

Workshop

NLP using
Generative Models:
What's Next?

Speaker

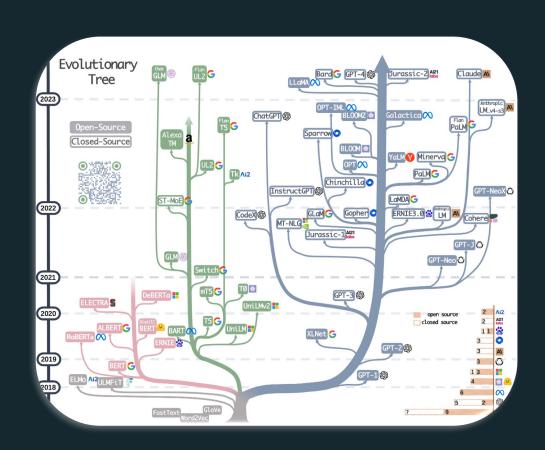
Raghav Bali Staff Data Scientist



Current State: Model Universe



- Huge improvements from the days of word2vec to current LLMs
- Extremely expensive to train and research (few key players)
- Far more accessible today than a few months ago
- Breakneck speed of innovation



• What is 2 + 2?

• Which object is smaller?



Complete the phrase "This Saturday I am"

Language Models



Pretrained Language Models

- Huge Datasets
- Very Large Compute
- Predict Next token

GPT, LLama, PaLM

Supervised Finetuned Models

- Small Dataset
- Medium Compute
- Predict Next token

Vicuna, MPT, ChatGLM

RLHF Models/Agents

- Smaller Dataset
- Medium Compute
- Predict Next token & Maximize Reward

chatGPT, Bard, Claude

- Extremely good at predicting the next token
- Prompts help structure the entropy for specific tasks

System 1 Characteristics







Error Prone

Fast

Automatic

Systems 2 Thinking



Write a blog post about LLM Landscape

- Research and list key areas in the LLM landscape
- Create an outline with key areas as headings
- Research and read papers and write understanding of each topic
- Rephrase paragraphs to make them understandable
- Re-read entire draft and rephrase to make it cohesive
- Proofread for any errors
- Add references
- Add illustrations and images
- Post

System 2 Characteristics







Conscious

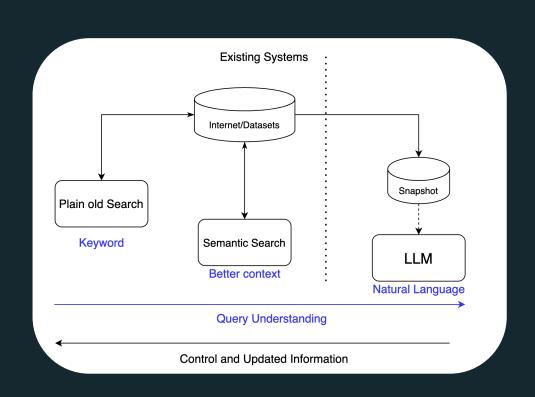
Effortful

Slow

Agents of Change

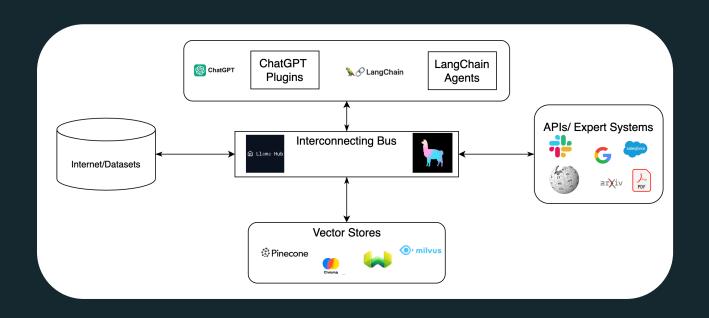


- Existing systems leverage keyword based or semantic search
- Current LLMs are far better at understanding natural language but not good at expert tasks and recent information



Agents of Change

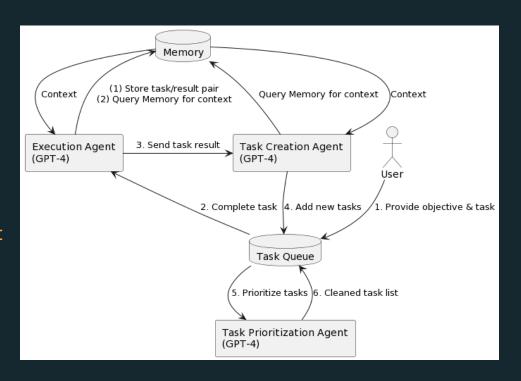




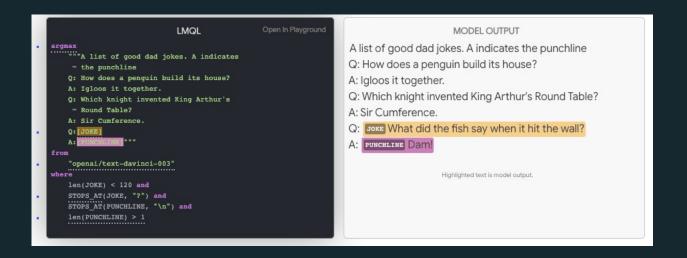
A *Retrieval Augmented System* leveraging LLMs for understanding and expert systems for specific tasks



- Agent driven system
- Ability to leverage external systems/APIs
- Ability to reference additional memory
- Ability to self-prompt/meta-prompt







- Language Model Query Language is a simplified interface for interacting with LLMs
- Expressiveness of Python combined with ease of Natural Language (COBOL YOU there?)
- Enables multi-part prompting, tool augmentation, scripting, constraint guided decoding, etc.

Source: https://github.com/microsoft/guidance

Source: https://lmql.ai/

chatGPT Plugins





Expedia

Bring your trip plans to life—get there, stay there, find things to see and do.



iscalNote

Provides and enables access to select market-leading, real-time data sets for legal, political, and regulatory data and information.



etacart

Order from your favorite local grocery stores.



KAYAK

Search for flights, stays and rental cars. Get recommendations for all the places you can go within your budget.



Klarna Shopping

Search and compare prices from thousands of online shops.



Milo Family Al

Giving parents superpowers to turn the manic to magic, 20 minutes each day. Ask: Hey Milo, what's magic today?



OpenTable

Provides restaurant recommendations, with a direct link to book.



Sho

Search for millions of products from the world's greatest brands.



Speak

Learn how to say anything in another language with Speak, your Al-powered language tutor.



Wolfrag

Access computation, math, curated knowledge & realtime data through Wolfram|Alpha and Wolfram Language.

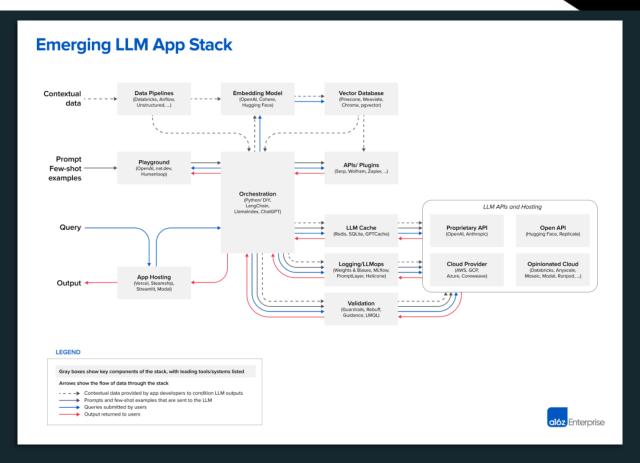


Zapier

Interact with over 5,000+ apps like Google Sheets, Trello, Gmail, HubSpot, Salesforce, and more.

LLM App Stack

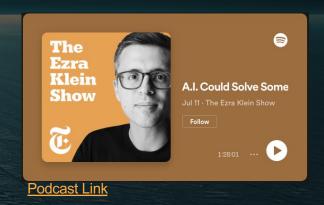




Beyond the Horizon



- Multi-modal Models (successors of MUM, GPT-4, VIMA, PaLM-E)
- Even Longer Context Windows
- Efficient and Faster Training/ Smaller Robust Pretrained Models
- Improvements and Better access to RLHF
- Improvements in LLOps
- Improvements in handling attacks
- Improvements in controlling hallucinations



Thank You!