# WebGPT and Visconde

Thiago Coelho Vieira - thiagotcvieira@gmail.com

# 1.1 main concepts

- **WebGPT**
  - **imitation learning**
    - submits search queries, follows links, and scrolls up and down web pages
  - **human preferences** approach from fine-tuning GPT-2 to produce answers more human-like
  - **long-form question-answering (LFQA)** : paragraph-length answer is generated in response to an open-ended question
  - **factual accuracy** : support answers references (mitigate model "*hallucination*")
  - **demonstrations**: examples of humans using the browser to answer questions
  - **comparisons**: pairs of model-generated answers to the same question, and asked humans which one they preferred to optimize answers quality (*better, worse or equally good overall*)

# 2.1 contributions

- **WebGPT**
  - leverage existing information retrieval (Bing Web Search API) and synthesis (GPT3 fine tunned) to solve LFQA
  - optimize answer quality using human feedback (**comparisons**)
  - provide answers with citation from the articles of the retrieval step as a factual proof
  - developed a end-to-end open-ended Q/A system using a text-based web browser WebGPT Answer Viewer

# 2.2 WebGPT

- GPT-3 to use a text-based web-browser trained on 6000 **demonstrations** and 21500 **comparisons** and 4 different training methods (BC, RM, RL and best-of-$n$)
- the model is provided with an open-ended question and a summary of the browser state and must issue commands (search, find in page, quote, scroll...)
- retrieved text from hits are used to compose a crafted (text-based web-browser state) prompt for *GPT-3* produce the answers with the citations
- trained a reward model to predict human preferences, and optimizing against it using either reinforcement learning or rejection sampling (*demonstrations* and *comparisons*)

1. **Behavior cloning (BC).** We fine-tuned on the demonstrations using supervised learning, with the commands issued by the human demonstrators as labels.
2. **Reward modeling (RM).** Starting from the BC model with the final unembedding layer removed, we trained a model to take in a question and an answer with references, and output a scalar reward. Following Stiennon et al. [2020], the reward represents an Elo score, scaled such that the difference between two scores represents the logit of the probability that one will be preferred to the other by the human labelers. The reward model is trained using a cross-entropy loss, with the comparisons as labels. Ties are treated as soft 50% labels.
3. **Reinforcement learning (RL).** Once again following Stiennon et al. [2020], we fine-tuned the BC model on our environment using PPO [Schulman et al., 2017]. For the environment reward, we took the reward model score at the end of each episode, and added this to a KL penalty from the BC model at each token to mitigate overoptimization of the reward model.
4. **Rejection sampling (best-of-*n*).** We sampled a fixed number of answers (4, 16 or 64) from either the BC model or the RL model (if left unspecified, we used the BC model), and selected the one that was ranked highest by the reward model. We used this as an alternative method of optimizing against the reward model, which requires no additional training, but instead uses more inference-time compute.

# 3.1 interesting/unexpected results

- **WebGPT**
  - retrieve 64 hits (Bing)
  - fine tuned *GPT-3* 760M, 13B and 175B
  - post-processing dataset, demonstration interface, comparison denoising (appendixes)
  - use moslty *ELI5* dataset of open-ended questions scraped from the "Explain Like I'm Five" subreddit
  - answers are factually accurate as those written by our human demonstrators
  - best model produced answers that were **prefered 56%** of time against the ones written by humans.
  - do not work very well in *OOD* questions from *ELI5* dataset
  - *TruthfulQA* an adversarially-constructed dataset of short-form questions (scored on *truthfulness* and *informativeness*)
  - model answers are **true 75% of the time**, and are **both true and informative 54% of the time**

# 4. basic doubts

- **WebGPT**
  - what are the human *demonstrations*? examples of humans using the browser to answer questions
  - what passages from the site are picked? simply picked passages with the term on it
  - how to cherry-pick the best sources? (Bing, Google, private IR)

# 5. advanced topics

- **WebGPT**
  - bring light to the challenges and risks of allowing general-purpose AI system like this to access and work directly on information from the web.
    - what sources are reliable
  - it's something like plugins of the ChatGPT and AgentGPT does nowadays
  - answers with citations can obscure the fact that our model still makes basic errors