Kevin Bryan

A User's Guide to GPT and LLMs for Economic Research

On Thursday, May 11, Kevin Bryan will join Markus' Academy for a lecture on A User's Guide to GPT and LLMs for Economics Research. Bryan is an Associate Professor, Strategic Management Area, Rotman School of Management at the University of Toronto.

A few highlights from the discussion:

- **A summary in four bullets**
    - LLMs have been shown to increase productivity in sales and programming in the past; academia should capitalize on this technology
    - 6 takeaways: (1) Controlling the output of LLMs is difficult, (2) the "Raw" ChatGPT online is far from state of the art, (3) hallucinations are mostly fixable, (4) the technology' rate of improvements is fast, (5) most use cases for economists require using API+code (this will give you much more control on the output), (6) it is cheap to do so
    - The main uses for economists are: (1) cleaning data, (2) programming/making graphs, (3) spelling checks, (4) summarizing literature
    - Some best practices: (1) provide the model with as many examples as you can of what you need to be done, (2) do not use LLMs to do math, (3) use an "ensemble approach" to find the optimal prompt to ask the model

- **[0:00] Introduction**
    - Korinek (2023) has argued that GPT and other LLMs help economic research through ideation, writing, coding, data analysis, and textual analysis
- **[4:45] Basics of LLMs**
    - Large language models are novel: a team at Google in 2017 introduced the fundamental insights in their paper "Attention is All You Need"
    - Astonishing results have come with GPT3 in 2021 and Dall-E/Stable Diffusion (image generators) in 2022. ChatGPT is released Nov 2022.
    - LLMs have been shown to increase productivity in sales and programming in the past; academia should capitalize on this technology
    - How do they work? Take a huge corpus of written language to predict the next word in a sentence. The model tries to adjust the posterior distribution of words that will follow, based on words that already exist, to produce a response. This is why LLMs are called "stochastic parrots"
- **[16:25] Main research uses**
    - Entity recognition in unstructured data
    - Speed up programming/graphics

    - Spell/grammar check
    - Summarizing the literature
    - Soon, it will be able to check mathematic calculations or derivations
    - But also think of “Architectural Innovation” (Henderson and Clark 1990), the first thing we do with a new technology is replacing tasks, but the long term benefit is the ability to do new tasks we can't think of doing yet
- **[19:50] How to control the output of an LLM**
    - When controlling the model, our goal is to shift that posterior distribution
    - Zero-shot prompt engineering: changing the prompt of the question
    - Zero-shot context definition. Give context about the thing we are interested in
    - Few-shot context definition. Give it examples of how we would like the model to respond. In industry this is the most common way to control LLMs
    - Fine-tuning. Train the model only on a certain set of text: training it on academic papers will yield academic language. Most effective but expensive
    - Keep in mind: LLMs do not understand your question, they only predict your next word in the context of your question. Because of this, you can't ask it how to better control its own output, since there has been no text on the topic
- **[32:06] Examples of LLM use cases**
    - In the first example, we used chatGPT to read through academic papers and determine why certain citations were included (for background, or for introducing a new tool/technique). An “ensemble approach”, where we tried different prompts and had the LLM choose the optimal one, improved results
    - In the second, we used it to clean data from an unreadable and old file format
    - In the third, we used it to clean a draft of a paper which, after OCRing it, had lost its math symbols and special characters. The output was clean Latex code
    - Lastly, we asked the chat to make us the basic chart of supply and demand (with perfect formatting)
- **[45:27] Practical fixes**
    - Do not use the web ChatGPT. Get access to the API
    - When you feed it data, use few-shot contexting. Give as many of the examples as possible of what you want to be done.
    - Play with prompts, use an ensemble approach
    - Ask follow-up, concrete, questions (“is this fact correct?”), to avoid hallucinations
    - For research, within the API set the “temperature” at zero. It makes answers more deterministic. Leads to less creativity, but much less hallucination. If you code with STATA, where there is much less content on the internet, it might be helpful to increase it
    - To handle the model’s memory limit, we can (1) break the content to be processed into different chunks, (2) use vector embedding and cosine similarity to find the relevant text for a given problem
    - Do not use LLMs to do math. Suppose you need to multiply two 8-digit numbers together. That computation has never been done before, so the LLM will give you the wrong answer

    - Soon, (1) OCR will be replaced by LLM-based reading of images, (2) the LLM will run code itself, (3) LLMs will be able to search the web to verify facts
    - Optimal prompting is not well understood yet

Timestamps:
[4:45] Basics of LLMs
[16:25] Main research uses
[19:50] How to control the output of an LLM
[32:06] Examples of LLM use cases
[45:27] Practical fixes