AI DIALOGUES

Navigating the Generative AI Landscape

10 QUESTIONS INFORMED EXECUTIVES ARE ASKING (AND OUR CURRENT ANSWERS)

DECEMBER 2023

WILLOWTREE®
a TELUS International Company

A Note from WillowTree

In my free time, I ride my bike. It's a way to slow down, reset, and interact with my surroundings more deliberately than when driving a car or hustling through an airport.

As head of WillowTree's Data & AI practice, it's easy to get caught in the breakneck pace at which this tech is moving. I enjoy this speed and the near-constant innovation — a space full of opportunities, pitfalls, and downright wonder.

This booklet is my attempt to briefly slow down, take a look around, and share what I'm hearing from clients and seeing from our leading-edge practitioners.

In the following dialogue, we aim to demystify ten questions we often hear from clients interested in applying generative AI, alongside our most current answers and approaches based on real-world experience.

To be sure, this is not an evergreen document. We'll be updating this dialogue regularly with the latest questions business leaders should be asking, alongside our most current suggestions.

Stick with us on this journey, and together, let's enjoy the ride.

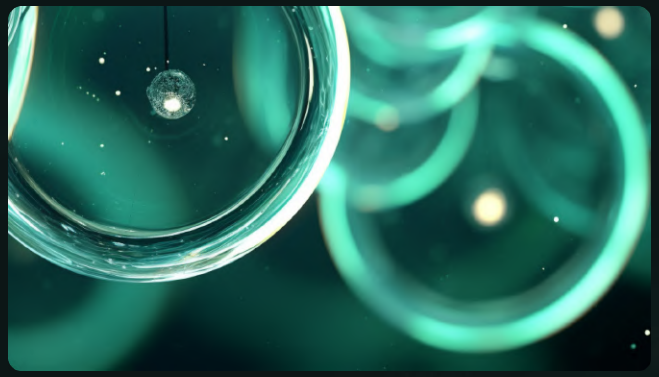
— Patrick Wright

Chief Data and AI Officer

WillowTree, a TELUS International Company

Q1

Should we invest in generative AI now, despite the risks and frequent changes, or should we wait to see how this all shakes out?



A: “

The answer is to do both. Choose where to invest based on risk, ease of implementation, and opportunity to learn.

It’s undeniable that generative AI is set to have a significant impact across many varied industries. This technology isn’t just about automation or efficiency; it’s about creating new possibilities in product development, customer engagement, and internal processes.

However, this potential carries some risk. Each of our clients has unique needs, business objectives, and varying risk tolerance levels. The generative AI landscape is changing quickly, with new developments and ethical considerations emerging weekly. That fluidity can be daunting, and it’s natural to think about adopting a “wait and see” approach, particularly in more heavily regulated industries.

Start with smaller, lower-risk projects that help your team learn or increase productivity, deploying tools like GitHub Copilot (a coding assistant) or ChatGPT. Next, experiment with conversational AI assistants loaded with applicable business content so employees can find information fast.

It’s tempting to wait on projects with higher risk. But tools for managing issues like AI hallucinations are improving fast, and inertia is its own risk, so push your team to grow their skills and prepare the foundations. Generative AI is heavily dependent on data quality and diversity, for instance. By initiating projects now, we can better prepare our data infrastructure, ensuring it’s robust and well-suited for future, more extensive AI endeavors.

Q2

How do we know if generative AI is the appropriate technology solution for a particular use case or business problem?

A: “

If the use case includes verbs like Find, Distill, or Create, that’s a generative AI problem. That’s where these non-deterministic systems are so strong.

- Find: If you want to build a semantic search, we’ll use a retrieval augmented generation (RAG) technique and ask an LLM to summarize the returned results.
- Distill: If you’ve got a megabyte of text about a given topic, and you want to draw some themes out of it, that’s a distillation problem where we’d use generative AI, passing the text to an LLM.
- Create: If you’re considering creative capability, prompting ChatGPT to write a poem is just the tip of the iceberg. There’s massive customer service potential here: suppose you have access to a given user’s buying history or their response to specific messages. In that case, you can take this history, understand their interests, and use creation to generate unique marketing messages that resonate well with them.

However, if the problem is Decide — here’s a large pile of data, make a decision based on it — you instead want to use something deterministic to crunch all that data and provide you with some sort of probabilistic answer about what to do. The classic example is looking at a funnel of leads and deciding which leads might convert. In that case, we’d recommend a predictive model using machine learning techniques to determine if a new lead is worth pursuing.

In short, there are many business problems that can be solved more immediately without generative AI, and it’s worth exploring these possibilities with an experienced partner before jumping straight to generative AI.

Q3

How do you guide clients to maximize ROI in their generative AI projects?

We always start with this series of questions:

- What segment of users are you hoping to benefit?
- How many users are you going to serve?
- How far can you move the needle for them?
- How much effort will you have to put into risk mitigation?

A: “

You can choose a vast user population with a huge potential benefit, but if you’re introducing a lot of risk along the way, it can backfire. So, here are a couple of sweet spots with high ROI:

1. Business-to-employee tools: when you’ve got a narrow population of people in your organization trying to accomplish a certain task or a certain kind of work, we can build generative AI tools that focus on enabling and supporting that task. These are still very human-in-the-loop, but we’re removing a bunch of the cognitive load or helping employees summarize information so they can move faster. What’s nice about these use cases is you’ve got a captive user audience, so you can survey them, you can see how it’s working for them, you can iterate and have them feel the change as you implement, and then you can adjust the system to maximize benefit.

2. End-user AI assistants: Maybe you’ve got a platform or site that’s used by tens or hundreds of thousands of people in the general public; they need information from you, and they’re going to ask different questions that need accurate but nuanced answers. You could use a traditional rules-based “chatbot,” but you might benefit from a “conversational AI assistant” approach for more complex information spaces.

We utilize a variety of established frameworks, such as the RICE Framework (Reach, Impact, Confidence, Effort), that we find really helpful in driving prioritization and go/no-go decisions for projects like these.

Q4

How do we ensure effective governance without sacrificing speed and innovation?

A: “

These are critical issues, particularly for clients operating in highly regulated spaces with heightened legal concerns around data sovereignty, security, and privacy. For any organization in healthcare or financial services, for instance, these are crucial, and there are data sovereignty laws around this that are crystal clear.

And then there’s another layer on top of that: consideration of user harm. This is where things get a little fuzzier. Every time we build a system, whether AI-backed or not, we have to think: is this system well architected? Is it easy to use? And, could it cause distress to a user?

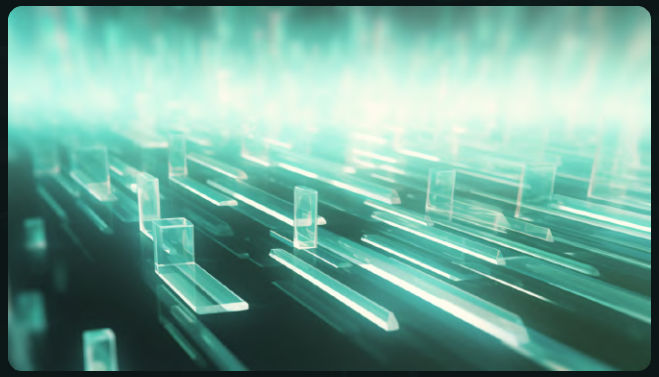
We take a pragmatic view, aiming to achieve our goals of data security while also managing for user harm. We use a couple of key techniques:

Internally, we advise clients to create an oversight and governance body within their organization. We help them create a set of standards to apply across all projects to ensure AI-backed projects are operating correctly. These standards should be communicated clearly and updated regularly.

We employ various established frameworks for responsible AI, like NIST, HHS’s Trustworthy AI Playbook, and WillowTree’s “defense-in-depth” approach to AI hallucinations, which we employ to help clients mitigate and minimize harm.

Q5

Should we only consider OpenAI's GPT because it's the current LLM leader?



A:

When building products and platforms, particularly early in the evolution of new capabilities, it's never been a good idea to lock your solution to a single vendor's offering. Recent upheavals at OpenAI and ensuing uncertainty about the future of their products bear that out.

OpenAI has built a Swiss Army Knife with GPT-3.5 and -4.0: a wide range of impressive capabilities, and they do a few things really well. But GPT-4, in particular, is expensive to run in production at scale.

That may be fine when it's just an individual interacting with the site, and it's driving productivity. But if you build an app that uses an LLM in the background to synthesize data and scale that up to 20,000 or 50,000 users per day, all of a sudden, the cloud computing bill gets quite pricey.

We can likely deliver the same or better performance if we build an ensemble solution that combines instances of smaller, more compact models we can train, fine-tune, and prompt to get similar qualities of answers for a given niche application.

Q6

How should we consider training & inference (query) costs vs. performance tradeoffs?

A: “

First, it’s important to realize that training a large language model from scratch is currently a costly, resource-intensive endeavor. What we’re frequently discussing with clients in practical scenarios is the process of “fine-tuning,” like customizing a high-performance car to suit specific racing conditions rather than building it from the ground up.

We advise an evaluation-driven approach. This means first assessing which out-of-the-box model is closest to your needs and then validating that any fine-tuning continues to provide better results. It’s not just about finding the most powerful model; it’s about finding the right fit. Sometimes, a smaller and less expensive model fine-tuned for a specific task outperforms a larger, more expensive model.

Fine-tuning is more effective for adjusting the format and style of responses rather than injecting new knowledge into the model. If you want to enhance the model’s fact-based knowledge, technologies like retrieval augmented generation (RAG) are more suitable. They combine the generative capabilities of LLMs with the ability to pull in information from external sources.

There’s also an interesting trade-off between fine-tuning and prompt engineering. Take, for example, the goal of aligning an LLM with your company’s specific tone and voice. While you can achieve this by continuously providing detailed context in each prompt (prompt engineering), this can become costly at scale. In such cases, fine-tuning the model to inherently understand and reflect your company’s style might be more cost-effective since you’re no longer paying to send the same style-based context with every prompt.

Q7

How should we consider using LLMs as a service versus deploying open-source models in the cloud?

A:

You can run models in two ways:

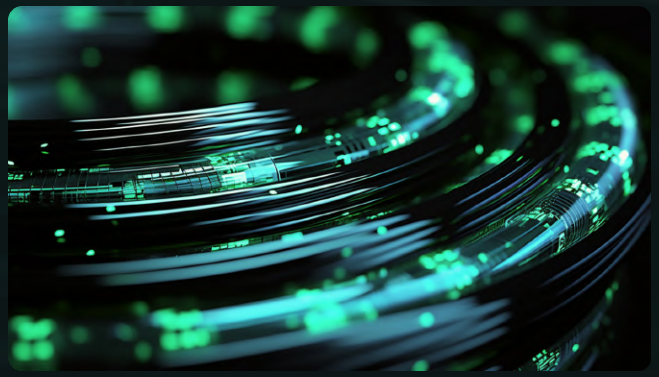
Proprietary (e.g., GPT) or open-source (e.g., Llama2) models can be accessed on services like Microsoft Azure via an API. You pay for usage — so tokens in, tokens out in response — and total cost is determined by how much data is moving.

You can run open-source models (or one you've fine-tuned yourself) on your own or cloud hardware. Using your own incurs an upfront purchasing cost, and you run some obsolescence risk, but long-term, it can be quite inexpensive. Cloud hosting is just renting the hardware — so there's less upfront cost and more portability but higher expense in the long run.

The bigger the model, the more hardware and memory you need to run it. So, again, where you can figure out how to get the same output quality from a smaller model that's trained more narrowly on the specific problem you face, you'll get better results, and you'll bring your inference (query) costs down. But you'll likely incur more training costs on the front end.

Q8

How do we approach data readiness?



A: “

Garbage in, garbage out: we’ve heard it 100 times. If you’re trying to build a model, you have to have good, clean data.

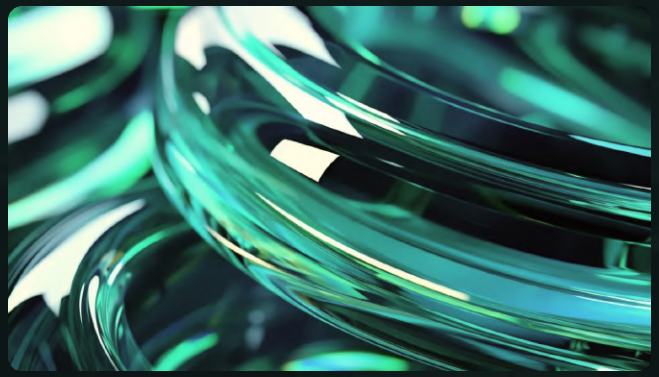
You may have to first deal with issues of duplication or contradiction. For example, imagine you’re building a system based on your company’s HR policies, and you have contradictory or ambiguous information in different parts of your documents. You have to go back and clean up that data. So, both in the unstructured text world (where you’ve got those kinds of policy documents) and in the world of structured data (where you want the LLM to query this data and pull it together), the data has to be clean.

It’s the same kinds of readiness tasks that you have to prioritize when doing analytics work for your business or building metrics around your business’s performance — data pipelines, data lakes, data prep, data visualization, all of that work counts towards this.

Where you have a specific problem and can narrow down the data sets required to make an LLM work, you can all go back upstream regarding your pipelining, skinny down your focus, and do less work to get value out.

Q9

How do we measure and optimize our model's performance?



A: “

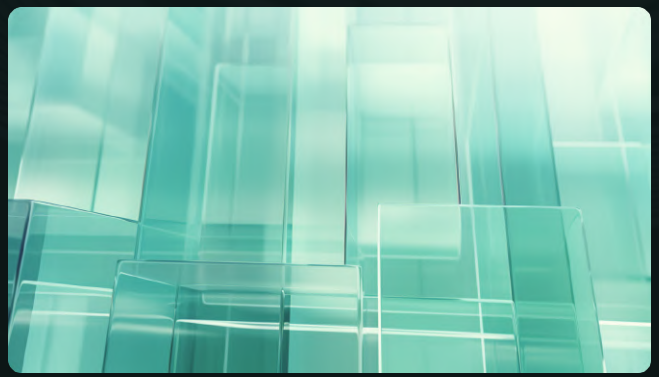
At WillowTree, we've built a benchmarking practice: for every client problem, we're evaluating and selecting models for custom uses, thinking deeply about the users, their expectations, how to shape prompts and responses, etc. Then, we build a benchmark around that, which also includes some measurement of cost.

We run this benchmarking against three or four models that look applicable, we take those results, and we pick the top two or settle on one if there's a clear winner. Then we get into fine-tuning, prompt–response optimization, etc., to drill down into what clients really want out of that model so that they can make a final choice.

Again, all of this is a multivariate problem. You can't just do simple math against it, but you can gather data to help you think about the problem and then work to solve it.

Q10

How do we ensure our system's safety, consistency, and integrity over time, especially for higher-risk industries?



A: "

As we work through development — integrating a RAG solution or fine-tuning, for instance — we use the concept of evaluation-driven development to keep us on track. We'll run our evaluation suite against the application frequently, quantifying improvements we've made or reverting our changes if we need to.

What we sometimes see, particularly with model-as-a-service, is occasional "model drift": a given LLM, over time, shifts its behavior. You may also see "prompt drift": for a given prompt, responses change slightly over time. And so we monitor for that as well.

For highly regulated industries, we also employ a Dual-LLM Safety System incorporating:

1) an application layer that takes user input and generates a response from an LLM trained to adhere to specific policies, and

2) a "Supervisor" moderation layer auditing that response to catch anything that gets through the initial prompt and ensures the response adheres to specified policies rather than broader, more general, "world knowledge."

We've found a lot of success here in instances where risks are high, whether it's concern around AI hallucinations (misinformation), jailbreaking (malicious prompting), or industry-specific rules (not providing financial advice).

Data and AI Services

No matter your starting point, WillowTree's multidisciplinary teams can accelerate your generative AI roadmap.

Offerings

AI Strategy and Governance

- AI Intent and Use Case Prioritization Workshops
- Executive Education
- AI Governance Frameworks and Enterprise Councils

Generative AI Proof of Concept

Production Build and Launch

- Data Readiness Audit
- LLM Benchmarking and Selection
- Multi-platform Software Development
- Systems of Record Integration

Conversational AI and Voice Technology

- Voice-first Experience Design
- Voice Tech Selection and Integration

AI Expertise Across Disciplines

- Data Scientists
- Data Engineers
- AI Architects
- MLOps Engineers
- AI Strategists
- UX Researchers
- Conversational UI Designers
- Cross-Platform Developers

GenAI Jumpstart

Unlock the power of generative AI and transform your organization with WillowTree's GenAI Jumpstart accelerator. Ready to turn your ideas into tangible outcomes, responsibly and quickly? Partner with WillowTree as we design and develop a generative AI-powered virtual assistant in just 8 weeks.

Deliverables

- Working prototype / proof of concept of a generative AI-powered virtual assistant
- Reference architecture and model selection
- Guardrail strategy for minimizing hallucinations
- Path to production implementation plan

Core activities in an 8-week **GenAI Jumpstart** engagement

- AI Outcomes Workshop
- Guardrail Strategy
- Prototype Development
- Model Fine-tuning
- Production Implementation Planning
- LLM Benchmarking & Model Selection

Get a jumpstart!

Trusted by the World's Most Admired Companies

Wawa · ABInBev · dexcom · DQ · synchrony · TELUS · Marriott · FOX

WILLOWTREE®
a TELUS International Company

AI DIALOGUES

Navigating the Generative AI Landscape

10 Questions Informed Executives are Asking (and Our Current Answers)