Dataset Generation

The new dataset was generated creating 3 types of conversations:

* Human -human
* Human – AI
* AI-AI

The models used for the AI are the state-of-the-art for Spring 2025, when we performed this data collection:

* Claude 3.7 from Anthropic
* Llama 3.3 from Meta
* GPT - 4o from OpenAI

**Generation of AI-AI conversations**

AI-AI conversations were generated with the code in this folder. Each model has its own notebook. Please scroll down to ‘Current Version’ to execute the most recent code.

For each, the word length is sampled from a distribution of typical human lengths, computed on the previous version of this dataset (conversations version 1). If the model generates a sentence that is too long with respect to the sampled distribution, the code checks if there is any truncation element in the sentence (a period, question mark…) and truncates the sentence so that it is lesser than the sampled length.

**Generation of AI-human and human-human conversations**

The human-human and the human-AI conversations were generated using a custom-made app called EsChat. Here, participants were anonymously paired with humans on the platform or AI, they were assigned a topic around which to focus their conversations and they were assigned a length of their conversations: 12 exchanges each (90% probability), or 24 exchanges each (10% proability).

The record of the comprehensive dataset generated in 2025 is reported in the excel sheet in this folder, called **EsChat.xlsx**

To process the dataset and make into the format that then you can see in the mturk code (statit/dataset), please refer to the codes: chunk\_dataset.ipynb