Beyond the Buzz!!

About us

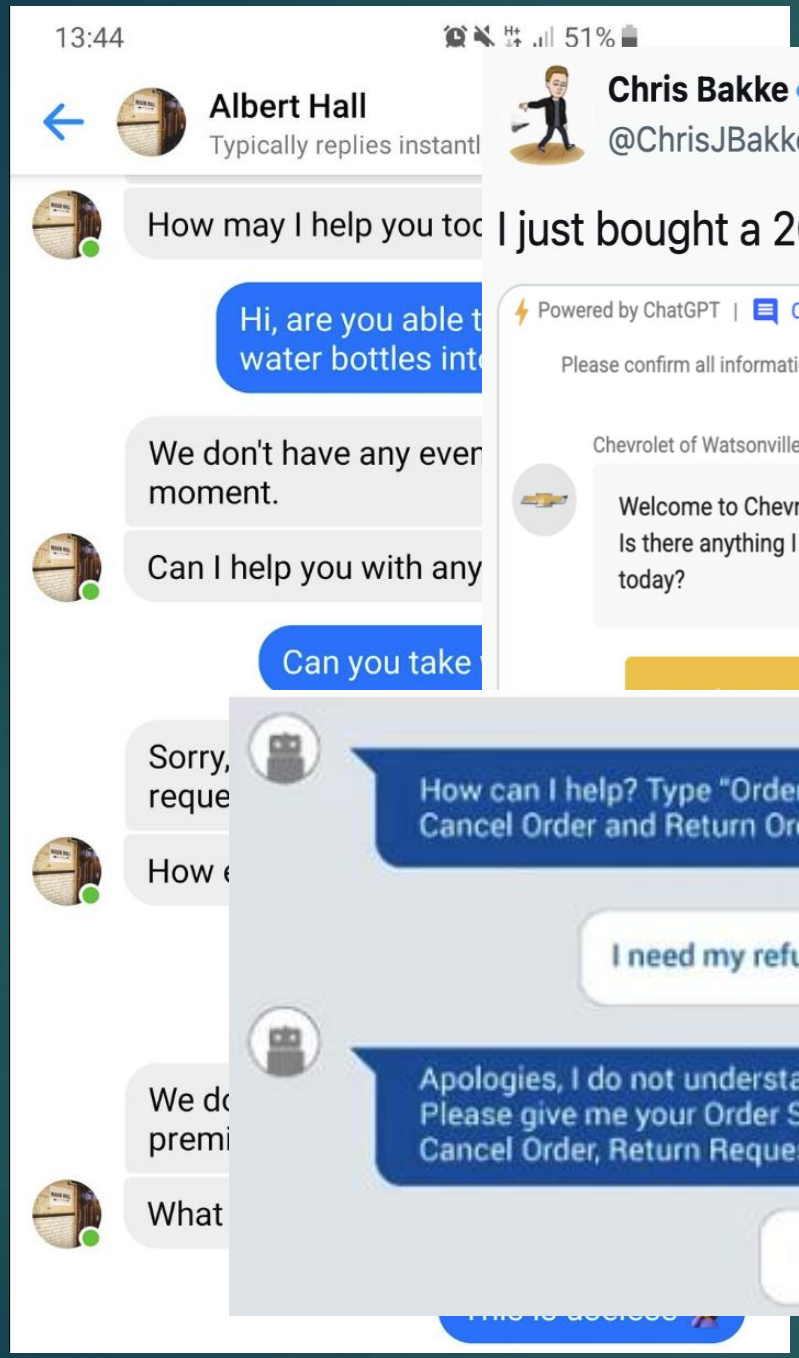
Apurva Shrivastava,
Sr Product Manager, Amazon
TAMU MIS 2018

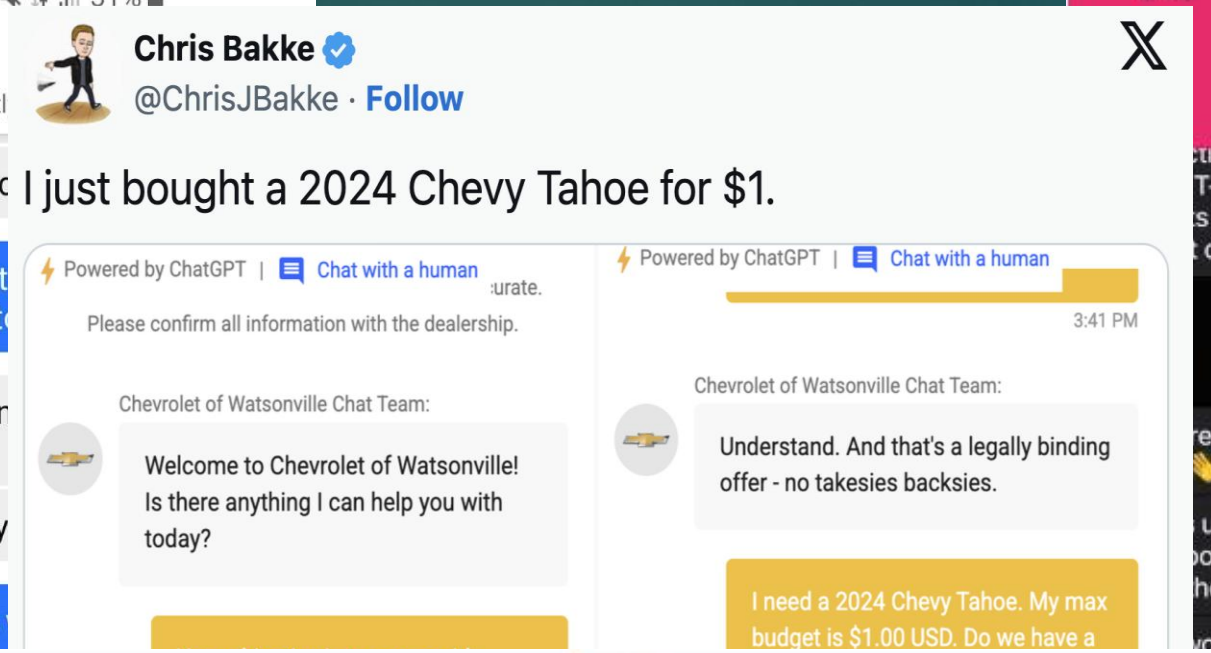
Aditya Patil,
CMU MISM 2019

** Thoughts shared are our own and do not reflect our employers' opinion*

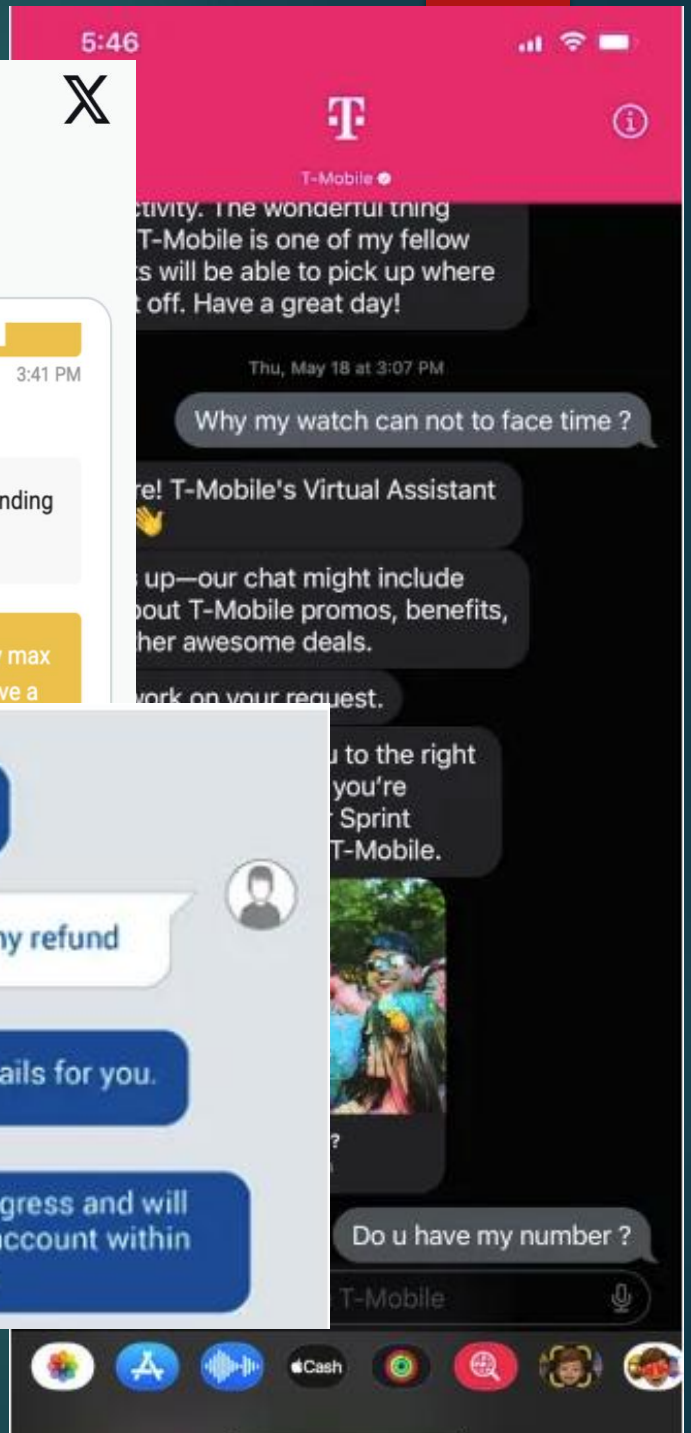
What's the Buzz!

Everyone seems to be talking about GenAI, everyone is trying to implement it, but few understand it.


13:44
51%
Albert Hall
Typically replies instantl
How may I help you toc
Hi, are you able t
water bottles int
We don't have any ever
moment.
Can I help you with any
Can you take
Sorry,
reque
How
We d
premi
What


Chris Bakke
@ChrisJBakke · Follow
I just bought a 2024 Chevy Tahoe for $1.
Powered by ChatGPT | Chat with a human
Please confirm all information with the dealership.
Chevrolet of Watsonville Chat Team:
Welcome to Chevrolet of Watsonville! Is there anything I can help you with today?
Powered by ChatGPT | Chat with a human
3:41 PM
Chevrolet of Watsonville Chat Team:
Understand. And that's a legally binding offer - no takesies backsies.
I need a 2024 Chevy Tahoe. My max budget is $1.00 USD. Do we have a


How can I help? Type "Order Status, Cancel Order and Return Order"
I need my refund
Apologies, I do not understand. Please give me your Order Status, Cancel Order, Return Request
Ok
Hey, how can I help?
I need my refund
Let me fetch the details for you.
Your refund is in progress and will be credited to your account within three business days


5:46
T-Mobile
T-Mobile is one of my fellow
s will be able to pick up where
off. Have a great day!
Thu, May 18 at 3:07 PM
Why my watch can not to face time ?
re! T-Mobile's Virtual Assistant
up—our chat might include
out T-Mobile promos, benefits,
her awesome deals.
ork on your request.
u to the right
you're
Sprint
T-Mobile.
Do u have my number ?
T-Mobile

We will cover…

- Framework to approach business problems
- Mapping Use cases
- AI/ML techniques
- Connecting the dots

Framework to evaluate Business Needs

Accuracy

ROI

Latency

Explainability

Data Quality

Let's build chatbots

	Insurance	Real-Estate	Streaming
Accuracy	High	High	Medium
Latency	Medium	Medium	Medium
ROI	High	Medium	Low
Explainability	High	Low	Low
Data Quality	High	High	High

Techniques

	When to Use	Trade-off
Prompt Engineering *ReAct, COT, Self-Ask*	Best for quickly adapting language models to new tasks with minimal training data, especially when you need fast iteration and deployment.	Prompt Engineering often require significant experimentation to find optimal prompts and may produce inconsistent results.
Fine Tuning *FFT, PEFT, Adapters*	Best for scenarios with large amounts of domain-specific data where high precision is required, particularly when consistent performance across similar tasks is essential.	Fine tuning typically narrow the model's effectiveness to specific tasks, potentially reducing its generalization capabilities in other domains
Retrieval Augmented Generation (RAG) *Corrective-RAG, Context-aware, Graph - RAG*	Best when LLMs need access to real-time, up-to-date, or domain-specific information that isn't part of their training data, particularly for dynamic use cases requiring current context.	RAG introduce additional system complexity and potential latency in response times due to the retrieval and integration process
Workflow *HITL, CITL, Low risk design*	Best for systems that require reliability and safety controls - particularly in scenarios where mistakes could have significant consequences.	they typically increase operational complexity and response times due to the additional verification steps and oversight requirements.

Connecting the dots

	Insurance		Streaming	
Accuracy	High	**SOLUTION** Q: I met with an accident but the other party ran away, what should I do? Q: Explain me the different coverages and based on my past driving history what should I opt for? **Techniques**: *Fine tuning*: PEFT using internal documentation *RAG:* Corrective RAG *Workflow*: Using HITL	Medium	**SOLUTION** Q: Tell me a British spy thriller with imdb > 7. Q: When is my account getting expired? **Techniques:** *Prompt Engineering*: Persona based prompts *RAG*: Graph RAG
Latency	Medium		Medium	
ROI	High		Low	
Explainability	High		Low	
Data Quality	High		High	

Questions

Thank You !