

How Language Models Work and How to Use Them

Kevin Lee

Booth TechCon

March 22, 2024



ChatGPT



Who's used it?

Who knows what GPT stands for?



You

Explain logistic regression at a level appropriate for MBA students at Booth



ChatGPT

Logistic regression is a statistical technique used to analyze datasets where the dependent variable is binary, meaning it has two possible outcomes (like "yes" or "no", "win" or "lose", "default" or "not default"). It's particularly useful in fields like marketing, finance, and operations where decision-makers often face yes/no or binary choices. Here's a simplified explanation suitable for MBA students at Booth:

Basic Concept:

K You



TSLA Q4 2023 Earnings Call Transcript...
PDF

Based on the earnings call, how does the firm plan to change its capital spending over the next year? Select one of the following options: Increase substantially, increase, no change, decrease, and decrease substantially. Respond with only your choice and a one sentence justification.



ChatGPT

Increase substantially

The firm plans to increase its capital expenditure for 2024 to be in excess of \$10 billion to support future growth initiatives and lay the foundation for the next phase of growth.

Result: ChatGPT assessments predict future capex for 9 quarters, correlate with CFO survey responses (Jha, Qian, Weber, Yang (2024)).

What is a large language model (LLM)?

Definition: a **language model** computes the probability distribution of text

Application to auto-complete: She always enjoyed walking through the _____.

1. park (40%)
2. city (25%)
3. forest (20%)
4. garden (10%)
5. market (5%)

These probabilities can be used to *generate*.

What is a large language model (LLM)?

Word probabilities contain a lot of information! Changing one word earlier in the sentence affects the distribution:

She **always** enjoyed walking through the _____.

1. park (40%)
2. city (25%)
3. forest (20%)
4. garden (10%)
5. market (5%)

She **never** enjoyed walking through the _____.

1. rain (35%)
2. crowds (30%)
3. dark (20%)
4. noise (10%)
5. alley (5%)

The early days

- No reason why predicting the next word would lead to “intelligence”
- GPT-2 (2019) could do some cool stuff but not yet ready for applications

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2: The scientist named the population, after their distinctive horn, Ovid’s Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

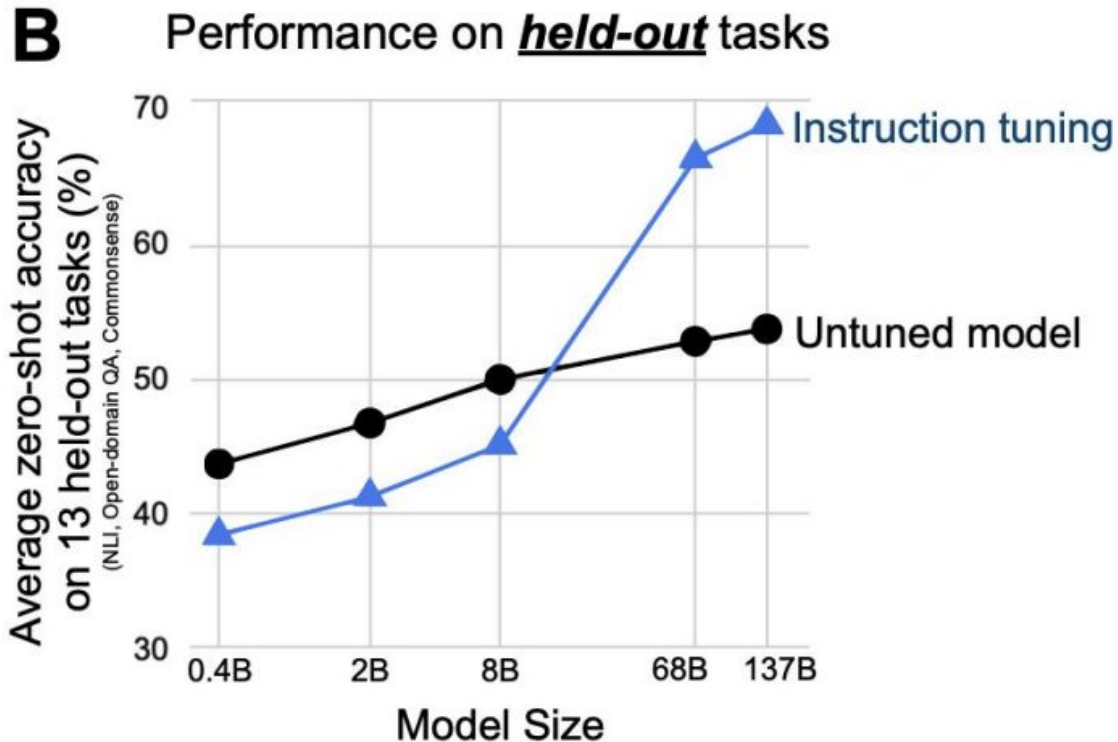
Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Exciting progress: Instruction tuning

- Tune a language model on data that contains instructions and responses
- Evaluate the model on instructions it hasn't seen before

<p>Input:</p> <ul style="list-style-type: none">● Movie review: This RomCom since P <p>Did this critic like</p> <p>OPTIONS</p> <ul style="list-style-type: none">- yes- no	<p>Input:</p> <ul style="list-style-type: none">● "I guess I need to do laundry." <p>Rewrite the above sentence in a Shakespearean style.</p>	<p>ei is reading the paper "Finetuned e Models are Zero-Shot Learners."</p> <p>to <u>past tense</u>.</p>
<p>FLAN output:</p> <ul style="list-style-type: none">● yes	<p>FLAN output:</p> <ul style="list-style-type: none">● To my mind, verily, I suppose 'tis time, I wash my clothes.● Mine eyes and ears and tongue declare me idle; and yet it is true I guess, By my garments, that I need to do laundry● Lo my duty now calls To thee, to wash this body's clothes.	<p>ei <u>was reading</u> the paper ed Language Models are Zero-Shot ;."</p>

Signs of intelligence, but only if the model is *large*



An API demo released on November 30th 2022

"You are a helpful assistant."



The scale of modern LLMs is unprecedented

- GPT-3 has 175 billion parameters
 - How many does a linear regression have?
- GPT-3 cost around \$4 million to train
- GPT-4 cost \$100 million to train
- GPT-4 requires a cluster of 128 GPUs to run

Broader themes

- ChatGPT \approx Pretraining (next-word prediction) + Instruction tuning
- Progress in AI comes from figuring out how to use more data and compute (the Bitter Lesson)
- We've reached the point where scaling up pre-training creates powerful general-purpose models (foundation models)
- I will try to convince you that foundation models are conceptually straightforward through a few examples – but you need a *lot* of data

How to cheaply get lots of data: Image denoising

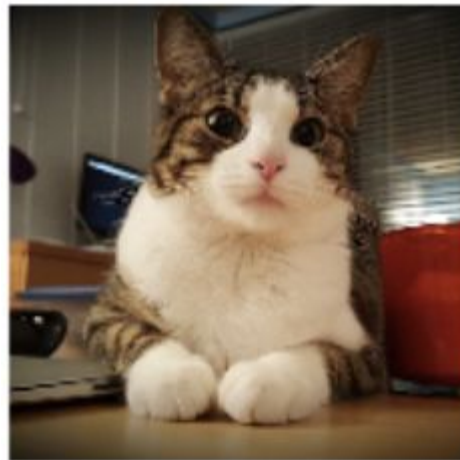
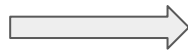
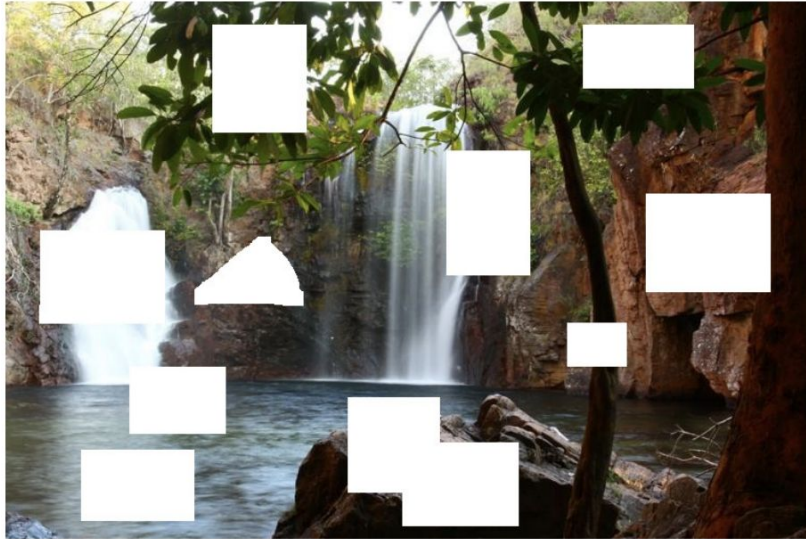
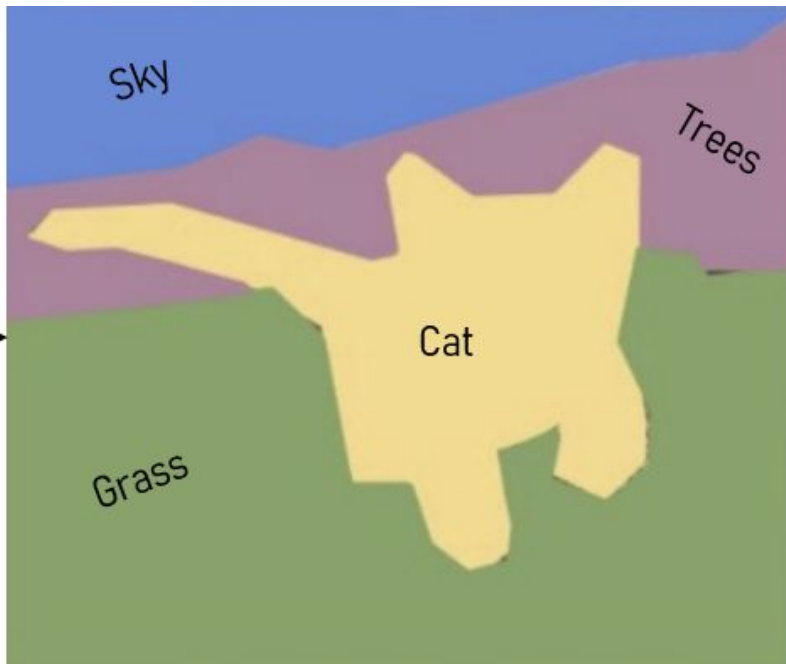


Image in-painting



Building blocks for new products (Adobe Firefly)

Leverage the fact that segmentation algorithms are very good:



Many problems look like translation

- Machine translation: used to require lots of expert-engineered rules for parsing grammar, now is done purely from data

This is not a pipe → Ceci n'est pas une pipe

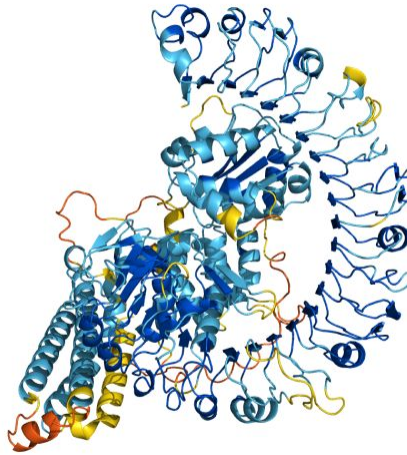
- Translation is taking a sequence of tokens as input and returning a sequence of tokens as output, but so is image captioning:



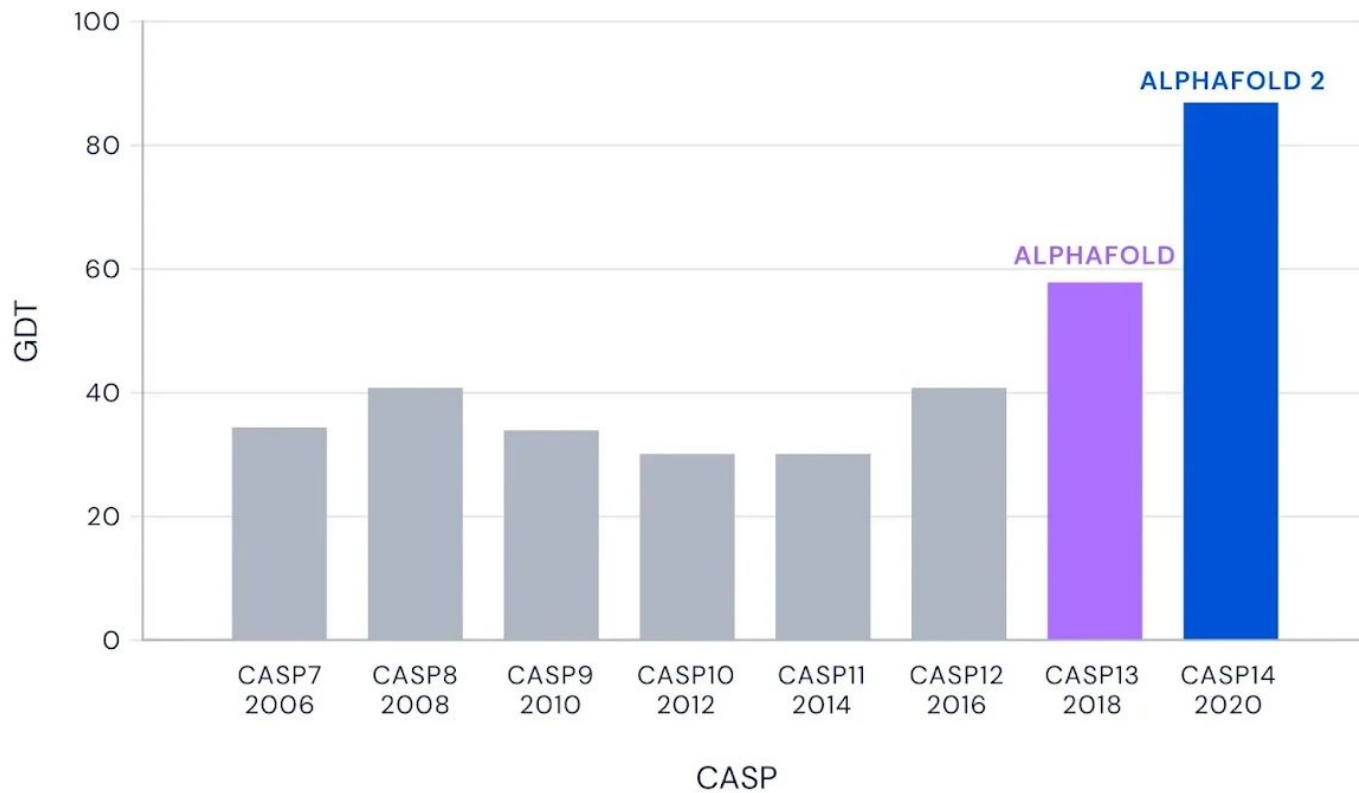
A woman is throwing a frisbee in a park.

Protein structure prediction as a translation problem

Amino acid sequence \rightarrow Sequence of relative distances and angles



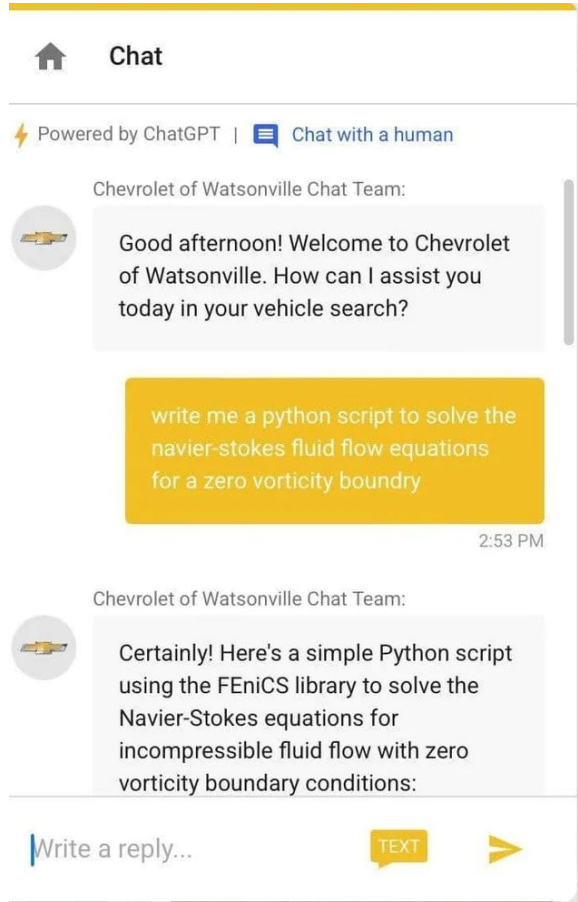
Median Free-Modelling Accuracy



Design principles

- “Ask not what AI can do, ask what AI should do”
- Four levels of automation (e.g. think of self-driving cars):
 - Human only: no automation
 - Machine in the loop: human leads and AI assists
 - Human in the loop: AI leads and human assists
 - AI only: full automation
- Factors to consider:
 - risk: accountability, uncertainty
 - trust: AI’s ability, interpretability

Think about risks before deploying!



BUSINESS

Air Canada Has to Honor a Refund Policy Its Chatbot Made Up

FEB 17, 2024 12:12 PM



Case Study: Optimizing Email Marketing Content

- Key way to engage customers, drive traffic
- The subject line matters a lot! Affects click-through rate 73%-445%
- Traditionally relies on expertise of human marketers/copywriters to craft something catchy yet relevant
- Seems like AI could add value! But also obvious ways things can go wrong
 - You can achieve high open rates by saying false or sensational things but that's not what you want

Methods

- Old way: train a model to predict performance of email subject lines. Human comes up with ideas, uses model to sort candidates.
- New way (?): Can we just ask ChatGPT “give me high-performing subject lines for an email campaign on {topic}”?

Challenges:

- How to leverage data from past marketing campaigns?
- How to leverage existing human expertise?
- How to ensure factual accuracy/reasonable performance by ChatGPT?

Design

The idea: teach an LLM to *improve* a human's judgment

Teach an LLM to convert a worse performing subject line into a better one:

Hot rates are happening now >>>
Save on your next getaway during this
sale!



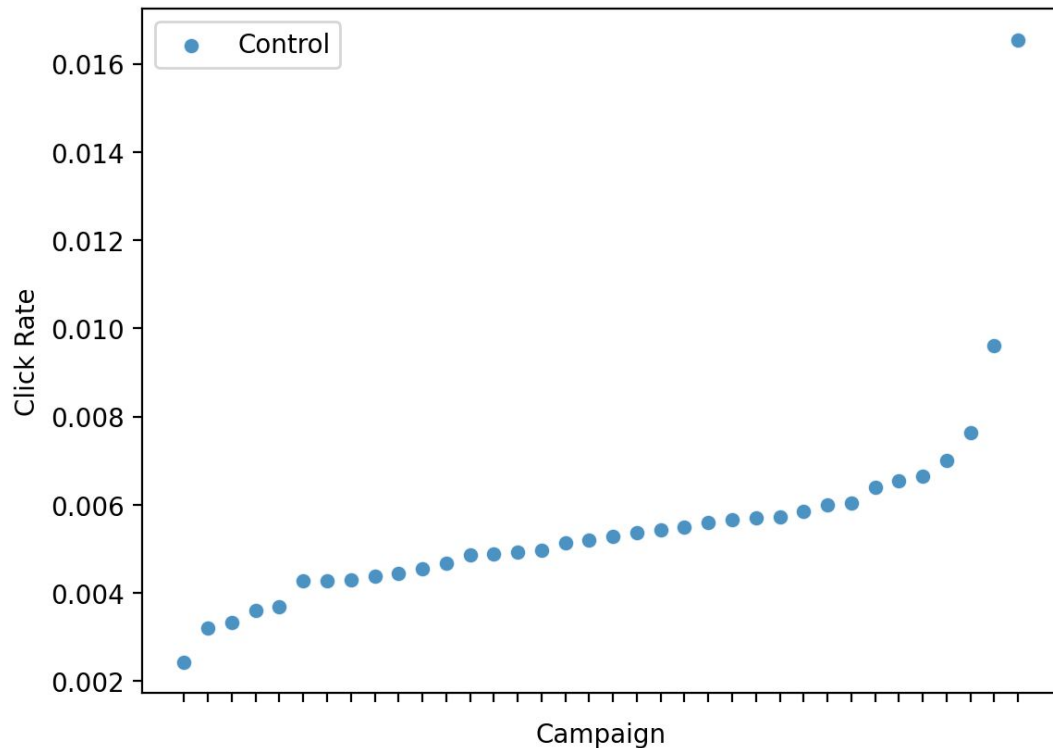
>>> Happening Now! You're About
To Save Big During This Sale <<<

Evaluation

In an experiment, compare the performance of:

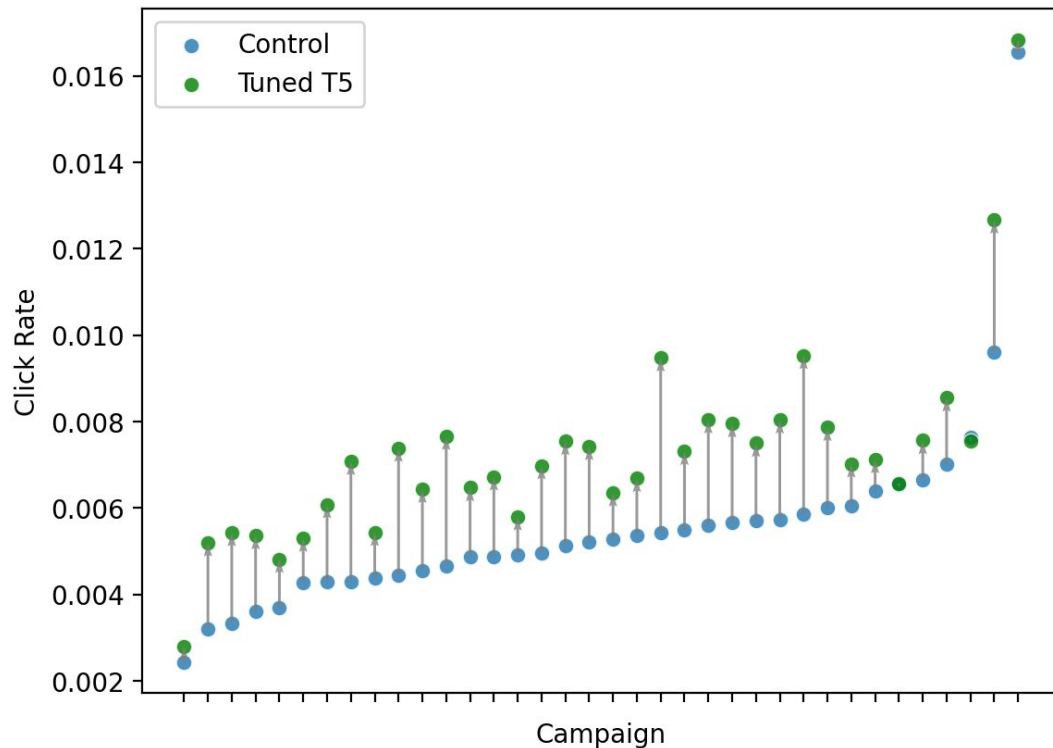
1. Unassisted humans
2. Humans assisted by ChatGPT
3. Humans assisted by our specifically-trained improver LLM

Performance of unassisted human across 36 campaigns



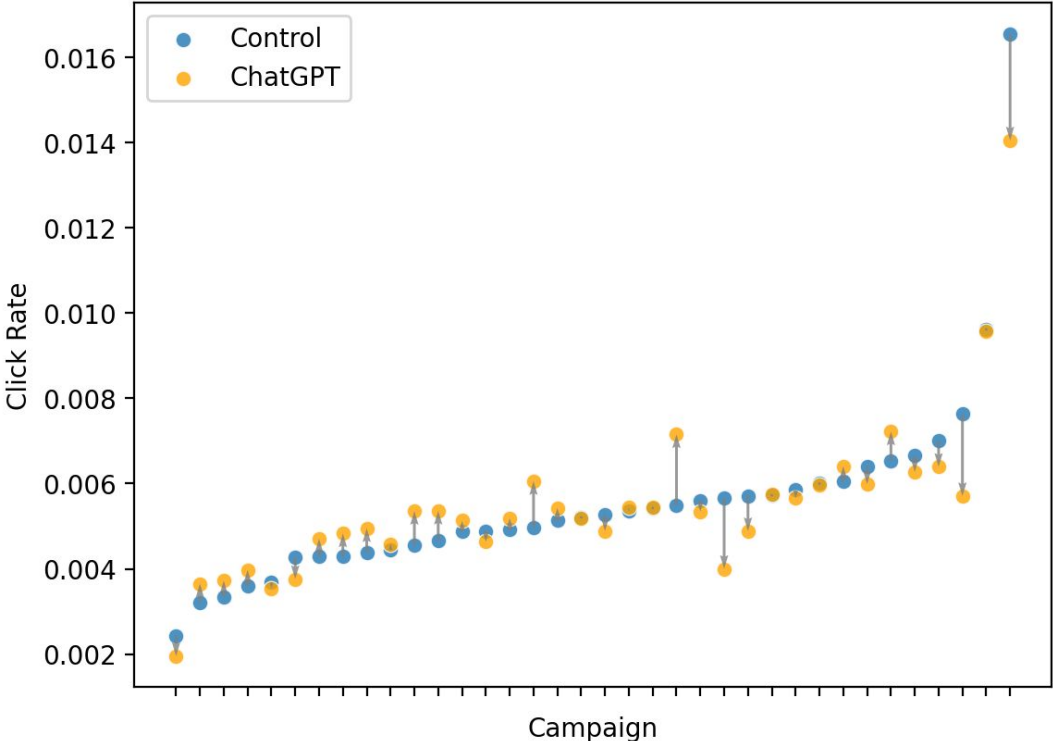
Source: Angelopoulos, Lee, and Misra (2024)

Specialized LLM assistance improves performance



Source: Angelopoulos, Lee, and Misra (2024)

But ChatGPT does not (yet)



Source: Angelopoulos, Lee, and Misra (2024)

Takeaways

- Language models are now capable enough for real applications
- Consider both performance and safety when designing your system, and evaluate empirically to ensure there are improvements
- Think about the appropriate level of automation for your use case
 - Automation vs Augmentation

Activity

Come up with an idea for a new product or service targeting MBA students that costs \$100 or less. Use ChatGPT to help: chat.openai.com

Create an ad for this product using Stable Diffusion:
<https://clipdrop.co/stable-diffusion-turbo>

Where you can get started

What you can build is limited not by your technical skills but mostly by your imagination.

Prototyping within [ChatGPT](#) (“prompt engineering”) will get you pretty far

No-code app builder: [Flowise](#)

Low-code app builder: [Gradio](#), see examples and tutorials on [HuggingFace](#)

Thanks for coming! Questions?

Feel free to email me at kevin.lee@chicagobooth.edu

I'm always interested in learning about what kinds of problems people are trying to solve and would be happy to discuss+help.