Voice-Activated AI Collaborators

A Hands-On Guide Using LLMs in IoT & Edge Devices

David vonThenen
Software Engr/Dev Advocate @ Deepgram
@dvonthenen

SCALE
Pasadena, CA
Mar 14-17, 2024

Deepgram

David vonThenen

- Are you Human or an AI?
- I want 5 Kubernetes
- Virtual Machines are Real
- Replacing Myself with Bots…
- Cloudy, cloudy, cloudy…
- There is storage for that!

[@dvonthenen](https://github.com/dvonthenen)

Agenda

- Voice-Activated AI Collaborators
- IoT and Edge Design Considerations
  - Constrained Resources on Devices
  - Speech, LLM and Model Considerations
- Building Your Device!
- Demo
- Q&A

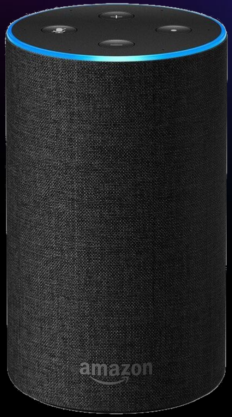
Voice-Activated AI Collaborators

SCaLE

Pasadena, CA
Mar 14-17, 2024

Deepgram

Voice-Activated AI Devices



Future Devices

Components of a Voice Assistant

STEP 1
Automatic Speech Recognition

STEP 2
Natural Language Processing

STEP 3
Desired business logic via hooks

STEP 4
Text to Speech

Voice Assistant Summary

- **Understanding the Voice Audio Data**
    - Speech-to-Text
    - Microphone, Vibration (The Thing), etc
- **Comprehending Context and Meaning**
    - Complex Business Logic
    - Large Language Models
    - Domain Specific Models
- **Producing an Audible Output**
    - Text-to-Speech
    - Speaker, Bone Conduction Device, etc

IoT + Edge Considerations

Constrained Resources

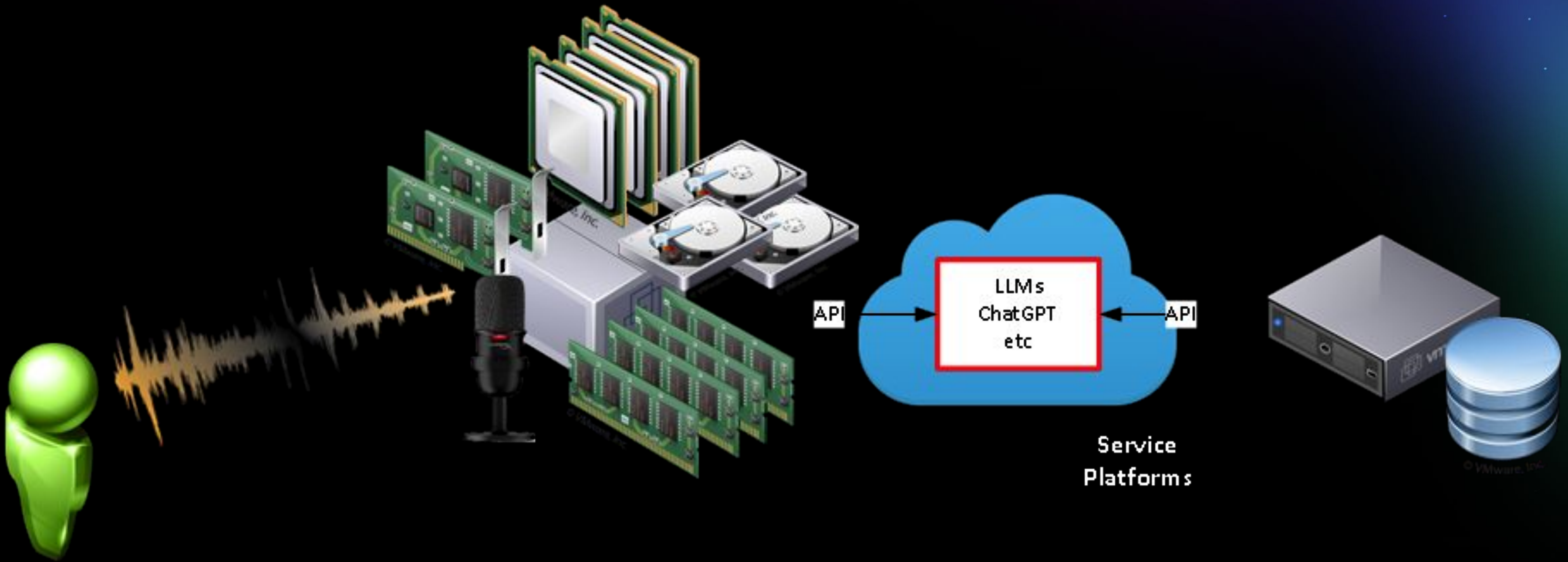
Resources on Device

- **Building an Edge Device Means:**
  - CPU Constrained
  - Memory Constrained
  - Graphic/AI Accelerator
  - Storage:
    - What Data Can We Bring?
    - How Fast to Access Storage?
  - Network: Access to the Internet?
- **What Architecture is Best? Depends on...**
  - Use Case - Scenario to Build For
  - Device - Attributes Required
  - Cost - 🤑🤑🤑

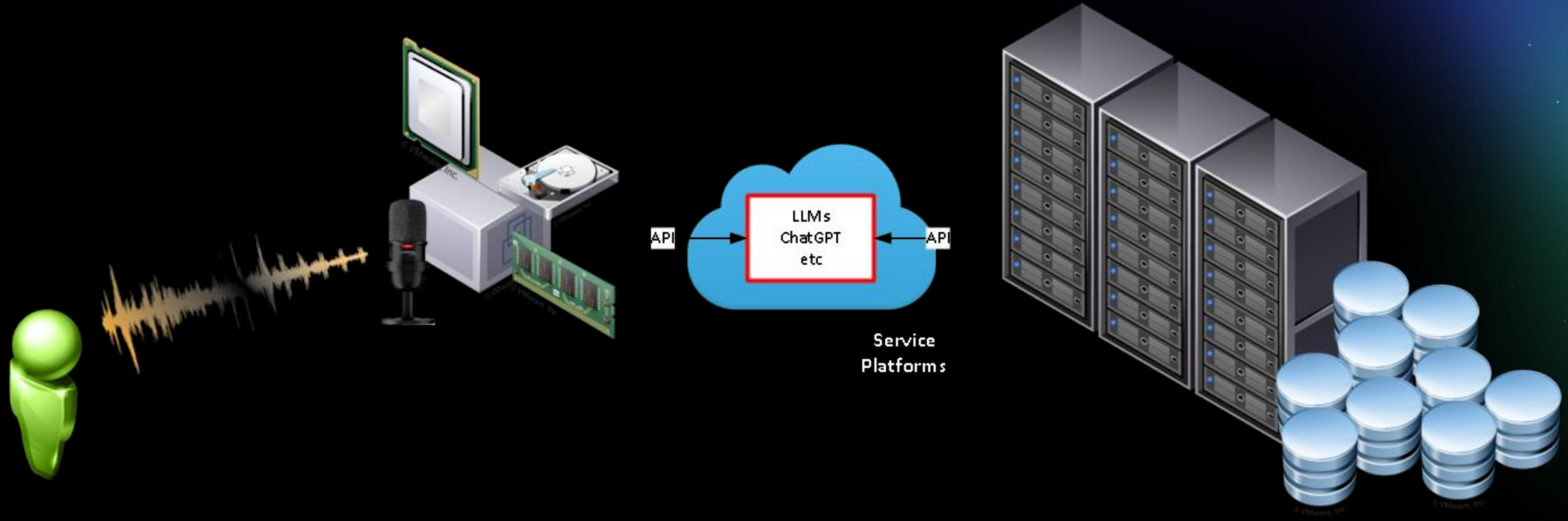
Use Cases and Goals

- **Understand the Goals:**
  - What Problem Are You Trying to Solve?
  - Are There Multi-Modal Considerations?
    - Multi Audio Channels or Streams
    - Images, Video, Data, Code, etc
  - Need to Recall History?
  - How “Good” is Good?
  - Minimum Viable Product (No Really)
  - Achieve Balance in Your Solution
- **Ok, Show Me Some Edge Devices!**

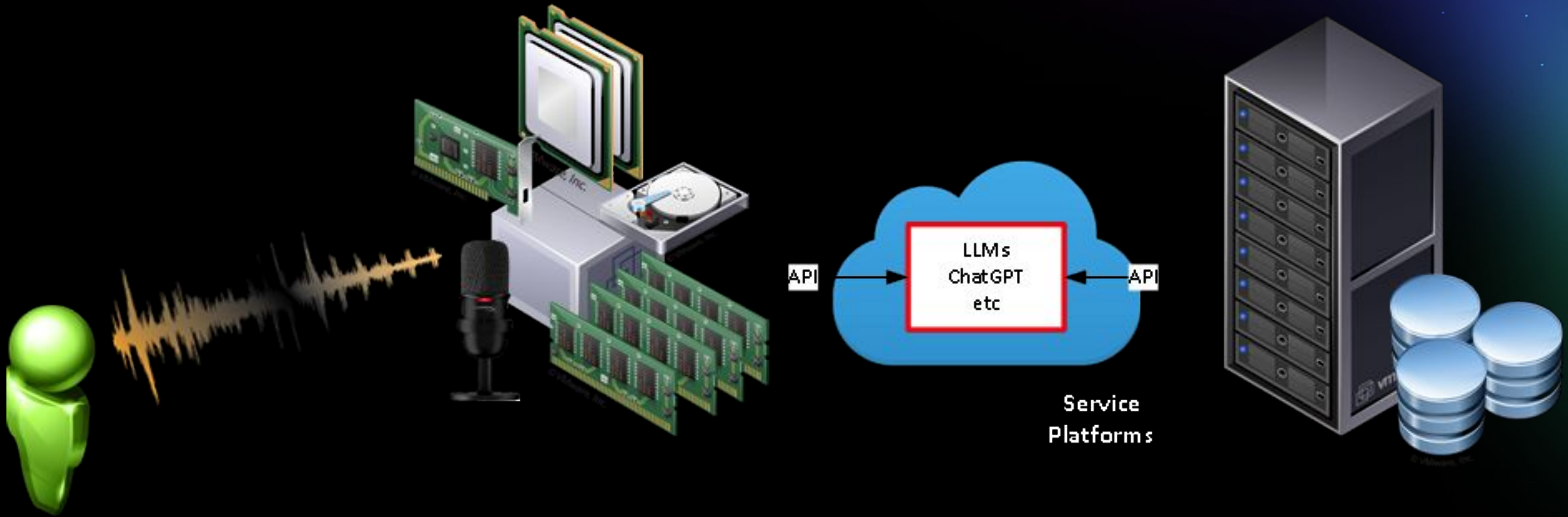
High Performance Edge Device



Lightweight Edge Device



Achieve Harmony



Device Design Recap

- **Lightweight Device**
  - Lower Cost Devices
  - “Collector” (ie Cloud Processing)
- **High Performance Device**
  - Need to Process Data at the Edge
  - Too Much Data (ex, Video Streams)
  - Lower Latency, Quicker Response
- **Hybrid Device**
  - Determine Processing Bottlenecks
  - Profile Components (Tracing + Metrics)
  - Optimize for Your Use Case

IoT + Edge Considerations

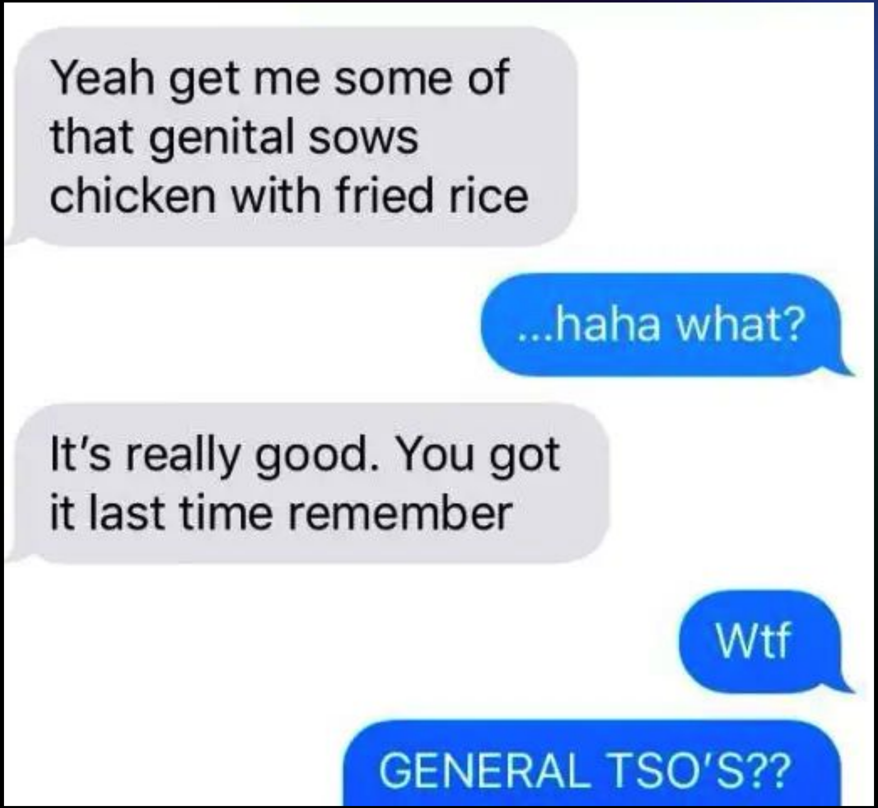
Speech, LLM and Model Considerations

Holding a Conversation

- **AI Voice Assistant MVP**
  - You Speak, It Replies
  - It Does Something Meaningful
- **Transcription Importance for Devices**
  - Errors, Not Critical+Retry
  - Errors, Impacts Results+Not Critical
  - High-Confidence, Impact Functionality
- **What Do You Want to Focus On?**
  - STT/TTS is Core to your Business?
  - Transcription Accuracy → Downstream Effects
  - Latency of Audio Transcription
  - And More...

Conversation Understanding

- Word Error Rate (WER)
- Ratio: Errors in Total Words (lower better)
  - 20% WER Means the Transcript is 80% Accurate
- Accuracy matters!
  - Correctness!
  - Context!
- Forgiving vs Critical
  - Jokes aside...



Yeah get me some of that genital sows chicken with fried rice

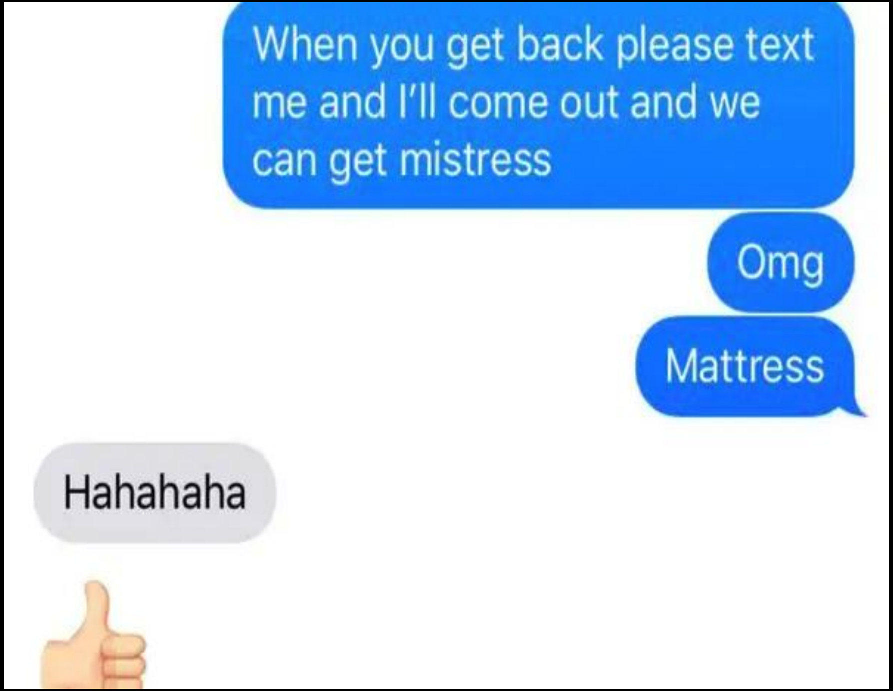
...haha what?

It's really good. You got it last time remember

Wtf

GENERAL TSO'S??

More Funnies...

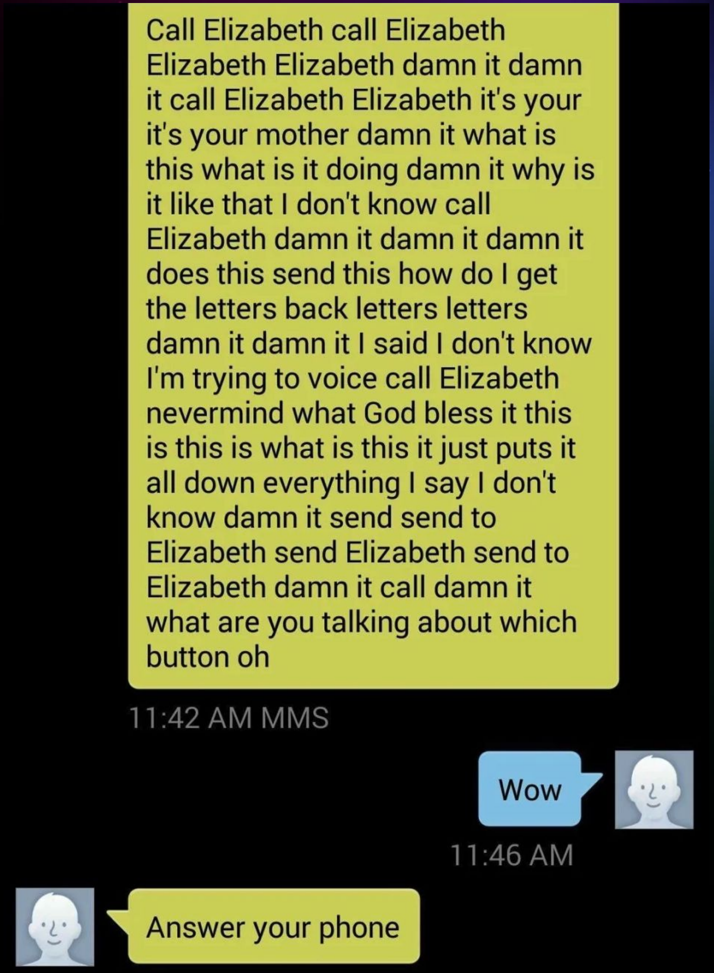


When you get back please text me and I'll come out and we can get mistress

Omg

Mattress

Hahahaha



Call Elizabeth call Elizabeth Elizabeth Elizabeth damn it damn it call Elizabeth Elizabeth it's your it's your mother damn it what is this what is it doing damn it why is it like that I don't know call Elizabeth damn it damn it damn it does this send this how do I get the letters back letters letters damn it damn it I said I don't know I'm trying to voice call Elizabeth nevermind what God bless it this is this is what is this it just puts it all down everything I say I don't know damn it send send to Elizabeth send Elizabeth send to Elizabeth damn it call damn it what are you talking about which button oh

11:42 AM MMS

Wow

11:46 AM

Answer your phone

1: Funny: https://www.buzzfeed.com/farrahpenn/voice-text-funny-mistakes

2: Funny: https://finance.yahoo.com/news/cant-stop-laughing-voice-text-223642046.html

Example: Conversation (Mis)Understandings

Summary on Actual Transcript:

The conversation is about vitamin D overdose and the potential risks associated with it. The topic is discussed, and the speakers mention that people may not realize the potential benefits of vitamin D and that it can cause nerve damage. The dangerous effects of vitamin D on health are also discussed, and the speakers suggest that people should be better informed by their current medications.

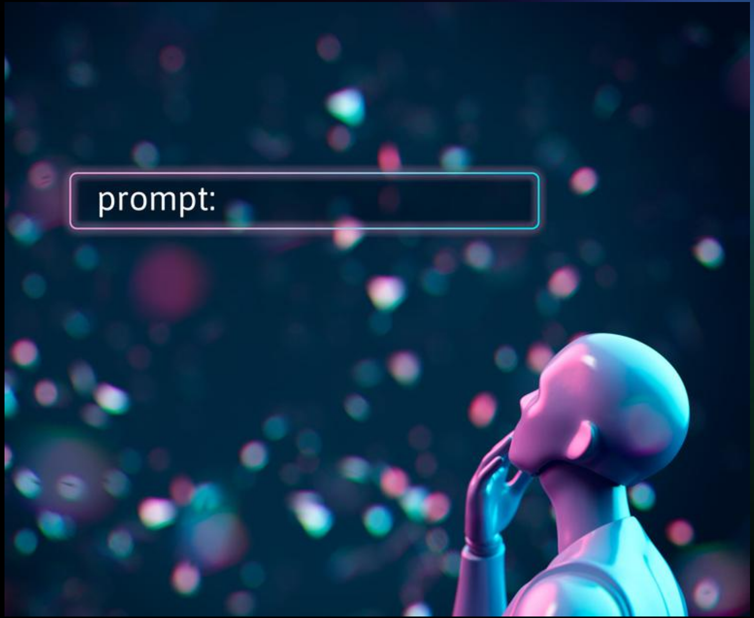
Summary from an ASR Transcription:

The host of a podcast discusses a man who was hospitalized with vomiting abdominal pain leg cramps and lost 13 kg. The host also mentions a woman who mistakenly thought she had multiple sclerosis and lost her vision. The podcast discusses the rising popularity of vitamin and the issue of pharmacist safety. The host also points out that people need to be better informed about safety and the need for more training for pharmacists.

Source Material by Sam Hardy:
Blog: Poor mans ASR pt. 1 and Poor mans ASR pt. 2
GitHub: https://github.com/samhardyhey/blog-os-asr

Accuracy with Prompts to LLMs

- **If the Answer Matters, This is Problematic Because:**
  - Voice Transcription is Your Prompt
  - Misinterpreted Data/Info
- **LLMs: Next Token Prediction Engines**
  - Prompt and Data = Words = Tokens
  - Higher WER Means Poor Response
- **Downstream Effects:**
  - Prompt Doesn’t Capture Intent
  - Inability to Harvest Language Data



[1] Ana Olssen - Accuracy Matters When Using GPT-4 and ChatGPT for Downstream Tasks

Example: Prompt Accuracy

Actual Transcription:

Q: Are there any specific vitamins that are being called in this article?

A: Yes, the specific vitamins mentioned in the article are vitamin D and vitamin B6.

Transcription from a Cloud Platform:

Q: Are there any specific vitamins that are being called in this article?

A: Yes, the specific vitamin mentioned in the article is vitamin B6.

Source Material by Sam Hardy:
Blog: Poor mans ASR pt. 1 and Poor mans ASR pt. 2
GitHub: https://github.com/samhardyhey/blog-os-asr

Example: Prompt Accuracy

Actual Transcription:

Q: Was there any mention of what could have caused the uptick in these vitamin overdoses?

A: ..was partly attributed to people believing it could be a preventative measure against Covid... individuals might consume multiple supplements and foods containing this vitamin..

Transcription from a Cloud Platform:

Q: Was there any mention of what could have caused the uptick in these vitamin overdoses?

A: ...attributed to people taking multiple supplements from different sources without realizing the cumulative doses...

Source Material by Sam Hardy:
Blog: Poor mans ASR pt. 1 and Poor mans ASR pt. 2
GitHub: https://github.com/samhardyhey/blog-os-asr

Example: Prompt Accuracy

Actual Transcription:

Q: Was there mention of diet being a contributor?

A: The article explicitly states that the chance of getting a B6 overdose from diet alone was marginal.

Transcription from a Cloud Platform:

Q: Was there mention of diet being a contributor?

A: The article does not specifically mention diet as a contributor to the vitamin overdoses discussed.

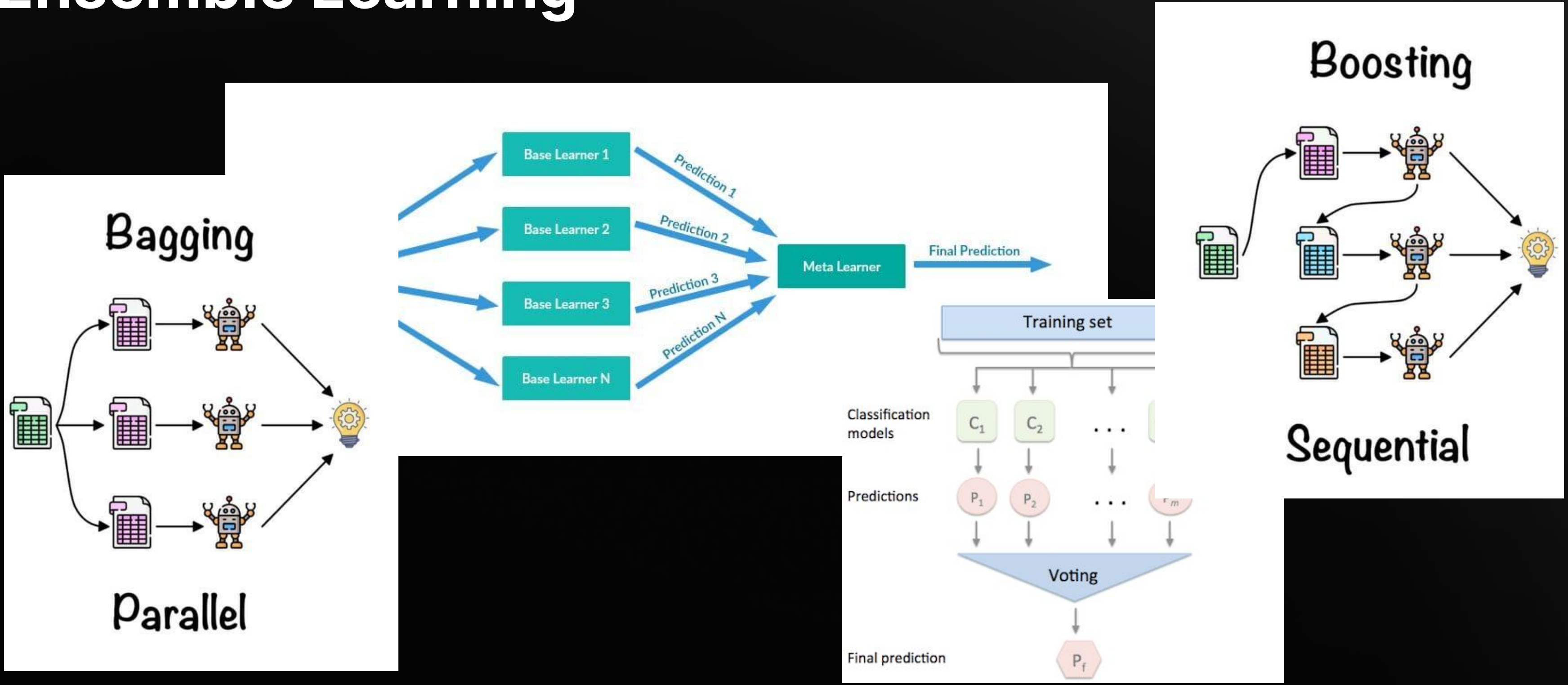
Source Material by Sam Hardy:
Blog: Poor mans ASR pt. 1 and Poor mans ASR pt. 2
GitHub: https://github.com/samhardyhey/blog-os-asr

Run-Time Resource Contention

- GPU vs CPU (hardware dependant) - ~4 Times as Fast Ref 1
- Resource Contention between AI models
  - CPU: “It is what it is”
  - Just In Time Models (Dynamically Load)
  - GPU Time Slicing (Serial)
  - GPU Partitioning (ex MIG) Ref 2
- Models as “Services”
  - Numerous requests a second
  - Ensemble Learning/Combining Ref 3
  - Solving Complex Problems Ref 4

1: Benchmarks: https://hackaday.com/2017/03/14/hands-on-nvidia-jetson-tx2-fast-processing-for-embedded-devices/
2: Partitioning: https://developer.nvidia.com/blog/running-multiple-applications-on-the-same-edge-devices/
3:Ensemble: https://medium.com/illumination/ensemble-learning-learn-the-power-of-multiple-models-for-enhanced-predictions-980b87b5f43c
4: Combining: https://machinelearningmastery.com/multiple-model-machine-learning/

Ensemble Learning



1: https://www.kdnuggets.com/2022/10/ensemble-learning-examples.html
2: https://scikit-learn.org/stable/modules/ensemble.html
3: https://machinelearningmastery.com/ensemble-machine-learning-with-python-7-day-mini-course/

Building Your Device

Coming to a Final Design

Complex Problems…

- **…Require Complex Solutions**
- **Voice Assistants Require 2 Models Min**
  - Speech-to-Text
  - (Optional, but Required) LLM
  - Text-to-Speech
- **What is Core to Your Business?**
  - Use Cases, Value Proposition, etc
  - Data Locality Stickiness
  - Quantity of Data to Collect vs Send
  - On-Device vs Offload
  - What Problem are Your Solving?

Limited vs Limitless

- **On-Device vs Cloud Resources**
  - Complex Models May Require More Resource
  - Smaller Models Miss Details/Nuance
- **Trading Off Model Run-Time Requirements**
  - Accuracy: How True is True? Ref 1
  - Speed: Ahead Full!
  - Nuance and Misidentifying
  - Edge Cases and Outliers
- **Understand Your Trade-Offs!**
  - Is This Good Enough?

1:Truth: https://www.linkedin.com/pulse/ai-struggles-detect-false-information-because-finding-vanessa-otero/

Example Trade Offs

- **Example 1: Text-to-Speech**
  - Data: “Hello, how are you doing?”
  - Send: 25 bytes + Transport Overhead
  - Receive: Linear16, mp3, etc Audio Stream
  - Maybe a Good Candidate to Offload
- **Example 2: (High Definition) Video Stream**
  - Data Stream for Video Camera = Large
  - Send: Various Answers (~10MB/Second)
  - Receive: Image with Identified Information
  - Maybe Not a Good Idea!
- **Different Use Cases w/ Different Considerations**
  - Understand the Problem and Trade Offs

Demo

Demo Components

[CLICK HERE] for All Material Contained in this Session

Speech-To-Text

- (Offline) PocketSphinx - https://github.com/cmusphinx/pocketsphinx
- (Internet) Deepgram - https://github.com/deepgram/deepgram-python-sdk

LLMs:

- (Offline) Falcon - https://huggingface.co/tiiuae/falcon-7b
- (Offline) OpenLlama - https://huggingface.co/openlm-research/open_llama_3b
- (Offline) Llama 2 - https://huggingface.co/TheBloke/Llama-2-7B-Chat-GGUF/tree/main

Text-to-Speech

- (Offline) pyttx3 - https://github.com/nateshmbhat/pyttsx3
- (Internet) Deepgram - https://github.com/deepgram/deepgram-python-sdk

Resources

[CLICK HERE] for All Material Contained in this Session

- Voice AI Demo
- Poor Transcription → Summarization Errors
- Poor Transcription → LLM Prompt Issues
- Example PocketSphinx App
- Example pyttsx3 App
- Additional Reading Materials

Thank you

David vonThenen
Software Engineer/Developer Advocate
@dvonthenen

Pasadena, CA
Mar 14-17, 2024