About LLMs, LMMs

Origins, Latest Developments
Administrative Details

● Finalize Team Formation Today
  ○ Will share the project template today
    ■ Create a Github repo and add github user cs224g to it
  ○ Any final questions before the main lecture starts?

● Biggest Question so far
  ○ How many and when do we get credits
  ○ Will get details from our partners starting today
Slide content due to:

- Monica Lam
- Eric Wallace
- Andrej Karpathy
\[ p(x_1, x_2, \ldots, x_L) \]
Language Modeling

\[ p(x_1, x_2, \ldots, x_L) \]

\[ p(\text{the, mouse, ate, the, cheese}) = 0.001 \]

\[ p(\text{mouse, the, cheese, ate, the}) = 10^{-20} \]
Language Modeling

\[ p(x_1, x_2, \ldots, x_L) = \prod_{i=1}^{L} p(x_i \mid x_{1:i-1}) \]
Neural Language Models

\[
\prod_{i=1}^{L} p(x_i \mid x_{1:i-1})
\]
Neural Language Models

Prompt

The mouse ate the

\[ \prod_{i=1}^{L} p(x_i \mid x_{1:i-1}) \]
The mouse ate the

Neural network

Prompt

The mouse ate the

\[
\prod_{i=1}^{L} p(x_i \mid x_{1:i-1})
\]
The mouse ate the

Neural network

Prompt
The mouse ate the

<table>
<thead>
<tr>
<th>Token</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>cheese</td>
<td>0.20</td>
</tr>
<tr>
<td>cookie</td>
<td>0.12</td>
</tr>
<tr>
<td>nibble</td>
<td>0.08</td>
</tr>
<tr>
<td>crumb</td>
<td>0.07</td>
</tr>
<tr>
<td>man</td>
<td>0.05</td>
</tr>
<tr>
<td>tail</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Jan Jannink, John Whaley

CS 224G   -   1/18/2024
Data for Pre-training

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sampling prop.</th>
<th>Epochs</th>
<th>Disk size</th>
</tr>
</thead>
<tbody>
<tr>
<td>CommonCrawl</td>
<td>67.0%</td>
<td>1.10</td>
<td>3.3 TB</td>
</tr>
<tr>
<td>C4</td>
<td>15.0%</td>
<td>1.06</td>
<td>783 GB</td>
</tr>
<tr>
<td>Github</td>
<td>4.5%</td>
<td>0.64</td>
<td>328 GB</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>4.5%</td>
<td>2.45</td>
<td>83 GB</td>
</tr>
<tr>
<td>Books</td>
<td>4.5%</td>
<td>2.23</td>
<td>85 GB</td>
</tr>
<tr>
<td>ArXiv</td>
<td>2.5%</td>
<td>1.06</td>
<td>92 GB</td>
</tr>
<tr>
<td>StackExchange</td>
<td>2.0%</td>
<td>1.03</td>
<td>78 GB</td>
</tr>
</tbody>
</table>

Table 1: **Pre-training data.** Data mixtures used for pre-training, for each subset we list the sampling proportion, number of epochs performed on the subset when training on 1.4T tokens, and disk size. The pre-training runs on 1T tokens have the same sampling proportion.
<table>
<thead>
<tr>
<th>Model</th>
<th>Year</th>
<th>Size (B)</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELMo</td>
<td>Feb 2018</td>
<td></td>
<td>&quot;First to pretrain on unlabeled text&quot;</td>
</tr>
<tr>
<td>GPT-2</td>
<td>Feb 2019</td>
<td>1.5B</td>
<td>&quot;Generative&quot;</td>
</tr>
<tr>
<td>BERT</td>
<td>Oct 2018</td>
<td>340M</td>
<td>&quot;Pretrain and fine-tune&quot;</td>
</tr>
<tr>
<td>GPT-3</td>
<td>June 2020</td>
<td>175B</td>
<td>&quot;First to remove fine-tuning&quot;</td>
</tr>
<tr>
<td>Megatron</td>
<td>Oct 2021</td>
<td>530B</td>
<td>&quot;Best open-release model at the moment&quot;</td>
</tr>
<tr>
<td>PaLM</td>
<td>Apr 2022</td>
<td>540B</td>
<td></td>
</tr>
<tr>
<td>LLAMA</td>
<td>Feb 2023</td>
<td>65B</td>
<td></td>
</tr>
<tr>
<td>Chinchilla</td>
<td>Mar 2022</td>
<td>70B</td>
<td>&quot;Data size is as important as parameter count&quot;</td>
</tr>
<tr>
<td>ChatGPT</td>
<td>Nov 2022</td>
<td>(Unknown)</td>
<td></td>
</tr>
<tr>
<td>GPT-4</td>
<td>Mar 2023</td>
<td>(Unknown)</td>
<td>&quot;Multimodal&quot;</td>
</tr>
</tbody>
</table>
LLMs Today

Jan Jannink, John Whaley
Failures of Pre-trained LMs

**Prompt**

*Explain the moon landing to a 6 year old in a few sentences.*
<table>
<thead>
<tr>
<th>PROMPT</th>
<th>Explain the moon landing to a 6 year old in a few sentences.</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPLETION</td>
<td>GPT-3</td>
</tr>
<tr>
<td>Explain the theory of gravity to a 6 year old.</td>
<td></td>
</tr>
<tr>
<td>Explain the theory of relativity to a 6 year old in a few sentences.</td>
<td></td>
</tr>
<tr>
<td>Explain the big bang theory to a 6 year old.</td>
<td></td>
</tr>
<tr>
<td>Explain evolution to a 6 year old.</td>
<td></td>
</tr>
</tbody>
</table>
## Failures of Pre-trained LMs

<table>
<thead>
<tr>
<th>PROMPT</th>
<th>Explain the moon landing to a 6 year old in a few sentences.</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPLETION</td>
<td>GPT-3</td>
</tr>
<tr>
<td></td>
<td>Explain the theory of gravity to a 6 year old.</td>
</tr>
<tr>
<td></td>
<td>Explain the theory of relativity to a 6 year old in a few sentences.</td>
</tr>
<tr>
<td></td>
<td>Explain the big bang theory to a 6 year old.</td>
</tr>
<tr>
<td></td>
<td>Explain evolution to a 6 year old.</td>
</tr>
</tbody>
</table>

Language models are not *aligned* with user intent [Ouyang et al., 2022]. Finetuning to the rescue!
## Failures of Pre-trained LMs

<table>
<thead>
<tr>
<th>PROMPT</th>
<th>Explain the moon landing to a 6 year old in a few sentences.</th>
</tr>
</thead>
</table>
| COMPLETION | **Human**  
A giant rocket ship blasted off from Earth carrying astronauts to the moon. The astronauts landed their spaceship on the moon and walked around exploring the lunar surface. Then they returned safely back to Earth, bringing home moon rocks to show everyone. |
Failures of Pre-trained LMs

<table>
<thead>
<tr>
<th>PROMPT</th>
<th>Explain the moon landing to a 6 year old in a few sentences.</th>
</tr>
</thead>
</table>
| COMPLETION | **Human**  
A giant rocket ship blasted off from Earth carrying astronauts to the moon. The astronauts landed their spaceship on the moon and walked around exploring the lunar surface. Then they returned safely back to Earth, bringing home moon rocks to show everyone. |
LLM Accomplishments

ChatGPT: US Medical Licensing Exam  Pass!

GPT-4: Uniform Bar Examination  Pass!

https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000198
Is GPT Ready to Practice?

Exams are designed for humans, not AI
Humans passing the bar can practice
  • They know when to look up info
But not AI bots
  • don’t know the meaning of their words
  • make mistakes no humans will
It Hallucinates!

Generative AI = Hallucination
As of September 2021, the monarch of the United Kingdom is Queen Elizabeth II. She has been the queen since February 6, 1952, following the death of her father, King George VI. Queen Elizabeth II is the longest-reigning monarch in British history and has played an important role in the country's political and cultural life over the past seven decades.
ChatGPT (7/2/2023)

+ Conversational
- Hallucinates on long-tail info.
  With details and conviction

1978  Hinton, G. E.
  Relaxation and its role in vision.
  PhD Thesis, University of Edinburgh

Geoffrey Hinton received his PhD in Artificial Intelligence from the University of Edinburgh in 1978. His doctoral thesis was titled "Relational Learning of Pattern-Matchable Templates". Hinton is widely regarded as one of the pioneers of deep learning and has made significant contributions to the field of artificial intelligence and machine learning throughout his career.
Buying a Ticket from GPT-3

User: I’d like to buy a plane ticket to New York.

Agent: Sure, I can help you with that. What date and time would you like to depart?

User: July 7th, 9am.

Agent: Alright, I can book a flight for you departing on July 7th at 9am. What airport would you like to depart from?

User: SFO

Agent: Great, I can book a flight for you departing from San Francisco on July 7th at 9am. Would you like to book this flight now?

User: What is the flight number?

Agent: The flight number is AA123.

+ Perfect flow without training
- Hallucinates: American AA123 DFW △ HNL
- Needs to consult DB
Examples of (Failed) Business Use Cases

- EPIC medical use cases (with expert reviews)
- A large manufacturer
  - “Our pretrained agent hallucinates components that don’t exist”
- Microsoft travel article (with human reviewers)
  - “Visit Food Bank if you are hungry in Ottawa”
  - Withdrawn
We All Have an “LLM” in Our Brain!

- Our “LLM”: Speech area
- Prefrontal cortex
  - Inhibition and attention
  - Planning
  - Long-term memory
  - Monitoring external signals
- If prefrontal cortex is damaged, “Speak without thinking”
Practical challenges

Associated with training
- Bias
- Stereotyping
- Costs
- Access
- Legal issues
Practical challenges

Associated with the operation of the technology

- Hallucinations
- Catastrophic forgetting
- Centralization (versus open source)
- Legal issues
- Prompt injection
- Guardrails

Jan Jannink, John Whaley  
CS 224G  -  1/18/2024