AI Models and Scalable Apps

The New Frontiers of the Possible
Administrative Details - 1

- **Homework assignment**: fill out the Project Team Form today at the latest
  - [https://forms.gle/zdTvYF1qno5zz6v86](https://forms.gle/zdTvYF1qno5zz6v86)
  - Not everyone filled a form for a project proposal (which was homework by the way)

- **Homework assignment**: use the slack workspace
  - [https://cs224g-winter-2024.slack.com/](https://cs224g-winter-2024.slack.com/)
  - There are still opportunities to find teams and adjust projects
  - Conversations generate ideas for improvements

- **Extra credit**: track the latest public info on the webpage
  - [https://cs224g.stanford.edu/](https://cs224g.stanford.edu/)
  - Be on time! We’ll have a link for attendance starting Thursday…
Administrative Details - 2

- **Grading**:
  - 10% participation
  - 15% Feb. 1 sprint
  - 15% Feb. 15 sprint
  - 15% Feb. 29 sprint
  - 45% Mar. 14 demo day

- **Participation** (also counts for extra credit)
  - Interaction with instructors, TAs, other students
  - In person, slack, email, github
  - Asking questions, sharing project ideas, details, constructive commenting, project reviews
Administrative Details - 3

- **Sprints**:
  - All projects on GitHub, having added user cs224G to your repo
  - Recommend merging code on Wednesday by noon
  - Testing and bug fixing on Wednesday afternoon and evening
  - We will clone your repos starting Thursday afternoon (after class)

- **README.md**
  - Describes how to use the app, link to test data, or interesting things to try
  - Every sprint, each team member adds a subsection describing their own contribution
    - Managing merges, documenting, data contributions, test harnesses all count
  - We review commits, so no squash commits, or other approaches that might mask your effort
Administrative Details - 4

- **Sprint Feedback**
  - We will offer teams a chance to share their progress with the class.
  - Sharing your repo and reviewing others increases learning, adds to your score.
  - Your score will combine code testability, personal contributions, commits.

- **Demo Day: March 14, 2024, 4PM-6PM**
  - Venue with space to present and also to show a poster or live demo (finalizing).
  - Guests from academia, incl. other students, industry, investment all welcome.
  - Present and interact with peers, getting their feedback and comments.
  - Participation Extra Credit: Feedback to other projects.

Jan Jannink, John Whaley       CS 224G      - 1/23/2024
Administrative Details - 5

- **Credits**
  - Anthropic accounts by Thursday
  - Working on setting up:
    - AWS
    - GCP
  - Waiting on answers from:
    - OpenAI
    - Google Gemini
    - Baseten
Administrative Details - 6

● GenAI Startup Demo Day Event this Sunday 4pm
  ○ At Stanford Frontier Technology Lab in the Engineering Quad - Y2E2 #292B
  ○ New startup companies built on LLMs
  ○ Register at: https://lu.ma/InceptionDemoDay8
Discussion

- What do you need to succeed?
- Are people still looking for projects?
- How can we be clearer about tasks/homework/project?
- What are next team steps?
  - Choose how you want to manage github responsibility
  - Do you want to do PRs, will you use something like trello
  - Create a GitHub repo, add all team members
  - Set up a commitments section in README.md
    - Have each team member commit to something up front at the beginning of each sprint
  - It is super powerful to review these at the end of each sprint
News

- Meta buying 350K H100 GPUs from NVIDIA this year (~$30K each)
  - ~20% of total H100 manufacturing capacity, over $10,000,000,000
- Alpha Geometry
  - Roughly gold medal standard for international math olympiad
- Raspberry Pi & Mobile Phones running LLMs
  - 7B parameter, quantized models
Foundations of AI Model Scalability

- Moore’s Law
  - $100 buys 100x the fastest computer of 50 years ago ($10,000,000)

- WWW i.e. Data
  - ~11,000 minutes of video added to YouTube per minute, same for TikTok

- Algorithms Advances
  - Deep Learning
  - Attention
  - Unsupervised learning
The AI Scalability Gradients

- PetaFLOPS
- PetaBytes
- TeraParams - (Gemini Ultra)

- Context window from 1K - 128K in 5 years
  - GPT-4-Turbo
- 50+% yearly inference price/performance improvement
- 7B parameters can compete with 600B parameters!
  - Hugging Face leaderboard
The Age of the GPU

- NVIDIA H100 Architecture
  - 80B transistors
  - >15K cores
  - Optimized for parallel multiply/add operations
  - >1 TeraFLOPS at 16bit precision
  - >900 GB/s bandwidth
LLM Training Costs

- GPT-2: $50K
- GPT-3: $4M
- GPT-4: > $100M
- Gemini Ultra: ~ $1-2B
Inference As A Service

- GPU shortage limiting self hosting and even cloud hosting options
  - Months in advance reservation times on AWS
- Inference APIs from major players and cloud vendors
  - OpenAI / Azure leading the pack
- Open source model hosting platforms
  - Possible to query LLAMA and Mistral using the OpenAI API
Emergence - A Primer

- 100 GigaNeurons
- 100 GigaLives
- 100 KiloYears
- Individuals
  - Albert Einstein, Marie Curie, Simone Biles, Alex Honnold
- Groups
  - Native Hawaiians, NASA moon landing, Silicon Valley
What is Emerging From LLMs?

- A world model
  - Details of the universe as understood from a vantage point recapitulating a dominant worldview

- A human model
  - A catalog of human archetypes and behaviors

- A communication model
  - A gradient of interactions going from the most fanciful to the most rigorous
What is Not Yet Emerging?

Reasoning

Not a truly solitary act

Autonomous innovation

The art of the mashup

Unconstrained self improvement

Beyond Alpha Zero

Feedback loops

Jan Jannink, John Whaley   CS 224G   -   1/23/2024
What We Are Not Going to Invent Today

- GPT in advanced R&D
  - Beyond Alpha Fold
- Alpha Riemann
  - Beyond Alpha Geometry
- GPTn builds GPTn+1
  - When GPTn fires OpenAI staff
What We Do Next

- Use Emergence Insight to Develop Tools
  - Capitalize on growth trends
  - Understand model strengths
  - Provide context to ground models
  - Counteract/Avoid bias if needed
  - Skip long tail problems
Context for the Early Win

- Prompts can be much more elaborate than before
- Token pricing is dropping fast allowing greater creativity
- Provide exactly the data you want the model to discuss
- Don’t exceed the scope of the model training
An Anthropomorphic Roadmap For Prompting

- Write a scenario for how you are going to solve your project
- Your project is a performance, and models are the actors
- The decor is the context and support code you provide
- The narrator provides the flow of data setting up the actors
- The actors perform the scenario and the curtain call is the successful outcome
An Agile Approach to Getting There

- Start with bare bones everything
- Build up the from no decor and a single actor performing a soliloquy
- Iteratively build the story from the central theme
- Add other main characters, then bit players
- Add narration and decor to flesh out the story