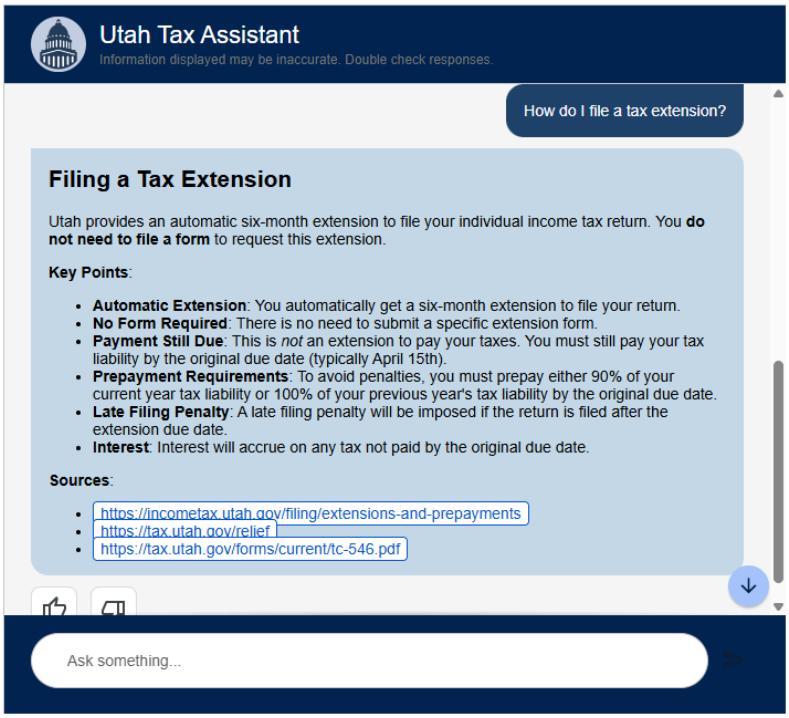
UTAH STATE TAX COMMISSION AI PILOT



Department of Government Operations, Division of Technology Services

Utah State Tax Commission, Income Tax Division, Motor Vehicle Division

Category: Artificial Intelligence

July 2024 - February 2025

Alan Fuller
Chief Information Officer
alanfuller@utah.gov
(801) 967-4603

Stephanie Weteling
Public Information Officer
stephanie@utah.gov
(435) 720-5315

The Utah Tax AI Pilot project, conducted from July 2024 through February 2025, evaluated the use of generative AI within the Utah State Tax Commission. The project's overriding purpose was to determine whether generative AI can output trustworthy, accurate, and consistent responses.

The project consisted of a unique "bake off" between four different vendor platforms across two phases. Vendor solutions demonstrated significant improvement in AI response accuracy from Phase I (73%) to Phase II (97%), leading to the current deployment of a production solution for the Tax Commission.

Additionally, the pilot project resulted in the creation of the State of Utah Division of Technology Services (DTS) AI Factory, streamlining future AI implementations within state agencies and across the enterprise. Importantly, lessons learned from the pilot are preparing DTS for citizen-facing generative AI solution pilots within the 2025 calendar year.

IDEA

Problem or opportunity addressed

ChatGPT introduced generative AI to the world in November 2022. Since then, individuals and organizations have explored ways to harness the power of this once-in-a-generation technology to increase efficiency, improve quality, ensure compliance, and enhance decision-making.

However, generative AI has a fundamental weakness: hallucination. DTS designed the Tax AI pilot to rigorously evaluate whether RAG-based generative AI solutions could meet the high standards of accuracy and reliability required by the Utah State Tax Commission specifically, and by extension, the State of Utah.

Why it matters

Trust is paramount for government operations, and the potential for generative AI models to hallucinate poses significant risks. Relying on inaccurate AI-generated information could negatively impact residents and erode public trust in government agencies.

That being said, the promise of generative AI deserves exploration. Utah executive branch agencies are entering into a period of financial austerity and uncertainty. If generative AI can live up to its potential, it could help agencies deliver essential services more efficiently while enhancing constituent experience.

What makes it different

The Utah Tax AI project was unique in its competitive "bake-off" approach, simultaneously evaluating multiple vendor platforms rather than a single solution. This approach brought the following benefits:

Direct comparative analysis	Instead of testing one vendor's capabilities in isolation, Utah designed the pilot to directly compare how different generative AI solutions performed using the same grounding data and the same set of test questions.
Data-driven decision making	Key metrics like accuracy, consistency, cost, ease of implementation, and handling of guardrails were evaluated across all platforms. This provided a much richer dataset for decision-making than a single-vendor pilot could offer.
Identifying the optimal fit	By testing multiple options concurrently, Utah could objectively determine which platform and underlying models offered the best balance of performance (accuracy, reliability) and cost for the specific Tax Commission use case.
Vendor-agnostic learning	The pilot provided broader insights into the state of RAG technology across the market, rather than just deep knowledge of one specific ecosystem. It

	addressed key questions around data ingestion, model selection, prompt engineering, response formatting, integration, latency, and deployment.

What makes it universal

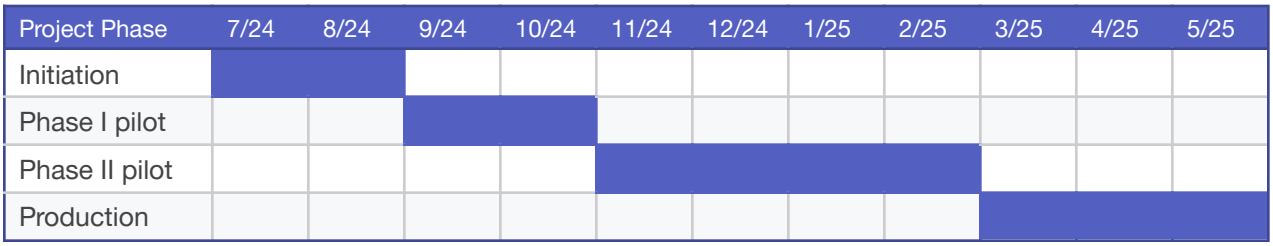
The Utah Tax AI project addresses fundamental challenges shared by most state governments concerning generative AI, aligned with CIO priorities identified in the 2024 NASCIO survey:

Priority #1 Cybersecurity and Risk Management	**Cybersecurity & Risk Management** Utah's pilot assessed how well different platforms adhered to guardrails and provided accurate, non-harmful responses, which relates directly to the universal priority of managing AI cybersecurity and technological risks.
Priority #2 AI/ML/RPA	**Evaluating AI trustworthiness** The core question Utah tackled – *Can we trust generative AI with RAG to provide accurate answers for government work?* – is universal.
Priority #3 Digital Government / Digital Services	**Improving government efficiency and services** The Utah Tax Commission aims to make call center agents more productive, a goal shared by agencies nationwide looking to enhance service delivery, reduce wait times, and manage workforce pressures.
Priority #4 Data Management and Analytics	**Data management and governance** Ensuring data quality, managing updates, and deciding how to structure data sources are universal challenges for states implementing AI.
Priority #6 Budget; #8 Cloud Services	**Navigating vendor solutions** States are often faced with multiple complex technology solutions from various vendors. Utah's approach offers a replicable strategy for any state evaluating AI technologies.

IMPLEMENTATION

Roadmap

During the summer of 2024, Utah envisioned the concept of a "bake off" to help determine whether chatbots could viably support internal state agency operations. When considering agencies, the Utah State Tax Commission was identified as an ideal organization for the pilot. From a project management and coordination perspective, DTS oversaw the project. The project commenced in August 2024 and concluded by early October 2024. After the conclusion of the evaluation, Utah decided to add a second phase, as reflected in the project timeline below:



Who was involved

DTS began meeting with vendors to gauge their interest in participating. Several were interested, as long as the process was equitable and transparent. DTS ensured all vendors received the same messaging, and proposed conducting blind testing to ensure all vendors were evaluated without bias.

Simultaneously, the DTS Director of AI engaged the Tax IT Director, who was instrumental in identifying and inviting key Tax Commission leaders and experts to participate. Tax nominated experts from the Income Tax Division as well as the Division of Motor Vehicles (DMV). Additionally, the DTS Cloud Hosting team was engaged to provision and configure environments as needed.

DTS prepared concise documentation outlining the project including proposed objectives, scope, timelines, roles and responsibilities, ensuring all parties involved had a common understanding of the project. Additionally, to secure support, the project was structured to minimize the level of effort required of the Tax / DMV experts and the Tax IT team.

How we did it

Phase I (August - October 2024)

An in-person project kick-off meeting with all vendors, DTS project team members, and Tax / DMV experts was held in late August 2024. All project elements were discussed and agreed upon by all parties.

The Tax / DMV experts provided 366 questions to be included in the pilot. To make it easier for vendors, the experts recommended using the publicly available tax.utah.gov and dmv.utah.gov subdomains as data sources for the pilot.

No DTS staff were experts in the generative AI platforms provided by the vendors for the pilot. Therefore, DTS asked vendors to manage their configurations directly, in their own environments. Each vendor was left to optimize their configuration as they saw fit.

To ensure transparency in testing, DTS requested direct access to vendor platforms to submit questions and retrieve responses. The DTS Director of AI imported questions into the platforms, downloaded responses, placed them into spreadsheets, and circulated them to the Tax / DMV experts for evaluation. This process saved Tax / DMV experts significant time and effort and removed the need for end-user training on the various platforms.

The Tax / DMV experts evaluated questions based on their areas of expertise and, for each question, scored them as follows:

- Ranked each vendor response from best to worst, assigning a 4 to the best and a 1 to the worst.
- Scored each response for accuracy:
    - 4 - equivalent or better than a knowledgeable human
    - 3 - accurate but perhaps missing some information (e.g., link to a source)
    - 2 - response contains inaccuracies
    - 1 - system did not answer the question

In addition, DTS conducted consistency testing to evaluate the similarity of responses over time and guardrail testing to evaluate how the solutions might handle inappropriate questions. Testing and evaluation took place during September and October 2024. DTS compiled and presented the results in a pilot debrief session held in mid October 2024.

Phase II (November 2024 - February 2025)

The Tax / DMV experts were encouraged by the pilot results. However, the vendors believed they could improve on the Phase I scores. The decision was made in November to initiate a second (previously unplanned) phase of the pilot.

DTS hired two AI Analysts in September and October. For Phase II, it was recommended that the new DTS AI Analysts work closely with the vendors to further optimize their RAG solutions. This required configuring environments within the DTS technology stack.

Provisioning environments took longer than anticipated. Two of four vendors were able to provide credits and the DTS Cloud Hosting team was able to stand up dev environments in December 2024. The other two vendors needed more time, and their environments were finally provisioned and configured in February 2025. One of those proved quite challenging to configure to a minimally viable level of accuracy, and further testing was discontinued.

During that time, the AI Analysts conducted testing using a subset of questions from Phase I that were scored lower by the experts. The AI Analysts employed a range of techniques to evaluate responses. Tuning parameters and configurations were adjusted through multiple iterations in Phase II to maximize accuracy.

Production deployment (March 2025 - present)

Analyses were concluded at the end of February 2025 and showed marked improvement. In March, the results were presented to DTS leadership along with a recommended vendor solution. The same were presented to Tax.

Tax was pleased with the results and agreed to move to a production solution. The Tax IT Director identified resources to support the deployment. As of this writing, production deployment is in progress.

IMPACT

What the project made better

The Tax AI pilot project proved that generative AI could be successfully used internally within the Tax Commission and, by extension, other state agencies.

The story: from risk to reliability

Like many government agencies, the Tax Commission saw the potential for generative AI to help its approximately 200 call center agents become more effective. However, providing inaccurate or “hallucinated” tax information carries significant risks for residents and could erode public confidence.

The pilot directly confronted this uncertainty. Systematically testing multiple vendor solutions using RAG architecture grounded in Utah's specific tax knowledge base answered the fundamental question: Can this technology be trusted for government work?

The results demonstrated that, yes, with careful implementation, RAG can achieve high levels of accuracy. The pilot showed significant improvement through refinement, with average top-accuracy scores increasing from 61% in Phase I to 83% in Phase II. It identified viable, cost-effective platforms capable of meeting the Tax Commission's needs.

Environment before and after the project

Prior to the Tax AI pilot, the Utah state government was uncertain whether generative AI solutions could produce trustworthy results. The pilot proved they could.

Before	**After**
High uncertainty: Concerns about the reliability and safety of using generative AI for sensitive Tax / DMV queries.	**Demonstrated trustworthiness**: Specific RAG solutions, when properly configured and grounded, can achieve high accuracy and reliably answer questions.

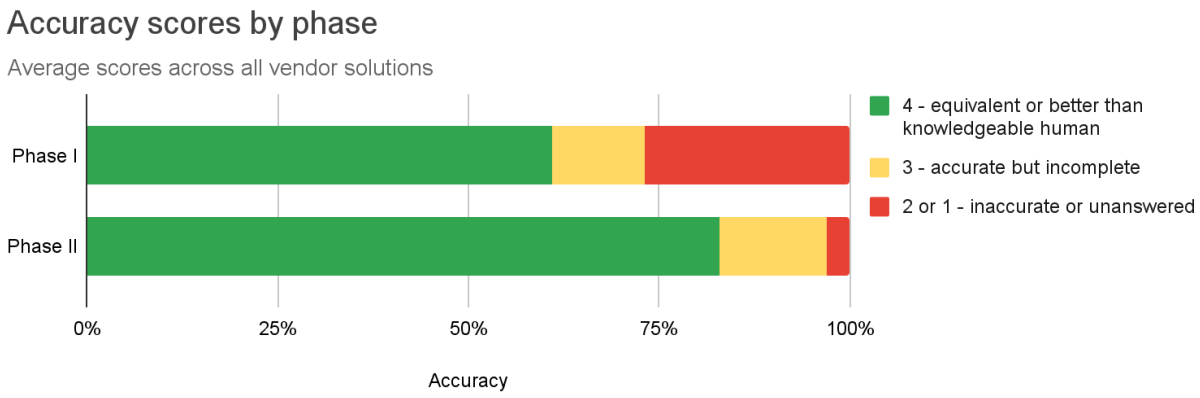
Lack of comparative data: No empirical basis for comparing how different major AI platforms would perform using Tax Commission data and the RAG approach.	**Evidence-based platform choice**: Clear, comparative performance data across established vendors on accuracy, consistency, cost, and implementation complexity, leading to a specific recommendation and selection decision.
Undefined path forward: No clear, low-risk pathway to implement generative AI assistance for call center agents. Potential benefits were overshadowed by the risks.	**Viable implementation path**: Specific, tested approach and platform for moving forward with AI agent assistance, de-risking next steps, and serving as the foundation for the establishment of the DTS AI Factory.
Unknown costs / ROI: Difficulty in budgeting or justifying investment due to unknown implementation and operational costs.	**Budget clarity**: Concrete estimates of annual operational costs enabling informed financial planning. Favorable cost trends were noted from Phase I to Phase II as vendor business models evolved.

How we know

The Tax AI pilot project built confidence in the viability and trustworthiness of using RAG generative AI solutions, based on assessments conducted throughout its structured evaluation phases.

Accuracy scores by phase

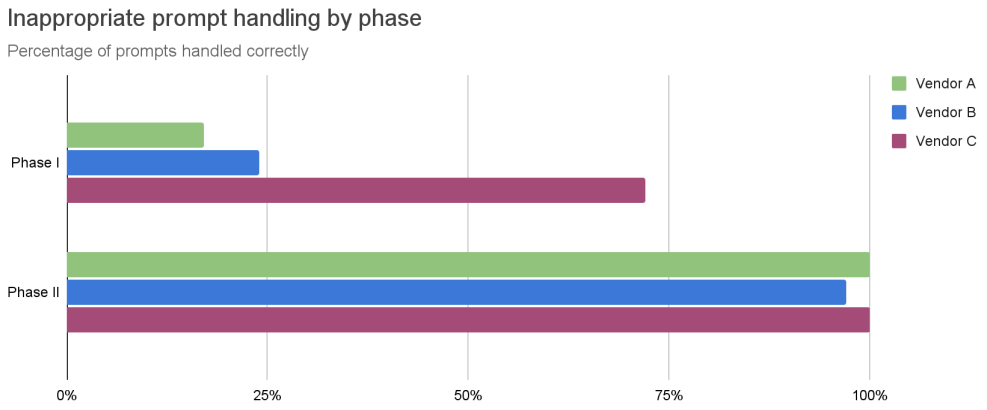
The following chart shows overall accuracy across all vendor solutions tested in both Phase I and Phase II. Responses rated as equivalent or better than a knowledgeable human increased from 61% in Phase I to 83% in Phase II. Responses rated as accurate (3s and 4s) increased from 73% to 97%.



Following Phase I, newer, better, and more cost-effective models and features have been introduced. DTS is working on further fine-tuning and configurations that may increase accuracy to 99%, well above the accuracy of the average call center agent.

Guardrails by phase

The following chart shows improvement in guardrails and inappropriate prompt handling from Phase I to Phase II. Thirty cleverly worded prompts that the systems shouldn't answer were processed.



Consistency

Delivering consistent responses over time can be challenging for generative AI. To that end, DTS conducted a range of consistency tests using the FuzzyWuzzy Python library, which offers several string comparison metrics, each tailored for different scenarios. The `ratio()` function computes the Levenshtein Distance, representing the minimum single-character edits needed to make two case-sensitive strings identical, and scales this to a 0-100 score. For situations where one string might be a substring of another, `partial_ratio()` identifies the best-matching substring. To disregard word order and case sensitivity, `token_sort_ratio()` sorts the words alphabetically and compares the lowercase strings. Finally, `token_set_ratio()` delivers a more resilient comparison by creating token sets, effectively ignoring duplicate words and word order, and then performing comparisons on these sets.

All vendors scored near or above 90 in the Token Set Ratio test, indicating a very high degree of similarity and a strong likelihood that strings share almost all the same important tokens.

Platform	Ratio	Partial Ratio	Token Sort Ratio	Token Set Ratio
Vendor A	79	78	84	91
Vendor B	78	78	87	92
Vendor C	78	78	85	89

Return on investment

Generative AI has brought with it unprecedented speed of change, not just in the technology but also vendor business models and pricing plans. At the conclusion of Phase I, solution pricing varied greatly, from high five-figure to mid six-figure annual pricing for a production Tax AI agent assist solution. By the end of Phase II, vendor pricing averaged around $10,000 per year.

Furthermore, the level of effort from the agency business staff, agency IT team and enterprise IT resources to deliver and deploy RAG-based solutions is low compared to most application development projects. That being said, ROI was not a primary element of the project scope. Instead, the scope was to determine whether responses could be trusted.

What now

The Tax AI pilot solution is currently deploying to production, where it will assist agents in quickly handling incoming resident queries. Moving forward, DTS is concentrating on four major initiatives spawned by the Tax AI pilot. First, preparing agencies for long-term maintenance of GenAI RAG solutions. Second, iterating the AI Factory to increase its efficiency and quality. Third, measuring the impact of generative AI solutions deployed both within specific agencies and at the enterprise level. And fourth, applying lessons learned from the Tax AI and other generative AI pilots to prepare for piloting citizen-facing solutions, beginning in late 2025.