- Introduction
- Past Trends
- Future Trends
- Swiss AI Initiative
- Q&A
Introduction

- Imanol Schlag
- Neural network research since 2016
- PhD at the Swiss AI Lab (IDSIA) under Jürgen Schmidhuber
- 20+ publications on ML/AI
- Research Scientist at the ETH AI Center
2020: GPT 3

- GPT 2, February 2019
- 175B Parameters, May 2020
- In-context learning -> excel at NLP tasks

Kaplan et al. (January 2020)
2021: First Copycats and API-based Research

- Jurassic-1, Yuan 1.0, Ernie, Gopher
- Sparse methods at a Trillion parameters: GShard, GLaM
- Research explosion based on OpenAI API
  - Many new benchmarks
  - GPT-3 is not truthful / hallucinations
  - Prompting is brittle and challenging
  - Increased discussion around ethics & safety
2022a: Bigger! Better?

500B scale models / first GPT-3 level open source copycats
2022b: Largest Models are Undertrained

March: Chinchilla & Scaling Laws

Hofmann et al. (2022)
2022c: Instruction Tuning and Alignment

- InstructGPT, Jan 2022
- RLHF, March 2022
2023a: Chatbots!

ChatGPT release: November 2022

Road To 100 Million Users For Various Platforms

- ChatGPT: 2 Months
- TikTok: 9 Months
- Youtube: 1.5 Years
- Instagram: 2.5 Years
- Facebook: 4.5 Years
- Twitter: 5 Years
- Spotify: 11 Years
- Netflix: 18 Years
2023b: GPT-4

- Released in March
- Significant boost in performance veiled in secrecy
- Multimodal: image inputs + image generation through dall-e
- Unofficial/leaked/rumoured details:
  - about 1.8T parameters, 120 layers, (x10 GPT-3)
  - 16 MLP-Experts each with ~111B parameters
  - 13T token training data
2023a: Chatbots!

- GPT-Turbo ChatGPT, March
- Claude 2, July
- Gemini, December
2023c: Competitive Open Weights LLMs

- Llama 1-65B, February
- MPT-30B, March
- Falcon-40B, June
- Llama 2-70B, July
- CodeLlama-70B, August

<table>
<thead>
<tr>
<th>LLM</th>
<th>Humanities</th>
<th>STEM</th>
<th>Social Sciences</th>
<th>Other</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPT</td>
<td>7B</td>
<td>26.7</td>
<td>25.3</td>
<td>27.1</td>
<td>28.2</td>
</tr>
<tr>
<td></td>
<td>30B</td>
<td>44.5</td>
<td>39.0</td>
<td>52.8</td>
<td>52.9</td>
</tr>
<tr>
<td>Falcon</td>
<td>7B</td>
<td>26.4</td>
<td>26.2</td>
<td>24.7</td>
<td>27.4</td>
</tr>
<tr>
<td></td>
<td>40B</td>
<td>49.3</td>
<td>45.5</td>
<td>65.4</td>
<td>65.0</td>
</tr>
<tr>
<td>LLAMA 1</td>
<td>7B</td>
<td>34.0</td>
<td>30.5</td>
<td>38.3</td>
<td>38.1</td>
</tr>
<tr>
<td></td>
<td>13B</td>
<td>45.0</td>
<td>35.8</td>
<td>53.8</td>
<td>53.3</td>
</tr>
<tr>
<td></td>
<td>33B</td>
<td>55.8</td>
<td>46.0</td>
<td>66.7</td>
<td>63.4</td>
</tr>
<tr>
<td></td>
<td>65B</td>
<td>61.8</td>
<td>51.7</td>
<td>72.9</td>
<td>67.4</td>
</tr>
<tr>
<td>LLAMA 2</td>
<td>7B</td>
<td>42.9</td>
<td>36.4</td>
<td>51.2</td>
<td>52.2</td>
</tr>
<tr>
<td></td>
<td>13B</td>
<td>52.8</td>
<td>44.1</td>
<td>62.6</td>
<td>61.1</td>
</tr>
<tr>
<td></td>
<td>34B</td>
<td>59.4</td>
<td>52.1</td>
<td>71.8</td>
<td>69.2</td>
</tr>
<tr>
<td></td>
<td>70B</td>
<td>65.0</td>
<td>58.0</td>
<td>80.3</td>
<td>74.6</td>
</tr>
</tbody>
</table>
2023c: Competitive Open Weights LLMs

- Llama 1, February 2023
  ~5,500 citations
- Llama 2, July 2023
  ~4,300 citations
  7B: ~1.3M downloads
  70B: ~380k downloads
- 7,000 GitHub projects mentioning LLama
2023d: Explosion in LLM Projects/Research

- Agents using software tools
- Increasingly sophisticated prompting techniques, e.g. tree of thoughts
2023d: Explosion in LLM Projects/Research

- Agents using software tools
- Increasingly sophisticated prompting techniques, e.g. tree of thoughts
- New alignment methods leads to many Llama derivatives:

Stanford Alpaca  Koala 13B  Vicuna  Orca  WizardLM  Llava

ETH AI CENTER
2024a: Next Generation General Purpose Assistants

- Multimodal (Voice, Video, Image)
- Long context (100k-1M tokens)
- Memory
- Strong coding and multilingual
- Tool use / execution environment
- Websearch / document upload

March 2024

February 2024

November 2023

March 2024

yesterday

literally now?
2024b: Open Source/Weights Catching Up

- Grok-1, 314B, March
- DBRX, 132B, March
- Mixtral 8x22B, April
- Llama 3, 70B, April

Labonne (2024)
2024b: Open Source/Weights Catching Up

- Grok-1, 314B, March
- DBRX, 132B, March
- Mixtral 8x22B, April
- Llama 3, 70B, April

Inference cost matters!

1. 15T (!) tokens
2. Beats/competitive with Gemini Pro 1.5 and Claude 3 Sonnet
2024b: Open Source/Weights Catching Up

- Grok-1, 314B, March
- DBRX, 132B, March
- Mixtral 8x22B, April
- Llama 3, 70B, April
2024b: Open Source/Weights Catching Up

- Grok-1, 314B, March
- DBRX, 132B, March
- Mixtral 8x22B, April
- Llama 3, 70B, April
- DeepSeek-V2 67B, May
2024b: Open Source/Weights Catching Up!

- Grok-1, 314B, March
- DBRX, 132B, March
- Mixtral 8x22B, April
- Llama 3, 70B, April
- DeepSeek-V2 67B, May
- Llama 3 405B, June?
Exciting Trend 1/6: MoE & Adaptive Computation
Exciting Trend 2/6: The Return of Recurrent Nets

Associative RNN cell allows us to “scan” the sequence

(a) Up: parent combines values of its children

(b) Down: right child combines statistics of its parent with the left sibling.

Blelloch (1990)
Exciting Trend 2/6: The Return of RNNs

Associative RNNs increasingly competitive with Transformers

(a) Scaling curve during training  
(b) Maximum throughput at 1B parameter scale.

De et al. (2024)
Exciting Trend 3/6: Low Precision & New Hardware

Precision
- fp16 (standard) -> fp8 training (ongoing) -> fp4 (soon)
- Inference with 4 bits, 2 bits, teneray (-1, 0, +1)

Hardware
- Specialised hardware for cheaper inference (groq.com)
- custom silicon (Apple, Meta, Google, Amazon, …)
- Nvidia Blackwell (FP4)
- Local LLMs -> embedded devices
Exciting Trend 4/6: Quality Data & Synthetic Data

- High-quality public data for pretrain: FineWeb (15T tokens; webcrawl)

- Synthetic data for reasoning and alignment
  1. Generate (E-step): The language model generates multiple output samples for each input context. Then, we filter these samples using a binary reward to collect the training dataset.
  2. Improve (M-step): The original language model is supervised fine-tuned on the training dataset from the previous Generate step. The fine-tuned model is then used in the next Generate step.

Singh et al. (2024)
Exciting Trend 5/6: Better Alignment

- RLHF: aligning LLMs via reinforcement learning with human feedback

- Alignment across modalities: images, videos, audio
Exciting Trend 6/6: System Interactions & Agents

- Agents interacting / Tool use
- Web/Database Search & RAG: Retrieval Augmented Generation
- Coding & execution environment
Swiss AI Initiative: swiss-ai.org

- National Research Initiative jointly lead by ETHZ and EPFL
- Scientific Council: 26 professors / researchers
- Assembly: >100 researchers
- 10M GPU hour commitment on Alps
Swiss AI Initiative: swiss-ai.org

Alps Supercomputer: 10’000 GH200 GPUs

<table>
<thead>
<tr>
<th>Rank</th>
<th>System</th>
<th>Cores</th>
<th>Rmax (PFlop/s)</th>
<th>Rpeak (PFlop/s)</th>
<th>Power (kW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Alps - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE, Swiss National Supercomputing Centre (CSCS), Switzerland</td>
<td>1,305,600</td>
<td>270.00</td>
<td>353.75</td>
<td>5,194</td>
</tr>
</tbody>
</table>
Swiss AI Initiative: swiss-ai.org
Swiss AI Initiative: swiss-ai.org

LLM Area:

- An LLM for Switzerland
- Trustworthy and Responsible
- Transparent and compliant (open source / open weights)
- Multilingual with Swiss societal values

- Attract and develop talent
- Startup fuel
- Teaching and sharing lessons, code, models, …

- Collaborations: users, developers, legal, …
Questions?

Thank you for your attention.

Feel free to get in touch:

Imanol Schlag
ischlag@ethz.ch