



Let's brainstorm how to train a “better” chatbot than ChatGPT

hululu.zhu@gmail.com

May 2023

github.com/hululuzhu/better-chatbot-than-chatgpt

Disclaimer

- This talk is my personal voluntary effort, prepared and conducted during my personal time outside of working hours.
- All content is derived from publicly available sources, and the views expressed herein only represent my personal opinions, and do not reflect the positions of Google.

hululu.zhu@gmail.com

May 2023

Agenda

First, get closer to ChatGPT

- Solid Pretrained models
- Mimic ChatGPT or Self-Align

Possibly Surpass ChatGPT in Selected Angle(s)?

- More knowledge in a subdomain
- Longer context, even longer than GPT4
- Lower cost of training and inference
- Reward Model(s) and Reinforcement Learning (RL)
- More modalities (e.g. vision, audio) than GPT4?

Welcome interruptions and discussion any time

Split to 2 sessions with 5 mins break in the middle



Before we delve into details, any thoughts on agenda?

Get Closer first

- Pretrained models
- Mimic ChatGPT or Self-Align

Surpass?

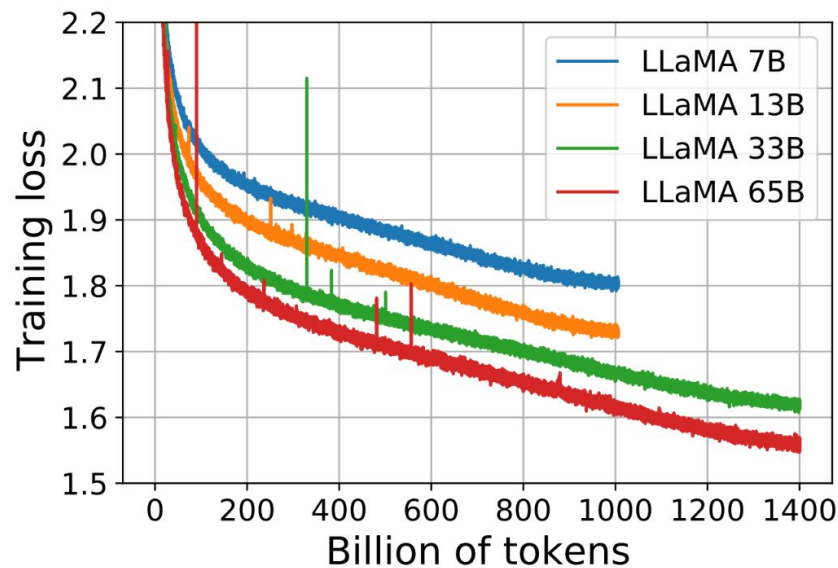
- Subdomain knowledge
- Longer context
- Lower cost of training and inference
- Reward Model(s) and Reinforcement Learning
- More modalities

Any thoughts? Open mic time

Get Closer to ChatGPT - Pretrained Models

LLaMa by Meta AI

- Released Feb 2023, **research only** use (in theory, cannot be used for commercial purposes)
- 7B, 13B, 33B & 65B, best **pretrained LLMs** of its size class until 05/20/2023
 - 65B LLaMa is better than GPT3 175B
- Best architectures
 - Pre-normalization [GPT3]
 - SwiGLU activation and RoPE [PaLM]
- 1T+ tokens for training!
- Max 2048 context length
 - 7B has 512 context length



Get Closer to ChatGPT - Pretrained Models

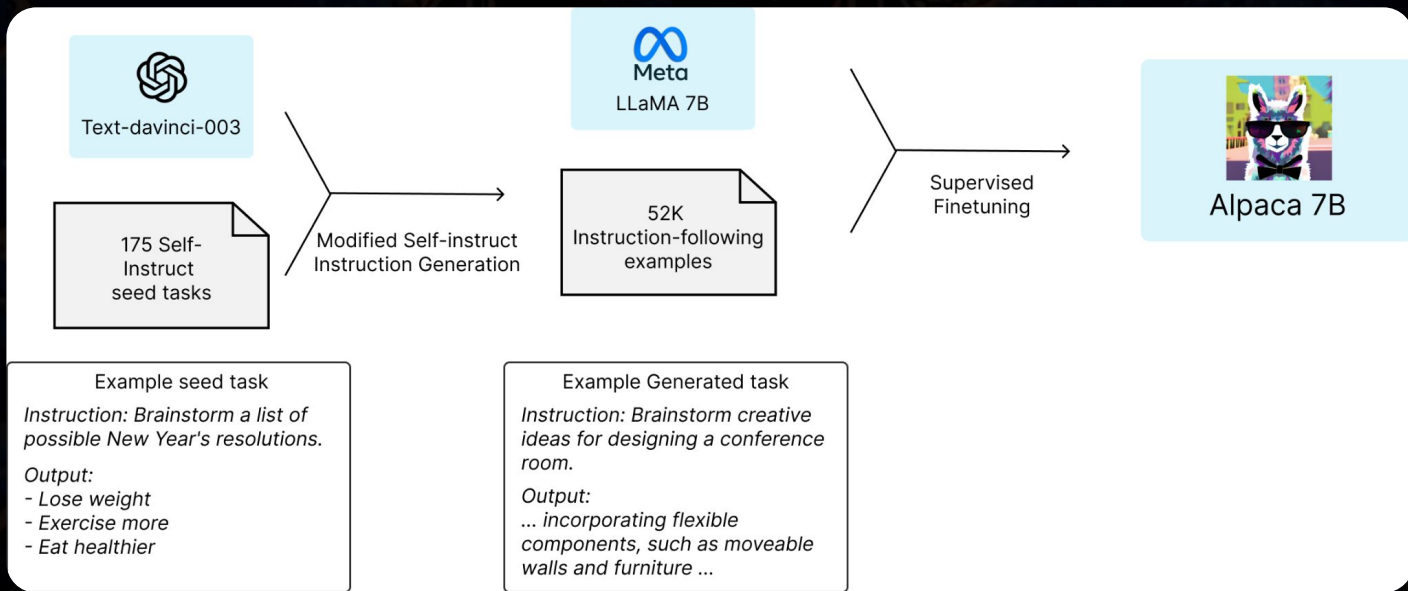
Other candidates

- [ChatGLM](#)-6B (finetuned [GLM](#)) by Tsinghua Univ, English+Chinese
 - Not for commercial
 - The most powerful [GLM130B](#) trained on 400 billion tokens ($\leq 40\%$ than LLaMa)
 - Mixed masked token predication and next token prediction training objectives
- [MPT-7b](#) by mosaicml.com
 - Commercial ok
 - Not best quality, but there is a 65k context length version! (2x context length than GPT4)
- [RedPajama](#) (reproduce LLaMa)
 - Commercial ok
 - Still training in progress, promising to be the best free candidate soon!
 - The preview one is close to LLaMa, 3b and 7b [released here](#)
- [WizardLM](#), [Pythia](#), and so on

Get Closer to ChatGPT - Mimic ChatGPT

Stanford Alpaca (~70% chatgpt)

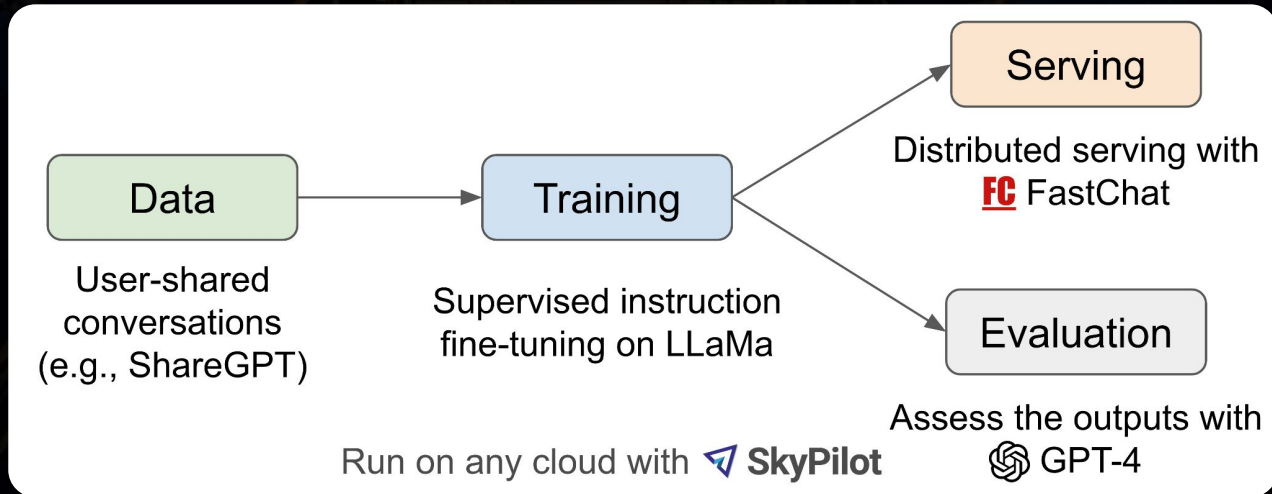
- Use GPT3.5 API as Oracle
- Sample questions (prompts) and sample answers
- Use the GPT3.5 data to finetune a 7B and 13B LLaMa
- Whole process only takes a couple of days and \$600!



Get Closer to ChatGPT - Mimic ChatGPT

Berkeley Vicuna (92% chatGPT)

- Rely on user-shared (selected high quality) conversations (with ChatGPT)
- Evaluation is through GPT-4 (***treating GPT-4 as human labeler***)



Get Closer to ChatGPT - Self Align

Dromedary (Self-Align)

- **To be verified by industry and academia**
- No dependency on ChatGPT API or data
- Starts with LLaMA-65b
- Similar approach as Alpaca to do seed prompts
- Similar to Constitutional AI to apply “Principle” alignment with 5-shot prompt
- Finetune by pruning principles out
- Make responses more verbose
- Claims on par with ChatGPT(GPT 3.5)



(Topic-Guided Red-Teaming) Self-Instruct

195 seed prompts

w/ **7** rules for new instruction generation



360k synthetic prompts



Principle-Driven Self-Alignment

16 principles for AI assistant to follow

w/ **5** in-context learning demonstrations



260k (after filtering) self-aligned responses
to synthetic prompts



(non-verbose)



360k self-aligned & verbose (by prompting) responses
to synthetic prompts



(final)

Verbose Cloning

Refining the model to produce in-depth and detailed responses

More knowledgeable than ChatGPT - Domain Knowledge

Codex: Coding on top of GPT3

- **175GB** github code finetuned on GPT3 (various sizes)
 - No quality difference observed using pretrained GPT3 or from scratch, but pretrained helps converging faster
- *Repeated sampling from the model is effective for producing working solutions*

Minerva: Math on top of PaLM

- **118GB** [...] scientific papers from arXiv [...] that contain mathematical expressions using LaTeX, MathJax
- *few-shot prompting*, *chain of thought* or *scratchpad prompting*, and *majority voting*, to achieve state-of-the-art performance

More knowledgeable than ChatGPT - Retrieval

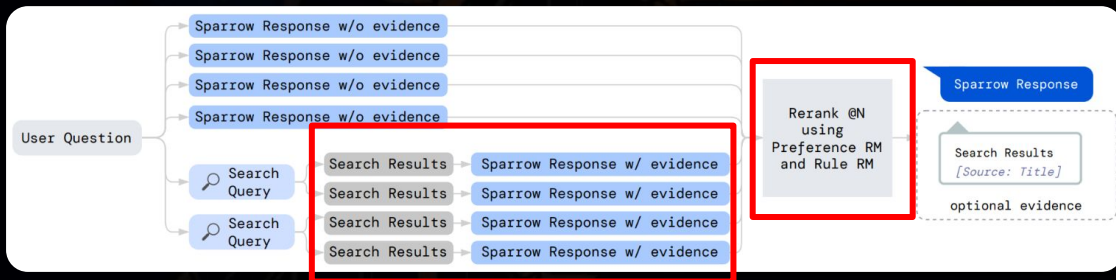
Retrieval in LM training

- [WebGPT](#): “allows the model to search and navigate the web”
 - Behavior cloning (BC)
 - Reward modeling (RM, for ELO)
 - Reinforcement learning (RL)
 - Rejection sampling (best-of-n)
- [Sparrow](#): “an information-seeking dialogue agent”
 - Search Results from Google and Reranker

Retrieval outside LM training

- Embedding similarity retrieval like [LangChain](#)
- But sometimes questions and answers may have different embedding spaces, [DPR](#) claimed better in Q&A (chatbot-like) scenarios

“Retrieval” is just one of the LLM “Tooling”



Longer context than GPT4 (32k tokens) - ALiBi

Attention with Linear Biases (ALiBi): Train Short, Test Long

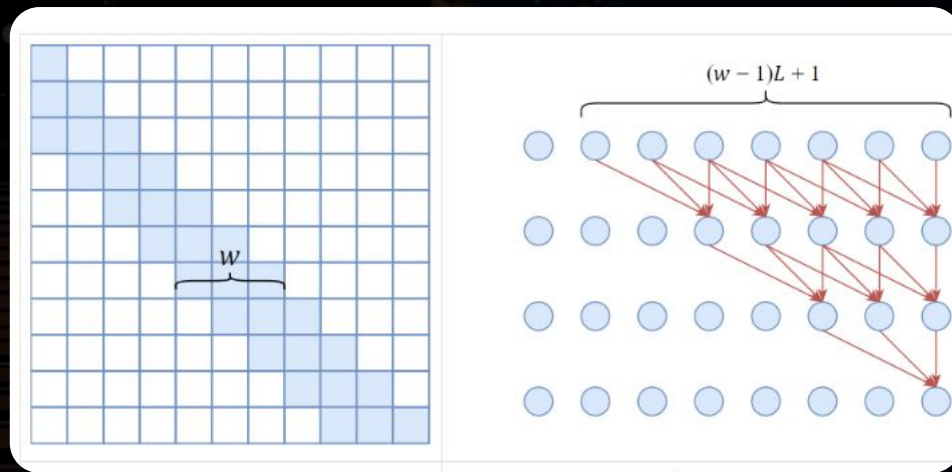
- Technically there is no max of context length given unlimited memory
- In practice, the “position encoding” for untrained positions is often bad
 - Thus bad context generalization
- In short:
 - Original: $\text{softmax}(\mathbf{QK}^T)\mathbf{V}$
 - ALiBi: $\text{softmax}(\mathbf{QK}^T - \lambda|m-n|)\mathbf{V}$
- MPT-7B Storywriter claimed to use ALiBi to provide 65k tokens
- Still to be verified by industry (*otherwise, why does GPT4 have 32k?*)

Longer context than GPT4 (32k tokens) - HWFA

Hybird Window-Full Attention (HWFA)

by Su, Jianlin (author of RoPE)

- “Window” (local) Attention + RoPE position encoding except last layer
- Last layer use original full attention with logn adjustment (why logn)
- $(w-1)L+1 = \alpha N$ ($0 < \alpha \leq 1$)
 - α suggested $\frac{3}{4}$
 - N is the training length (not target length which would be much longer)
- Actual usefulness to be verified



Longer context than GPT4 (32k tokens) - Other related

- KERPLE: [Kernelized Relative Positional Embedding for Length Extrapolation](#)
- Sandwich: [Receptive Field Alignment Enables Transformer Length Extrapolation](#)
- XPOS: [A Length-Extrapolatable Transformer](#)
- Meta: [Recurrent Memory to Extend the Model's Context Length to 1 million!](#)
- More questions
 - How does [Claude \(Anthropic\) achieve 100k tokens?](#)
 - How does GPT4 achieve [32k Tokens?](#)

More efficient training/serving - Multi-query attention

Multi-Query Attention from [Transformer](#)

$Q = \text{tf.einsum} ("bnd, hdk \rightarrow bhnk", X, P_q)$

$K = \text{tf.einsum} ("bmd, hdk \rightarrow \textcolor{red}{bhmk}", M, P_k)$

$V = \text{tf.einsum} ("bmd, hdv \rightarrow \textcolor{red}{bhmv}", M, P_v)$

Multi-Head Attention by Noam Shazeer

$Q = \text{tf.einsum} ("bnd, hdk \rightarrow bhnk", X, P_q)$

$K = \text{tf.einsum} ("bmd, dk \rightarrow \textcolor{red}{bmk}", M, P_k)$

$V = \text{tf.einsum} ("bmd, dv \rightarrow \textcolor{red}{bmv}", M, P_v)$

Difference: Reuse the Q/K projections for each “query” attention

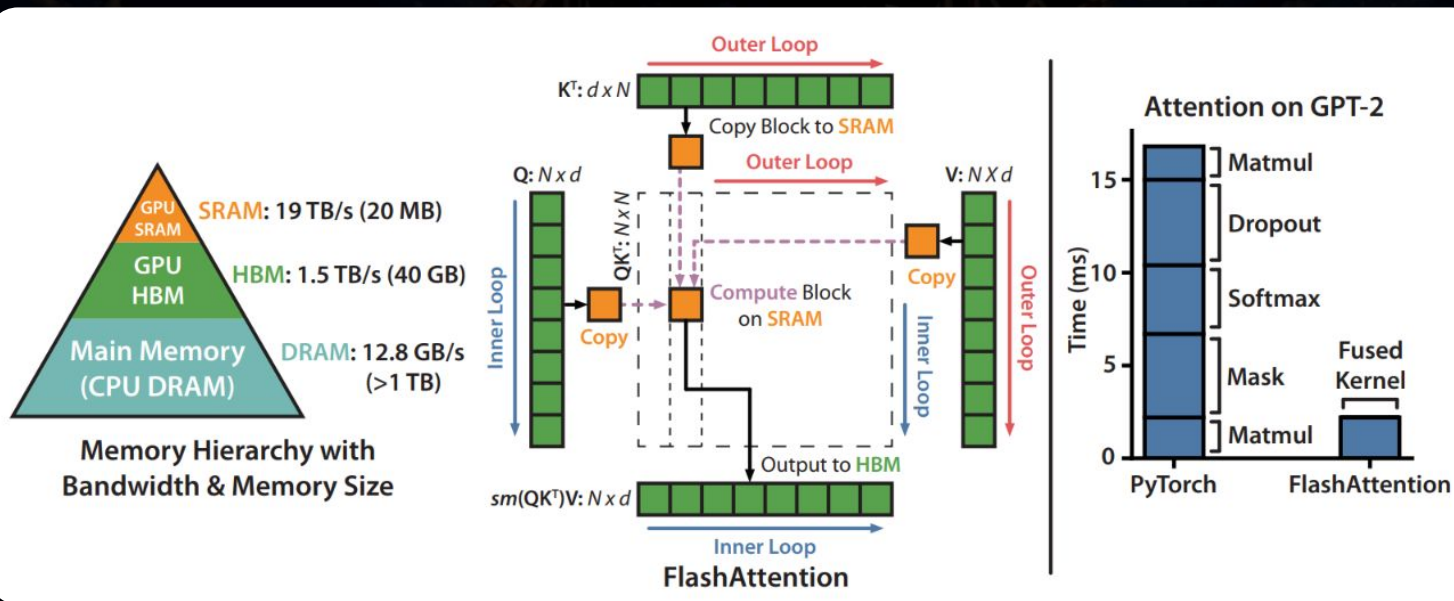
Result: Significantly reduced memory bandwidth requirements of incremental decoding, thus improved the incremental inference speed by 10x!

Used in [PaLM paper](#) (540B model)

More efficient training/serving - FlashAttention

IO-Aware Fast and Memory-Efficient Exact Attention by Stanford

- 2-4 speedup on training! Even with more tflops!
- Tiling to rely more on fastest SRAM and read less from HBM



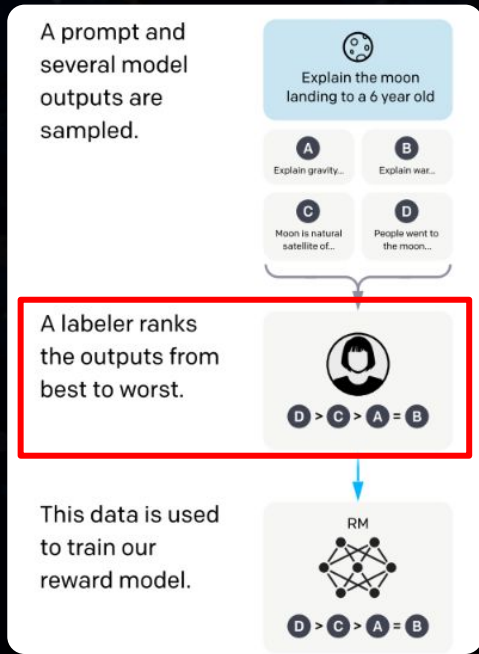
More efficient training/Serving - Other related work

- [FastTranformer](#) by NVidia (inference only)
- [PEFT](#) by HuggingFace (training only)
- [DeepSpeedChat](#) by Microsoft
- [ColossalAI](#) (the founder, Yang You, also invented [LAMB optimizer](#))

Better Reward Model(s) and Reinforcement Learning (RL)

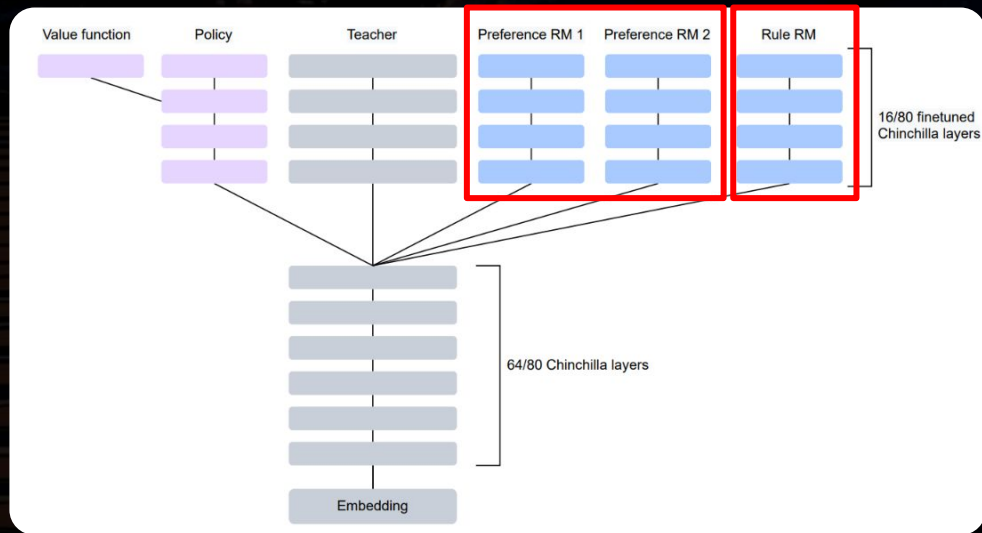
Simple Preference model

- [InstructGPT/ChatGPT](#)
- [Anthropic](#)



Preference model

- [Sparrow](#)



Why is Reinforcement Learning (RL) important?

[John Schulman](#) (ChatGPT architect, [PPO/TRPO](#) inventor) [Berkeley talk](#) (20:51)

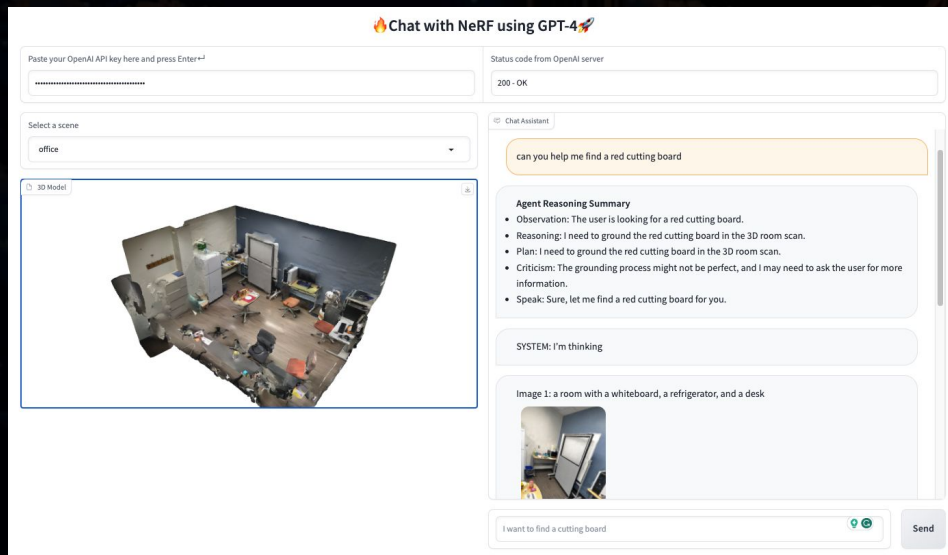
How to Fix with RL

- 1) Adjust output distribution so model is allowed to express uncertainty, challenge premise, admit error. (Can use behavior cloning.)
- 2) Use RL to precisely learn behavior boundary.
 - $\text{Reward}(x) = \{$
 - 1 if unhedged correct (The answer is y)
 - 0.5 if hedged correct (The answer is likely y)
 - 0 if uninformative (I don't know)
 - 2 if hedged wrong (The answer is likely z)
 - 4 wrong (The answer is z) $\}$
 - This reward is similar to log loss, or a proper scoring rule

Slides by John, 20:51 timestamp, fetched 05/21

More modalities than GPT4 (Text + Vision)?

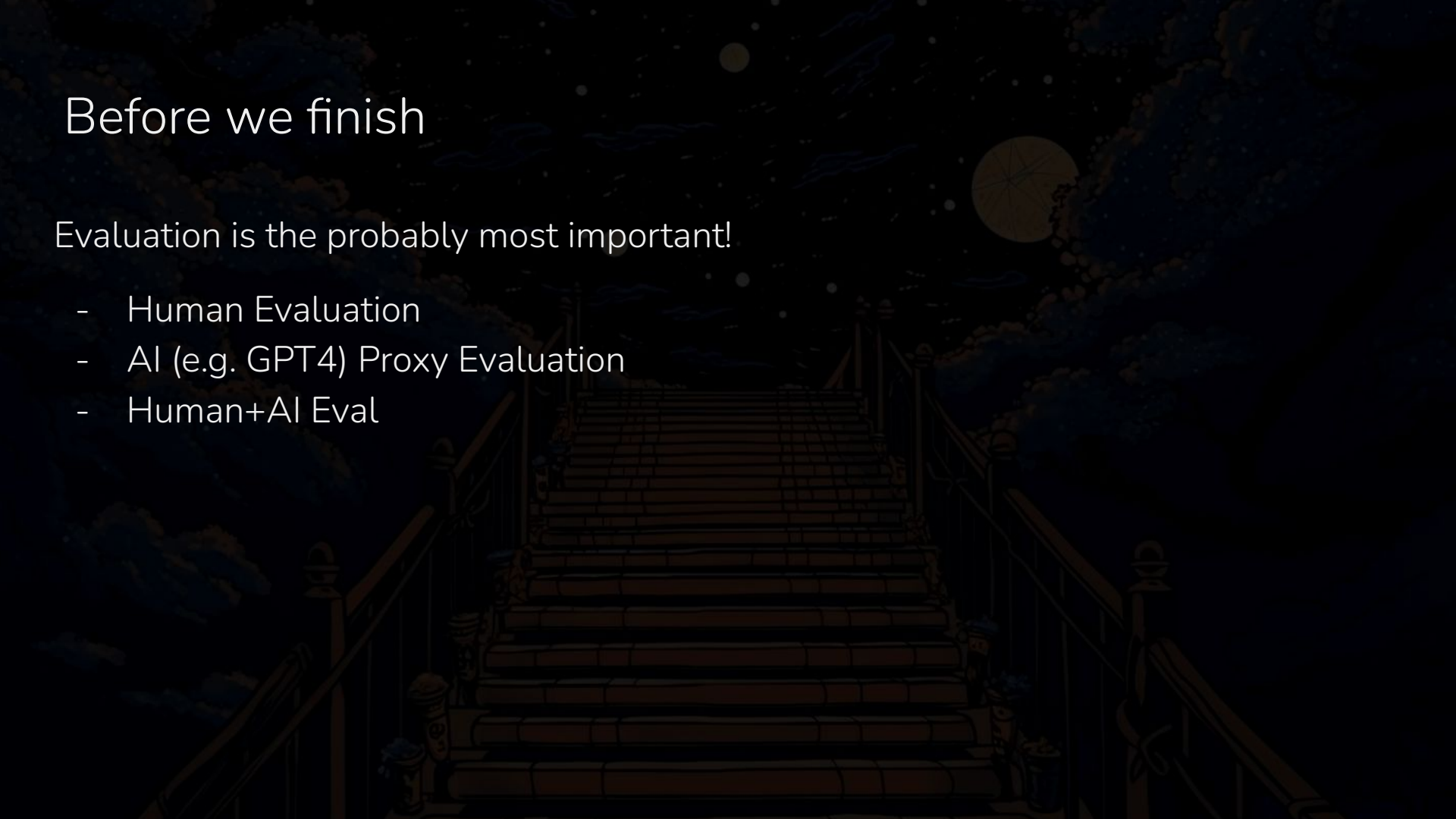
- Multi-modal engineering
 - E.g. [HuggingGPT](#)
- [Blip2](#)-based (*Frozen Image Encoders and Large Language Models*)
 - E.g. [MiniGPT4](#) and [Visual-GLM6B](#)
- 3D! [Chat with NeRF](#): Grounding 3D Objects in Neural Radiance Field through Dialog



Before we finish


Evaluation is the probably most important!

- Human Evaluation
- AI (e.g. GPT4) Proxy Evaluation
- Human+AI Eval



One more thing





fun-ai-talk

Time for more discussion!

Brainstorm: Train a “better” chatbot than ChatGPT

hululu.zhu@gmail.com

May 2023