midwest
architecture community
collaboration

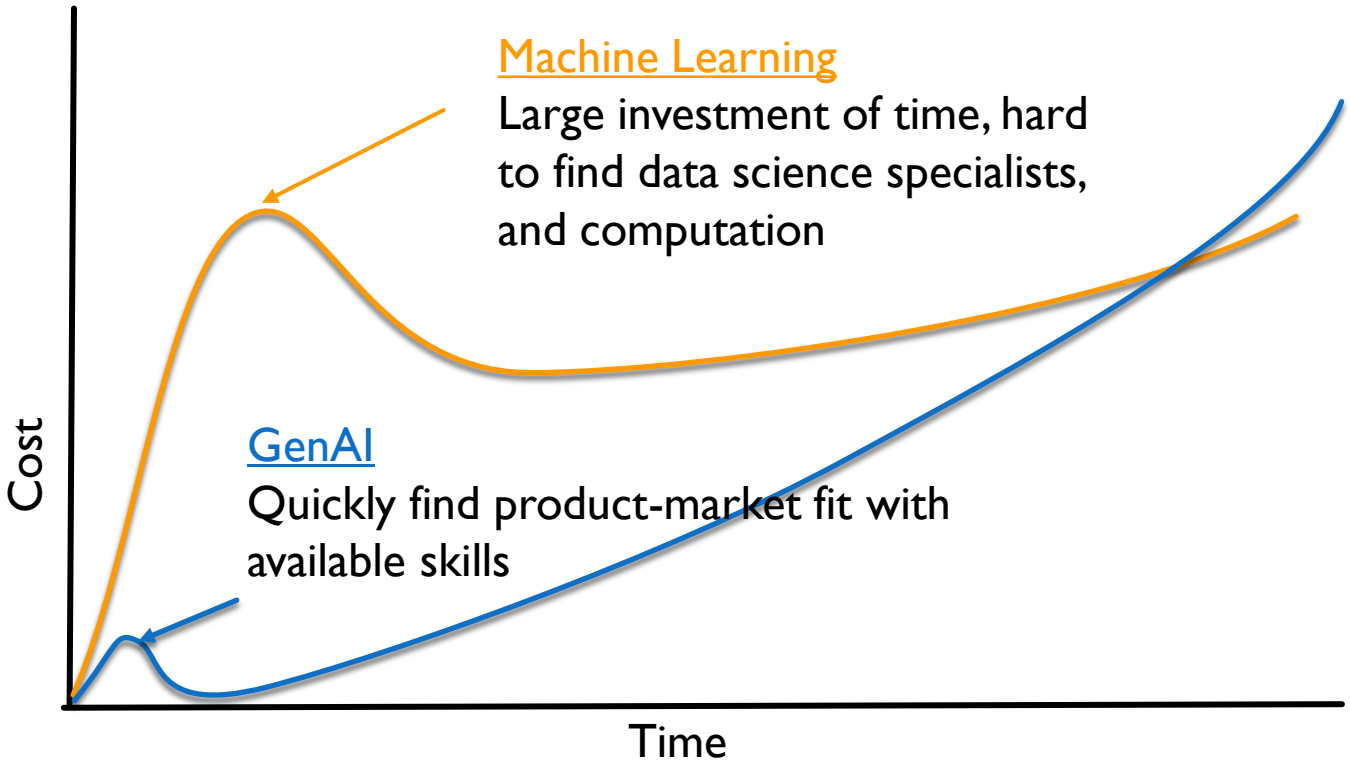
BEYOND THE HYPE

NOVEMBER 8, 2023

USING GENERATIVE AI TO SOLVE REAL WORLD PROBLEMS

MIKE GILLESPIE

WHY DID GENERATIVE AI GET SO MUCH HYPE?



pixar, brown and white Australian Shepherd, brown eyes

midwest architecture community collaboration

WARNING: NOT ALL PROBLEMS ARE GEN AI PROBLEMS

Large Language Models are versatile – they can solve many problems.

Great way to get started, but often are often outperformed by smaller, purpose build models.

Ask: What's being generated? Why is an LLM the right tool for the job?

Results are not deterministic. Can your application tolerate variance?

watercolor, modern, clouds, landscape

midwest architecture community collaboration

FRAMEWORK FOR GETTING MOST FROM GEN AI

Look across your organization. Where do you spend time?

Look at the expenses in your organization and categorize them by type of work.

Where are highly-trained people doing repetitive 'mindless' work.

When is mediocre 'good enough'?

night sky, camera, tripod, impressionist

#1 GROWING PAIN: STARTING WITH A TOOL INSTEAD OF PROBLEM!

midwest architecture community collaboration

FRAMEWORK FOR GETTING MOST FROM GEN AI

Do you have staff that requires light-moderate training to get up to speed. Especially seasonal clerical work.

Focus on enhancing productivity, not complete elimination.

What tasks are the bottlenecks in a process?

Data lookup, synthesis, and summarization.

graphic productive employee increasing

CONSUMING AI SERVICES

INTEGRATING GEN AI SERVICES INTO YOUR ORGANIZATION

midwest architecture community collaboration

100 LEVEL – GENERAL PURPOSE CHATBOTS

Using a general purpose chatbot:

- Create code snippets
- Natural language web search
- Writing a summary of a document
- Writing a term paper
- Many More!

monochromatic, greece, acropolis, sunset

midwest architecture community collaboration

100 LEVEL – GENERAL PURPOSE CHATBOTS

Concerns:

- Hallucinations – Don't use for legal briefs
- Doesn't include your organization's data
- Data Security / Data Classification
- Intellectual Property

concern, warning, caution, graphic

midwest architecture community collaboration

200 LEVEL – PURPOSE-BUILT TOOLS

Purpose built tools are generative AI tools designed and tuned for a specific purpose. These are mass consumption tools for common tasks like:

- Coding Assistants integrated into an IDE
- Email / Document analysis
- Chatbot interface to existing knowledge bases
- Natural language search on a document repository
- Natural language business intelligence platforms
- Product description from image

cloud computing, meetings, laptop

THESE TOOLS MAKE AI ACCESSIBLE TO EVERYONE

midwest architecture community collaboration

EVALUATING GEN AI SERVICES

Evaluation Criteria

- **Security**: with whom is the data being shared?
- **Cost**: How are services being charged? Per user? Unit of consumption? Is the cost predictable?
- **Accuracy**: How is the vendor ensuring accuracy of the responses?
- **Intellectual Property**: Was the model build on copyrighted material? If no, what exposure do we have using this to generate content?

software AI evaluation graphic

midwest architecture community collaboration

BUILDING AI POWERED PRODUCTS

BUILDING GEN AI SERVICES WITH YOUR DATA ASSETS

midwest architecture community collaboration

300 LEVEL – CUSTOMIZE GEN AI TOOLS

TOOLS DEVELOPED WITH ACCESS YOUR ORGANIZATION'S DATA

Building Gen AI enabled applications

- Integrate your data and APIs into Gen AI
- Fine tuning / model building is not required. Prompting can go a long way.
- Start with larger models, find the smallest model that meets the need.
- Keep security in mind. Would you let an intern answer this question?
- Measure how much this will cost on each model

builder craftsman tools artisan graphic

midwest architecture community collaboration

300 LEVEL – CUSTOMIZE GEN AI TOOLS
TOOLS DEVELOPED WITH ACCESS YOUR ORGANIZATION'S DATA

Customized Chatbot – Bank Chat Case Study

- Add context to the user's account: balance, recent transactions, profile data
- Allow customer to ask about their account.
- Only answer questions the from the customer profile.
- Be wary of **prompt injection**. How do you prevent data leakage? Giving recipe to napalm?

bank virtual teller chat

midwest architecture community collaboration

300 LEVEL – CUSTOMIZE GEN AI TOOLS

TOOLS DEVELOPED WITH ACCESS YOUR ORGANIZATION'S DATA

Case: Contact center call summarization

- Transcribe the call audio, or collect the text of the chat.
- Prompt the LLM to create a summary of the call, capturing sentiment and any action items.
- Person in the middle – the agent will edit / correct the transcript.
- **High value**: Contact center agents can spend 10-30% of their time summarizing and dispositioning contacts.

contact center agent typing, bored

midwest
architecture community
collaboration

300 LEVEL – CUSTOMIZE GEN AI TOOLS
TOOLS DEVELOPED WITH ACCESS YOUR ORGANIZATION'S DATA

Case: Car repair knowledge base

- Car repairs are complex and require multiple steps from multiple manuals.
- Load repairs into a knowledge base.
- Include vehicle make / model / description of serving
- Synthesis specific repair from different manually, looking at next steps based on make/model

high tech auto repair

midwest architecture community collaboration

300 LEVEL – CUSTOMIZE GEN AI TOOLS
TOOLS DEVELOPED WITH ACCESS YOUR ORGANIZATION'S DATA

Smart integration for SaaS

- Integration API for a SaaS provider.
- Provide a text interface to what the customer would like to do: "Provide a CSV export for the last month of sales data, written in java"
- Generates source code for the integration.
- Customer can use the code to interact with the SaaS.

software developer relaxing on a beach cartoon

300 LEVEL – CUSTOMIZE GEN AI TOOLS
TOOLS DEVELOPED WITH ACCESS YOUR ORGANIZATION'S DATA

Clinical Trials R & D:

- Large volumes of clinical data in research papers.
- Searching and summarization of research
- Report results adhering FDA guidelines.
- Find similar studies, compare conclusions

DNA scientist helix graphic

midwest architecture community collaboration

300 LEVEL – CUSTOMIZE GEN AI TOOLS
TOOLS DEVELOPED WITH ACCESS YOUR ORGANIZATION'S DATA

Contact agent "Next best action":

- Get a live transcript of the call.
- Have a knowledge base of the scripts, standard operating procedures, and escalations
- Provide the contact history and profile in the context.
- Provide 3 "Next best actions" to the agent.

contact center agent checklist ipad cartoon

midwest architecture community collaboration

ADVANCED GEN AI PRODUCTS

NATURAL LANGUAGE IS THE NEW JAVA

midwest architecture community collaboration

400 LEVEL – ADVANCED GEN AI TOOLS
USING REASONING AND ACTIONS FROM LLMS

Solve problems on behalf of users

- Provide context with information on how to interpret data, standard operating procedures
- Provide data specific to the scenario and ask how to best resolve
- Develop a plan on next steps, and execute those actions.
- Iteratively interpret the result of the action and determine next steps.

advanced AI tools graphic

midwest architecture community collaboration

400 LEVEL – ADVANCED GEN AI TOOLS
USING REASONING AND ACTIONS FROM LLMS

Case study: Cloud Financial Management

- Library of CFM best practices
- Tools to query an account's billing and consumption data
- Start with "Create a 5 step plan on reducing cloud costs"
- Taking each of the 5 steps, use billing data to provide additional context on action items. What specific actions need to happen.

cloud finance guru graphic

midwest architecture community collaboration

400 LEVEL – ADVANCED GEN AI TOOLS
USING REASONING AND ACTIONS FROM LLMS

Case study: Threat Detection Incident Response

- Library of security and TDIR best practices
- Access to security tools and logs
- Analyze a threat: Are we under a ransomware attack?
- Propose remediations. Write python code that will execute the containment and remediation.
- Execute the proposed steps?

IT security expert graphic

midwest architecture community collaboration

KEY CAPABILITIES
SKILLS NEEDED TO IMPROVE SUCCESS

Key skills to develop:

- Prompt engineering: asking the right way makes all the difference
- LLMs: Knowing which LLMs are best for each use case. Landscape is changing quickly.
- Model evaluation: How do you know the output is good? User based feedback loops and automated quality assessment.

SLR Camera, waterfall, lake superior, northwoods

midwest architecture community collaboration

KEY CAPABILITIES
SKILLS NEEDED TO IMPROVE SUCCESS

Key skills to develop:

- Fine Tuning: Improving the quality of the output of your prompts
- Optimization: Finding the balance of model selection, fine tuning, and prompt optimization

SLR Camera, waterfall, lake superior, northwoods

midwest architecture community collaboration

WHICH IS ‘REAL’?

lake, fog, northern lights. night

midwest architecture community collaboration

GETTING STARTED

THIS IS ALL NEW TO ME

midwest architecture community collaboration

GETTING STARTED
GET FAMILIAR WITH GEN AI

Step 1: Get your hands dirty!

- Don't start by reading books on neural networks and LLMs.
- Find a real problem you think GenAI can help.
- Get familiar with prompting an LLM. Go beyond signing up for a chat, and work with the API.
- Jupyter Notebooks are great for getting started.
- Practice writing prompts. See how each model differs in prompt structure.

starting line race running graphic

midwest architecture community collaboration

GETTING STARTED
GET FAMILIAR WITH GEN AI

Step 2: Build something

- Have a hypothesis. "We can solve problem 'x' with an LLM
- Provide context data into the prompt. You can use copy and paste to start, you are just proving a hypothesis. Fail fast!
- Create a metric for model output. Initially it can be human scored, but for scale, you will need to automation model quality assessment
- Get feedback from SMEs

fail fast graphic

GETTING STARTED
GET FAMILIAR WITH GEN AI

Step 3: Iterate and scale

- Iterate prompt and data until you can validate if the hypothesis is correct
- Iterate and select the smallest model that generates acceptable-quality results.
- Incorporate user feedback into the model evaluation.
- Have a person in the middle approach to start. Successful applications don't have to be zero touch. If you can accomplish a task in half the time, that's a big win.

iterate and scale graphic

midwest architecture community collaboration

QUESTIONS AND ANSWERS

STUMP THE CHUMP

midwest architecture community collaboration

THANK YOU!

MIKE GILLESPIE