A **GOVERNMENT TECHNOLOGY** THOUGHT LEADERSHIP PAPER | AWS AND DELOITTE

The case for domain-specific generative AI

Why agencies should explore smaller, more targeted AI models

Generative AI (GenAI) applications have the powerful ability to examine massive data sources and generate quick and relevant responses to questions in a conversational format.

Unfortunately, a consumer-facing large language model (LLM) often lacks precision and poses data-protection risks because user inputs in the chat interface might be used to further train and expand the model itself. The possibility of leaking personally identifiable information (PII) or protected health information is a major concern — which limits the usefulness of general-purpose LLMs for government.

"General-purpose LLMs do a lot of things very well, but might provide overgeneralized or even inaccurate information, especially for complex, nuanced, or highly specialized topics" says Ronnie Thomas, a Deloitte Consulting specialist leader with more than 15 years of experience in digital transformation, artificial intelligence, and machine learning.

One attractive alternative for agencies is a domain-specific LLM that can be used as a service or locally hosted to improve data protection, manage costs, and increase precision.

Why domain-specific LLMs?

Domain knowledge. LLMs can be trained to understand government terminology or nuances. For example, in human services programs, terms like SNAP and CHIP must be understood in the correct context.

Security and control. For use cases with sensitive data, agencies can fine-tune a smaller LLM for the specific task and domain. Smaller LLMs can be hosted on the agency's own cloud instance with a cloud service provider like Amazon Web Services (AWS). There, the agency can maintain better control of where the data is transmitted and stored.

Cost optimization. Large-scale LLMs offer great flexibility but can be expensive to run. While smaller, domain-specific LLMs may incur higher costs for training and hosting, they can offer lower compute costs, which may prove more economical for large-volume usage.

Four promising use cases

1. Knowledge retrieval. "Today, if a caseworker has to figure out the policies for a situation, they have to either go ask somebody experienced with that scenario or look through several pages of policy documents," Thomas says. A domain-specific LLM coupled with retrieval-augmented generation — which improves the accuracy of an LLM by linking it to authoritative external sources — can answer such a question almost instantly. Fast answers to complex queries help agency staff make better decisions and provide quicker turnaround to constituent questions.

2. Content generation. Agencies can use GenAI to create knowledge bases for staff and the public based on existing documentation. Domain-specific LLMs can be used to draft social media campaigns, marketing material, press releases, client communication, meeting summaries, and more. This can save significant time and effort for staff and provide frequent and proactive communication.

3. Insights from unstructured data. A significant amount of information is unstructured data that can be difficult to interpret. Domain-specific LLMs can identify hidden insights from unstructured data and help streamline business processes. For example, by identifying critical themes from case notes, unstructured data in case management systems, and telephone conversations with constituents, caseworkers can get insights into dynamics of a family and develop more tailored and data-driven plans.

4. Constituent engagement. Domain-specific LLMs can respond to queries from constituents quicker and reflect the agency's style and values. Additionally, by using multi-language capabilities in LLMs, agencies can more easily engage diverse populations and non-English speakers, helping them get access to the right services at the right time.

Governance and best practices

To optimize domain-specific GenAI capabilities, an organization should establish sound governance policy in the following key areas. "A strong data governance framework should align not only with your internal policies but also with the new AI rules and regulations from different governance agencies," Thomas says.

- **Fairness.** Continuously monitor and update an LLM to ensure its responses do not perpetuate biases or discrimination. Carefully curate training datasets to represent diverse perspectives and demographics.

- **Security.** Safeguard models from attacks. Make sure the model does not generate sensitive or personal information. Using a secure GenAI service like Amazon Bedrock, implement strong data governance during training to prevent sensitive or PII data from being distributed to third parties. Your LLM should also include robust protocols to resist malicious attacks and any attempts to misuse the model. Because threats evolve, assess your security and data governance approach on an ongoing basis.

- **Accuracy.** Consumer LLMs are notorious for occasional "hallucinations" — confident responses that contain false or imaginary information. To achieve high accuracy in responses, fine-tune the model with high-quality, domain-specific datasets. "You also need a concept called 'grounding' that forces the LLM to use your knowledge base — and only your knowledge base — for answering questions," Thomas says. A solution should provide citations and reasoning for its answers and inform the user when it lacks the data to answer a question.

- **Feedback.** LLMs should not replace human judgment. Integrate feedback mechanisms into the deployment of the model to collect users' experiences, suggestions, and criticisms. Analyze this feedback for insights into the model's performance, user satisfaction, and areas for improvement. "Your users should be able to flag something and say, 'This answer does not seem correct.' That is crucial for building trust in your solution," Thomas says.

Looking ahead

Most government agencies oversee immense quantities of data, which makes fine-tuning LLMs an attractive proposition. But what about smaller jurisdictions with smaller data sets? "Even a small local agency that does not have a huge amount of data may be able to tackle a specific use case," Thomas says. Finding the right partner for planning and implementation is essential to answering these kinds of questions and realizing your GenAI ambitions.

This piece was written and produced by the Government Technology Content Studio, with information and input from AWS and Deloitte.

government technology

Produced by Government Technology
Government Technology is about solving problems in state and local government through the smart use of technology. Government Technology is a division of e.Republic, the nation's only media and research company focused exclusively on state and local government and education.

www.govtech.com

aws

Sponsored by AWS
Amazon Web Services (AWS) Worldwide Public Sector helps government, education, and nonprofit customers deploy cloud services to reduce costs, drive efficiencies, and increase innovation across the globe. With AWS, you only pay for what you use, with no up-front physical infrastructure expenses or long-term commitments. Public Sector organizations of all sizes use AWS to build applications, host websites, harness big data, store information, conduct research, improve online access for citizens, and more. AWS has dedicated teams focused on helping our customers pave the way for innovation and, ultimately, make the world a better place through technology. To learn more about AWS in the public sector, visit us at **aws.amazon.com/stateandlocal**.

Deloitte.

Sponsored by Deloitte
Deloitte provides industry-leading audit, consulting, tax and advisory services to many of the world's most admired brands, including nearly 90% of the Fortune 500® and more than 8,500 U.S.-based private companies. At Deloitte, we strive to live our purpose of making an **impact that matters** by creating trust and confidence in a more equitable society. We leverage our unique blend of business acumen, command of technology, and strategic technology alliances to advise our clients across industries as they **build their future**. Deloitte is proud to be part of the largest global professional services network serving our clients in the markets that are most important to them. Bringing more than 175 years of service, our network of member firms spans more than 150 countries and territories. Learn how Deloitte's approximately 457,000 people worldwide connect for impact at **www.deloitte.com.**

ADOBESTOCK/AI GENERATED