MSI at the ARF
MARKETING SCIENCE INSTITUTE

MSI Webinar: What is ChatGPT and How Will It Affect Marketing Practice?

July 20, 2023 | Virtual | 12:00 pm – 1:00 pm EDT

Speaker:

Zhou Yu – *Associate Professor, Computer Science, Columbia University.*

Overview:

Zhou Yu, Associate Professor at Columbia University and CEO of Articulate AI Inc. provides a comprehensive overview of large language models, starting from the era before GPT3 and ChatGPT when models required expert data curation to the current state. She discusses the current GPT3 and ChatGPT era which uses multi-layer transformer decoders, yielding output based on probability and prediction using datasets from sources like Common Crawl, WebText, books1, books2 and Wikipedia, rather than human curation. In particular, ChatGPT is a conversation model built on top of the latest version of GPT3 (GPT3.5), which provides a more robust experience using reinforcement learning with human feedback (RLHF) through up or down votes. The human feedback helps to train and supervise the model leading to improvements.

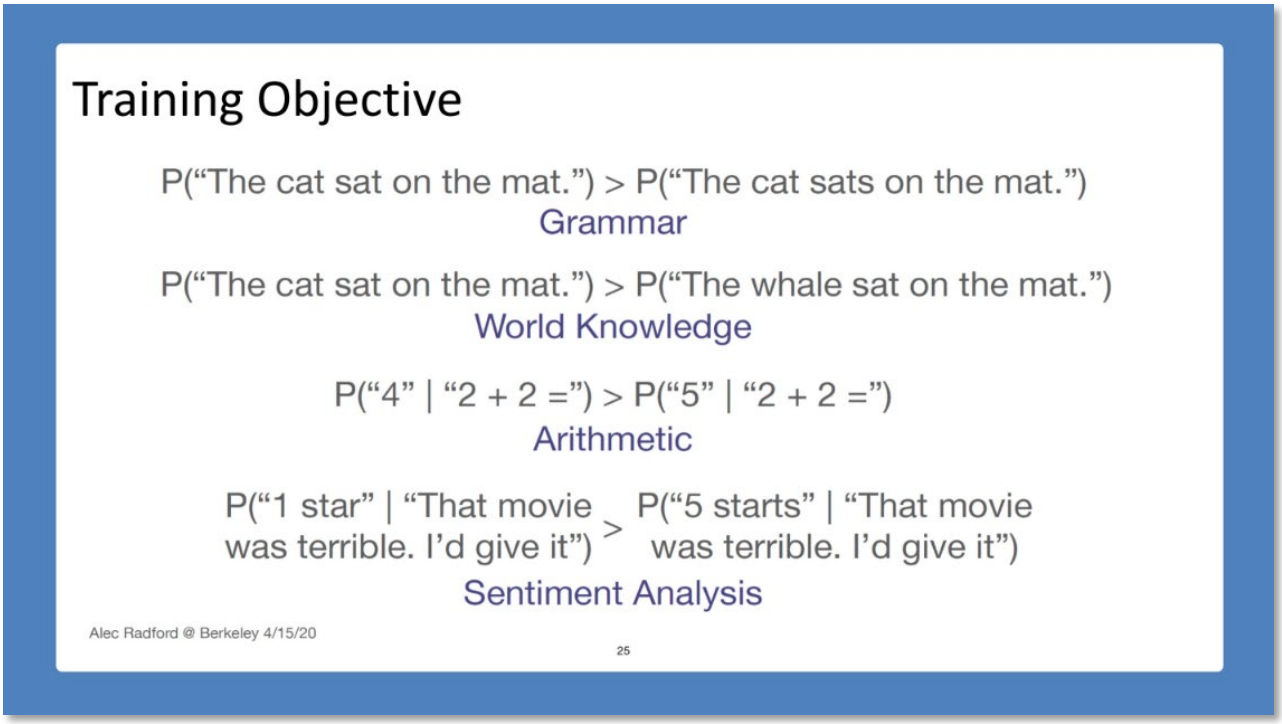
Yu points out that like other computer programs, GPT has areas it excels in and situations that pose challenges for it. GPT models do very well in the areas of knowledge-based questions and trivia question tasks. ChatGPT in particular, can handle up to 8000 characters (as opposed to 2000 for GPT3) providing a more humanlike experience. ChatGPT can be used to perform functions such as compiler, interactive debugging tool, marketing (messages), writing assistance, etc. GPT models become more limited in ability when attempting to complete more complicated tasks (trick tasks) where answers require more complex processing. Yu notes that ChatGPT, as with other AI tools, tends to be costly to run, does not reason well, and can return false or biased information, among other drawbacks. In the future, ChapGPT will be able to accommodate more interactive tasks and encompass increased modalities such as audio, video and images as input.

Takeaways:

Theories Behind Large Language Models

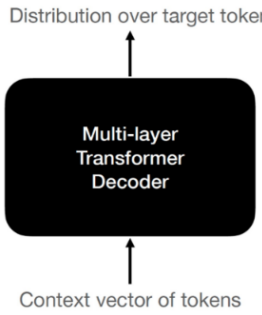
- **Before GPT3 and ChatGPT** models required much more human interaction and supervision. for every new task and were challenging to maintain. Questions and answers within these models **relied on expert data curation.** Experts did not necessarily know what users were going to ask which, led to results that were not always causally related (Spurious Correlation).

- In the **large language model GPT3,** the goal is to **imitate how humans learn** (demonstrations and instructions). This model can understand a lot of general-purpose instructions, does not require the breadth of data points for specific tasks, and has massive capacities (GPT3 model size: 175 billion parameters).
    - Large language models are based on training objectives (predict the next word in sequence) or language modeling and learning by association.
        - In training data sets, the probability of correct grammar or math solutions is typically greater than mistakes.



    - Large language models are trained by using a **multi-layer transformer decoder**, currently the most popular structure of neural networks. These transformers are stacked together forming a very large language model.
        - Tokens (words) are fed to this model which outputs a distribution over target tokens (predicting future values from past values) providing the greatest probability of predicting the next word (Autoregressive Generation). Datasets (text) for these models are derived from Common Crawl, WebText, books1, books2 and Wikipedia.

Model



Autoregressive

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

- **When the model is fed data, the output undergoes prompting** (the instruction given to the model) which can come in one of three instructions or prompts:
  - Zero-shot - No examples are given and the model is given a natural language instruction to figure it out on its own.
  - One-shot - The model is given one example.
  - Few-shot – The model is given many examples. The more demonstrations (examples) the better the results or output.
- When the model is trained or tested with similar tasks, the output will be of higher quality (better answers).
- Large language GPT models excel in **knowledge-based Q&A and trivia question tasks**.

Complete list of tasks

Language Modeling
- PTB

Close and Completion
- ROC Stories
- HellaSwag
- LAMBADA

Winograd-style
- Winograd
- Winogrande

Commonsense Reasoning
- PiQA
- ARC
- OpenBookQA

Reading Comprehension
- QuAC
- SQuADv2
- DROP
- CoQA
- RACE

Trivia-style Questions
- NaturalQs
- WebQs
- TriviaQA

Inference
- ANLI
- RTE

Comprehensive Benchmarks
- SuperGLUE

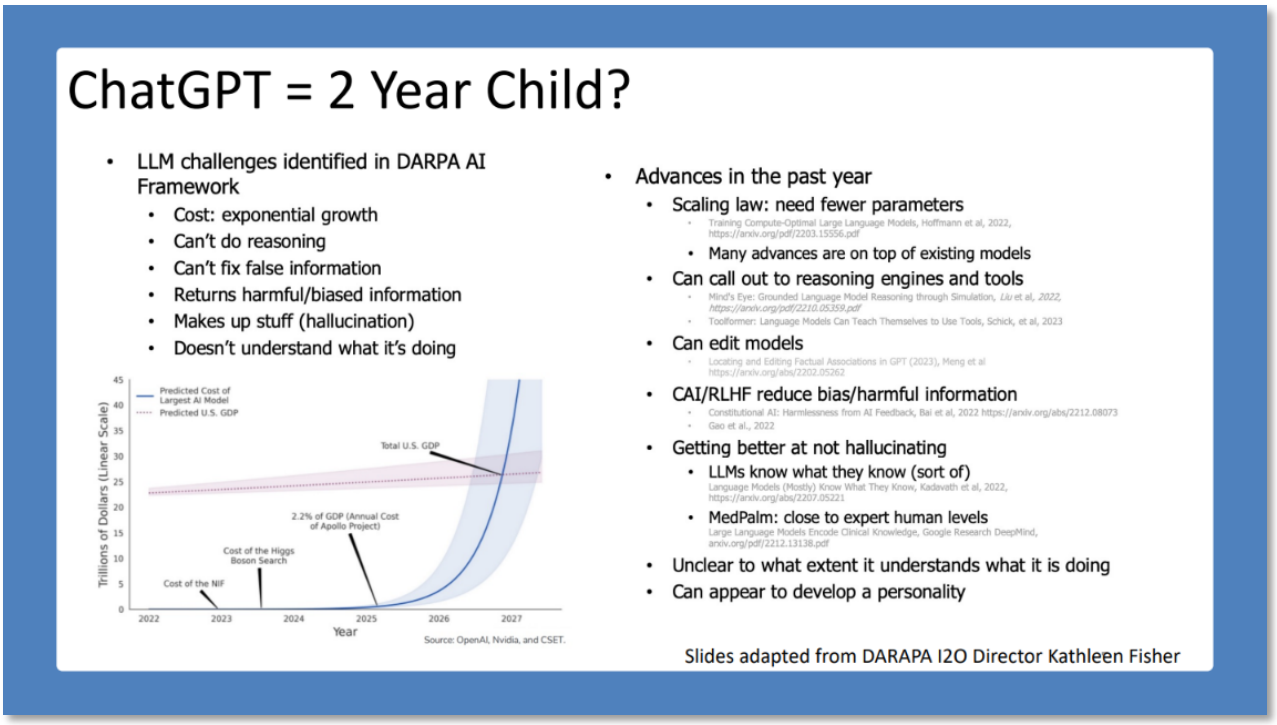
Translation
- En <-> Fr
- En <-> De
- En <-> Ro

Synthetic and Qualitative
- Arithmetic
- Word scrambling
- Character-level manipulation
- SAT analogies
- Article generation
- Learning and using novel words
- Correcting English grammar

55

  - GPT models have **limitations in completing more complicated tasks** (trick tasks) where answers require more complex processing.

- **ChatGPT is a conversation model built on top of the latest version of** GPT3 (GPT3.5). While GPT3 has an architectural limit of 2000 characters to give as context, ChatGPT introduced a chatbot model which can have 8000 characters as context, creating a more human-like conversational experience. This has increased with GPT4, in addition to the ability to handle some images.
  - In the future, it will be able to handle increased modalities such as audio, video and images as input.
  - What makes ChatGPT more robust than GPT3 is that it relies on one extra step which is **reinforcement learning with human feedback (RLHF).**
    - With ChatGPT, humans play the role of both user and AI assistant. Reinforced learning (RL) from human feedback informs preferences (e.g. up or down votes).
    - Human feedback helps to train and supervise the model and use it as the reward function for RL, leading to model improvements.
    - ChatGPT can be used to perform a variety of functions such as compiler, interactive debugging tool, marketing (messages), writing assistance, etc.
  - Running ChatGPT can be very cost prohibitive, as with other AI tools. Additionally, it does not reason well, can't fix false information, can return biased information or false information, etc.



The Evolutionary Tree

- The evolutionary tree follows the history and status of large language models and demonstrates the open-source versus closed-source models and the competition between the two.