Title: Sparrow (chatbot)

URL: https://en.wikipedia.org/wiki/Sparrow_(chatbot)

PageID: 72750759

Categories: Category:Chatbots, Category:Google DeepMind, Category:Language modeling, Category:Large language models, Category:Natural language processing

---

Sparrow is a chatbot developed by the artificial intelligence research lab DeepMind , a subsidiary of Alphabet Inc. It is designed to answer users' questions correctly, while reducing the risk of unsafe and inappropriate answers. [1] One motivation behind Sparrow is to address the problem of language models producing incorrect, biased or potentially harmful outputs. [1] [2] Sparrow is trained using human judgements, in order to be more “Helpful, Correct and Harmless” compared to baseline pre-trained language models. [1] The development of Sparrow involved asking paid study participants to interact with Sparrow, and collecting their preferences to train a model of how useful an answer is. [2]

To improve accuracy and help avoid the problem of hallucinating incorrect answers, Sparrow has the ability to search the Internet using Google Search [1] [2] [3] in order to find and cite evidence for any factual claims it makes.

To make the model safer, its behaviour is constrained by a set of rules, for example "don't make threatening statements" and "don't make hateful or insulting comments", as well as rules about possibly harmful advice, and not claiming to be a person. [1] During development study participants were asked to converse with the system and try to trick it into breaking these rules. [2] A 'rule model' was trained on judgements from these participants, which was used for further training.

Sparrow was introduced in a paper in September 2022, titled "Improving alignment of dialogue agents via targeted human judgements"; [4] however, the bot was not released publicly. [1] [3] DeepMind CEO Demis Hassabis said DeepMind is considering releasing Sparrow for a "private beta" some time in 2023. [4] [5] [6]

Training

Sparrow is a deep neural network based on the transformer machine learning model architecture. It is fine-tuned from DeepMind's Chinchilla AI pre-trained large language model (LLM), [1] which has 70 Billion parameters. [7]

Sparrow is trained using reinforcement learning from human feedback (RLHF), [1] [3] although some supervised fine-tuning techniques are also used. The RLHF training utilizes two reward models to capture human judgements: a “preference model” that predicts what a human study participant would prefer and a “rule model” that predicts if the model has broken one of the rules. [3]

Limitations

Sparrow's training data corpus is mainly in English, meaning it performs worse in other languages. [citation needed]

When adversarially probed by study participants it breaks the rules 8% of the time; [2] however, this is still three times lower than the baseline prompted pre-trained model (Chinchilla).

See also

AI safety

Commonsense reasoning

Ethics of artificial intelligence

Natural language processing

Prompt engineering

References

External links

White paper

Blog post

v

t

e

Google

Google Brain

Google DeepMind

AlphaGo (2015)

Master (2016)

AlphaGo Zero (2017)

AlphaZero (2017)

MuZero (2019)

Fan Hui (2015)

Lee Sedol (2016)

Ke Jie (2017)

AlphaGo (2017)

The MANIAC (2023)

AlphaFold (2018)

AlphaStar (2019)

AlphaDev (2023)

AlphaGeometry (2024)

AlphaGenome (2025)

Inception (2014)

WaveNet (2016)

MobileNet (2017)

Transformer (2017)

EfficientNet (2019)

Gato (2022)

Quantum Artificial Intelligence Lab

TensorFlow

Tensor Processing Unit

Assistant (2016)

Sparrow (2022)

Gemini (2023)

BERT (2018)

XLNet (2019)

T5 (2019)

LaMDA (2021)

Chinchilla (2022)

PaLM (2022)

Imagen (2023)

Gemini (2023)

VideoPoet (2024)

Gemma (2024)

Veo (2024)

DreamBooth (2022)

NotebookLM (2023)

Vids (2024)

Gemini Robotics (2025)

" Attention Is All You Need "

Future of Go Summit

Generative pre-trained transformer

Google Labs

Google Pixel

Google Workspace

Robot Constitution

Category

Commons