OFFLINE GENERATIVE AI

BRET MEIER

APRIL 8TH, 2025

OVERVIEW

1. **BACKGROUND INFORMATION**
    - ABOUT THE PRESENTER
    - FOUNDATION MODELS AND GEN AI
    - AUTOMOTIVE USE CASES FOR GEN AI
2. **OFFLINE MODELS WHAT, WHY, & WHEN**
    - WHAT IS AN OFFLINE MODEL
    - WHY USE AN OFFLINE MODEL
    - LIMITATIONS OF OFFLINE MODELS
3. **USING AND EVALUATING OFFLINE MODELS**
    - HOW TO CREATE & IMPROVE OFFLINE MODELS
    - MODEL EVALUATION
    - FINAL TIPS & TAKEAWAYS

BACKGROUND INFORMATION

BACKGROUND AND EDUCATION

Lead Data Scientist, Intelligent Products Division at Oshkosh Corporation

Educational History

- BS Mathematics, Physics UW-Eau Claire
- MS Data Science, UW-Oshkosh

Work History

Zywave

- Database Analyst: 1 year

Oshkosh Corporation

- Provisioning Analyst: 6 years
- Data Scientist: 4 years
- Lead Data Scientist: 2 years

FOUNDATION MODELS & GENERATIVE AI

What is a Foundation Model

Foundational Models are large scale machine learning models trained on broad data that generally need further training and adaption to perform specific tasks effectively

"Foundation models are powerful resource-intensive models that have emerged as the dominant paradigm for AI in the 2020s"

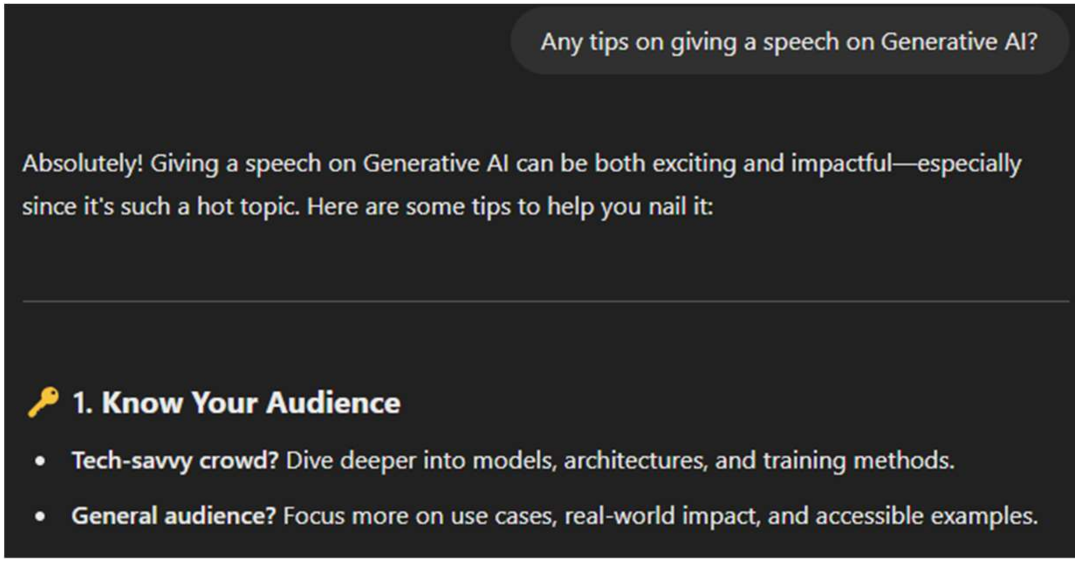
-Stanford Institute for Human-Centered AI

What is Generative AI

Generative AI models are trained to generate text, images and other content

"Generative AI can be thought of as a machine-learning model that is trained to create new data, rather than making a prediction about a specific dataset. A generative AI system is one that learns to generate more objects that look like the data it was trained on."

-MIT News



Any tips on giving a speech on Generative AI?

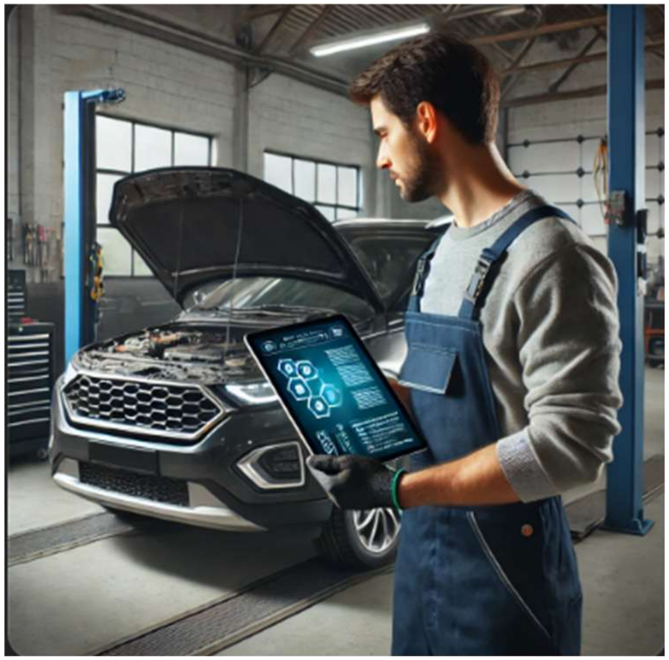
Absolutely! Giving a speech on Generative AI can be both exciting and impactful—especially since it's such a hot topic. Here are some tips to help you nail it:

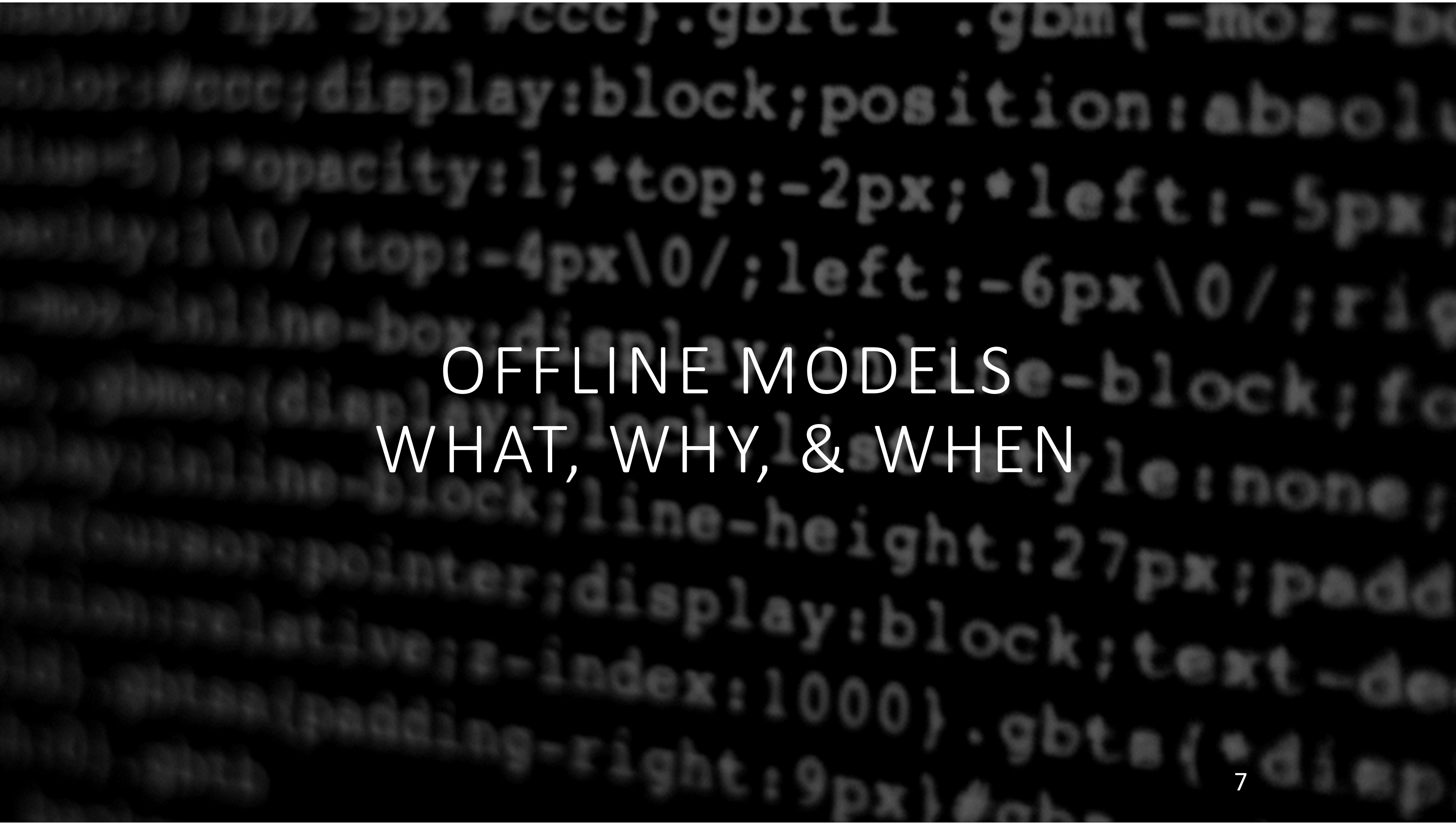
🔑 **1. Know Your Audience**

- **Tech-savvy crowd?** Dive deeper into models, architectures, and training methods.
- **General audience?** Focus more on use cases, real-world impact, and accessible examples.

AUTOMOTIVE USE CASES FOR GENERATIVE AI

1. Predictive Maintenance
2. Optimized Routing
3. Driver Behavior
4. Anomaly Detection
5. Improved Service Efficiency
6. Analyzing Data





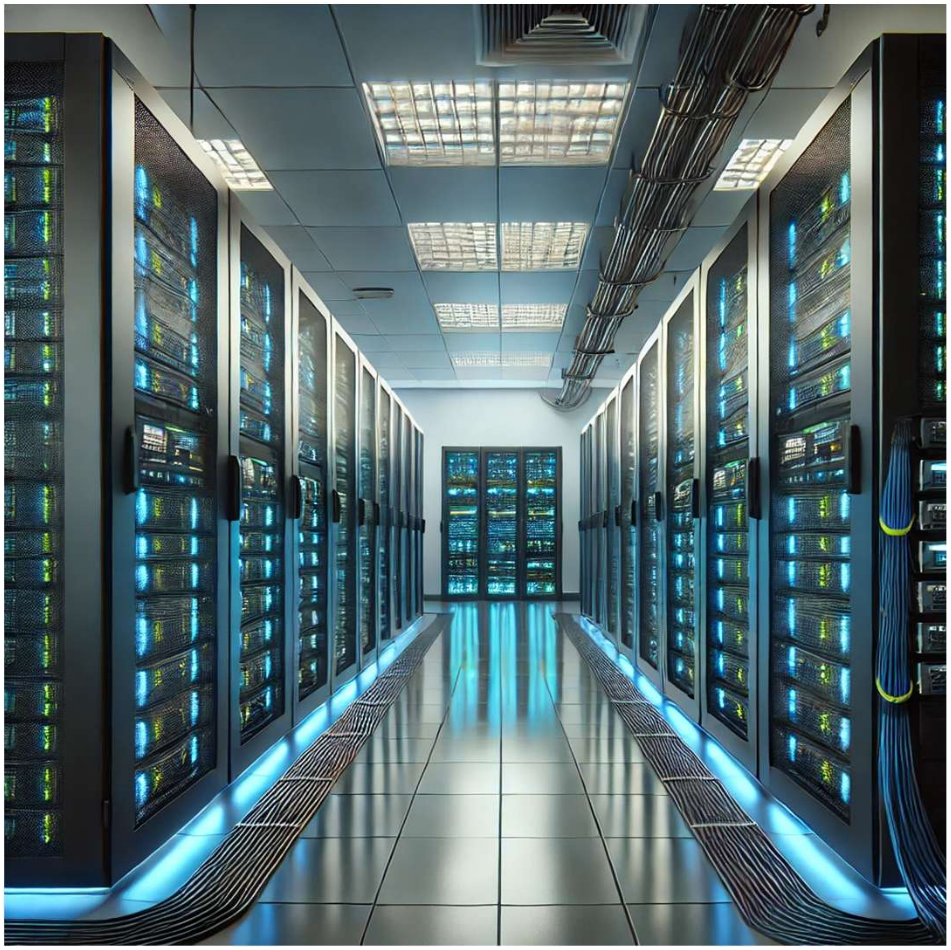
OFFLINE MODELS
WHAT, WHY, & WHEN

WHAT IS AN OFFLINE GEN AI MODEL

An Offline generative AI model is a model that can run without an internet connection. The model is stored and runs locally on controlled hardware

Examples of controlled hardware

- Laptop
- Tablet
- Local server system

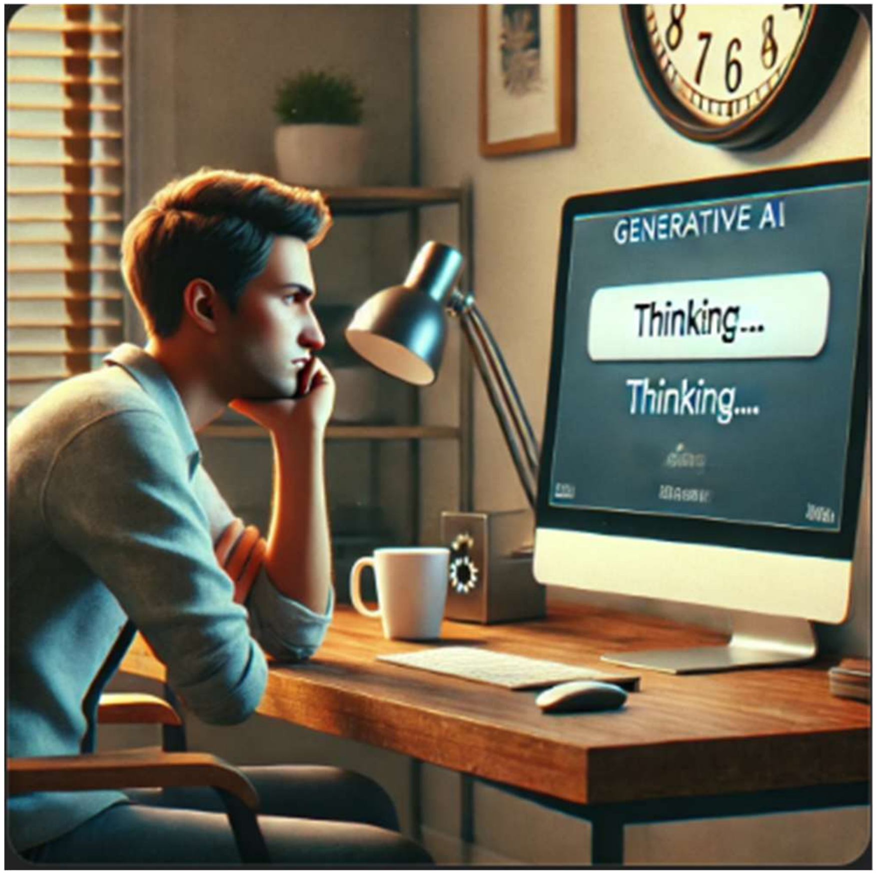


WHY USE AN OFFLINE MODEL

1. Expanded access and use cases for gen AI,
2. can put offline models on edge devices such as on-board a vehicle or on a phone
3. Improve data security
4. Reduce data leakage
5. Easier data compliance
6. Improved data privacy
7. Poor Internet connection or throttle concerns
8. Complete control over data you own
9. Control over the model outputs

LIMITATIONS OF OFFLINE MODELS

- Model is limited by hardware
- Slower response times
- Less Robust Models
- Less accurate responses
- Harder to maintain and update model



USING AND EVALUATING OFFLINE MODELS

HOW TO CREATE & IMPROVE OFFLINE MODELS

Get a Foundational Model

1. Can download many free versions of foundational models
    - https://ollama.com/library?sort=newest
2. Select model and number of parameters

Training methods

1. Active Learning: model can flag unsure responses for review
2. Reinforcement Learning: model receives feedback from user interactions. This seeks to maximize an objective function
3. Fine tuning: Feed the model a prepared dataset.

Improve performance

1. Prompt engineering can improve response and accuracy times
2. How you upload datasets for analysis can affect the model accuracy and response time

MODEL EVALUATION

Data types

- PDF
- Tabular

Evaluation Criteria

- Response time
- Accuracy/completion of answer

Models Evaluated

- Llama 3.2 (3b, 7b, & 70b parameter versions)
- OpenAI

Experimental Factors

- Size of data uploaded
- Data uploading methods
- Sparsity of tabular data

FINAL TIPS & TAKEAWAYS

Pros

- Offline Generative AI models can be used to expand access and create more use cases for Gen AI
- Offline models provide strong data control
- Offline models make data and regulatory compliance easier

Cons

- Offline models have worse performance
- Offline models require good hardware
- Offline models are less robust and may require more preprocessing for model functionality
- Only practical for smaller, specific tasks or to comply with laws and regulations

Cautions

- There are open Patent and IP questions without clear answers
- Trust your foundational model before downloading
- The data economy landscape is changing rapidly

THANK YOU

Bret Meier

bmeier@oshkoshcorp.com