# Directions and Challenges in Generative Modeling

Christian Szegedy
xAI

# Generative Language Modeling

OpenAI/ChatGPT

Google/Bard

Inflection

Anthropic

Microsoft/Bing

Hugging Face

Personal assistants, text/code generators.

All are using LLMs (large language models) with 10-100s of Bn parameters trained on most of the public data on the internet.

# Challenges

- **Staleness**: past LLMs could only argue about information in their training set
- Running out of **training data**
- **Correctness**: Is reasoning any good?
  - No guarantees of correctness
  - No references
- **Textual input** is restrictive
- **Textual output** is restrictive
- For some tasks (**symbolic computation**) neural networks are inefficient

# Current Trends

- **Utilizing human feedback**
  - Fine-tuning on human feedback
  - Constitutional AI (train an extra model)
- **Retrieval Augmentation**
  - Can update their knowledge in real time
  - Refer to facts after their training.
- **Multimodality**
  - Language models the can interpret other media (images, spoken text, music, video,...)
  - Generating other modalities
- **Generate own training data**
  - Tool use (programming environment, reasoning tools, external databases,..)
  - Feedback loops (hand-engineered synthetic tasks, self-critique,...)

# Future Directions

- Synthetic data generation will hinge on inference throughput.
    - Autoregressive inference is slower than feed-forward pass through the network.
    - Distillation?
    - Diffusion for text?
    - Specialist models?
    - Generate data for retrieval?
- Are 100s of billions of parameters really necessary?
    - … for each and every token?
- Self-improvement.
    - Should the system learn from its own mistakes? … Automatically?
    - Self-reflection, self-critique
    - Run the generated code, debug its own output
    - Create provably correct output.