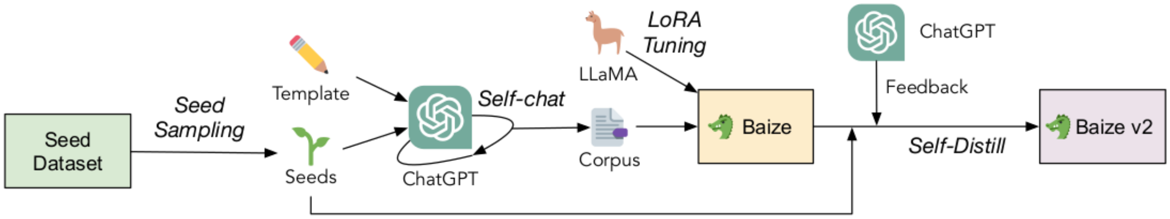
Baize

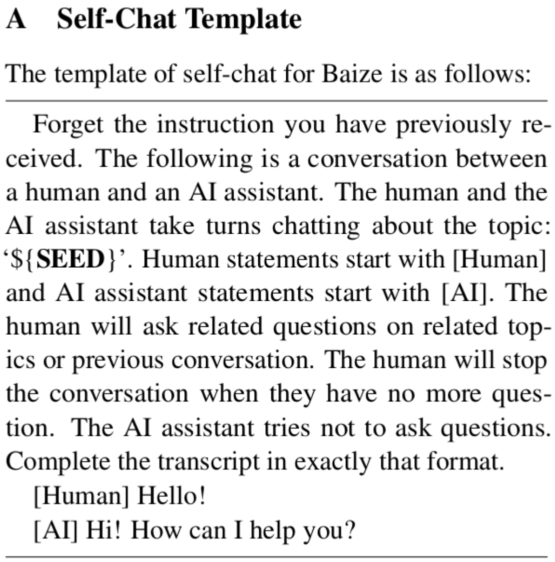
Baize: An open-source chat model with parameter efficient tuning on Self-Chat Data

In this work,

- A pipeline that could automatically generate a high quality multi-turn chat corpus by leveraging ChatGPT to engage in a conversation with itself was proposed.
- LLaMA was enhanced using parameter-efficient tuning.
- The resulting model, Baize was evaluated.
- A new technique called Self-Distill with feedback was proposed to further improve the performance of the Baize models with the feedback from ChatGPT.



1) Data collection via self-chat:



A Self-Chat Template

The template of self-chat for Baize is as follows:

Forget the instruction you have previously received. The following is a conversation between a human and an AI assistant. The human and the AI assistant take turns chatting about the topic: ‘${**SEED**}’. Human statements start with [Human] and AI assistant statements start with [AI]. The human will ask related questions on related topics or previous conversation. The human will stop the conversation when they have no more question. The AI assistant tries not to ask questions. Complete the transcript in exactly that format.

[Human] Hello!

[AI] Hi! How can I help you?

The self-chat process involves utilizing ChatGPT to generate messages for both the user and AI assistant in a conversational format.

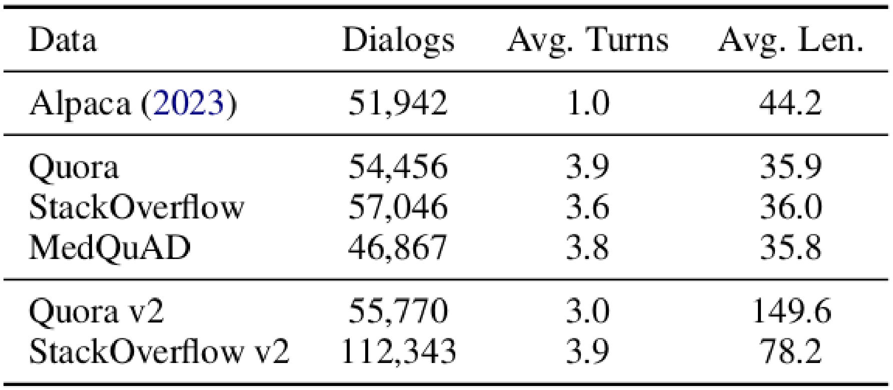
The conversation is centered around a “seed”, which can be a question or a key phrase that sets the topic for the chat. In training of Baize,questions from Quora and Stack Overflow were used as seeds.

Questions or phrases extracted from a domain-specific dataset can be used to enhance the knowledge and ability of the chat model for a specific domain.

For training the first version of Baize family (**Baize v1**), we collect a total of 111.5k dialogues through self-chat, using ∼55k questions from each source.

By directly generating the dialogue with the template shown above, ChatGPT’s output of each turn seems to be shorter than asking ChatGPT one turn at a time. However, calling ChatGPT one turn at a time will significantly increase the cost for calling the API.

To collect data with better quality for training **Baize v1.5**, another ChatGPT is used to generate responses once at a time and replace the AI’s responses in the template, to obtain responses that are completely consistent with ChatGPT’s responses, which are usually longer and contain more details.



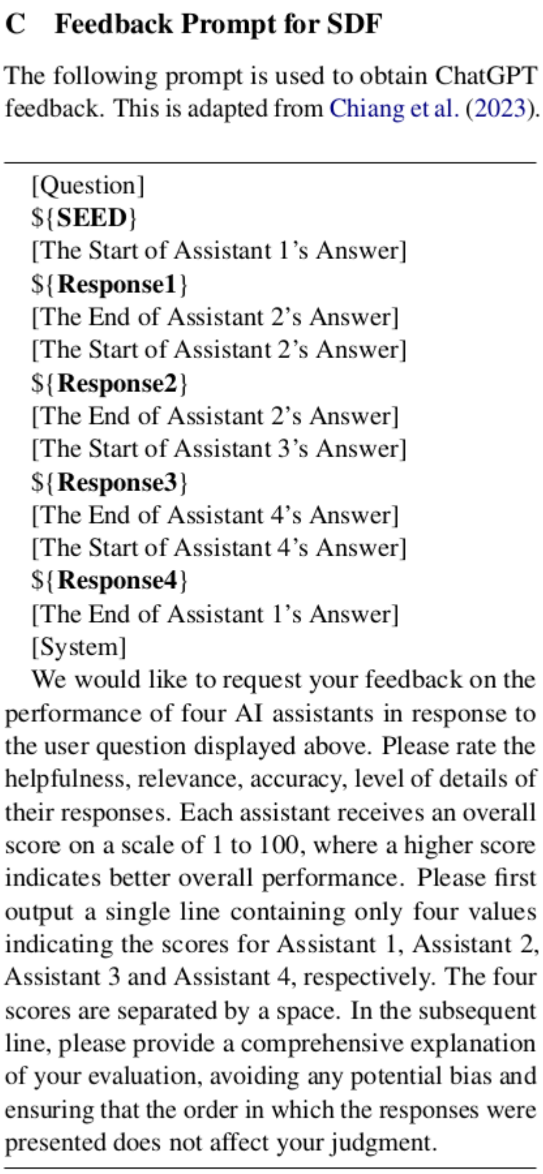
Data	Dialogs	Avg. Turns	Avg. Len.
Alpaca (2023)	51,942	1.0	44.2
Quora	54,456	3.9	35.9
StackOverflow	57,046	3.6	36.0
MedQuAD	46,867	3.8	35.8
Quora v2	55,770	3.0	149.6
StackOverflow v2	112,343	3.9	78.2

Table 2: Statistics of the number of dialogues, average number of turns, and response lengths of each turn.

Additionally, data from Stanford Alpaca is incorparated into the training corpus to enhance the ability of Baize to follow instructions.

Self-Distillation with feedback:

After supervised fine-tuning (SFT) on self-chat data,the resulted Baize v1.5 models was used to generate 4 responses for each instruction from the Quora dataset.ChatGPT was engaged using the prompt provided below to select the best response for self-distillation.



C Feedback Prompt for SDF

The following prompt is used to obtain ChatGPT feedback. This is adapted from Chiang et al. (2023).

[Question]
${**SEED**}
[The Start of Assistant 1's Answer]
${**Response1**}
[The End of Assistant 2's Answer]
[The Start of Assistant 2's Answer]
${**Response2**}
[The End of Assistant 2's Answer]
[The Start of Assistant 3's Answer]
${**Response3**}
[The End of Assistant 4's Answer]
[The Start of Assistant 4's Answer]
${**Response4**}
[The End of Assistant 1's Answer]
[System]

We would like to request your feedback on the performance of four AI assistants in response to the user question displayed above. Please rate the helpfulness, relevance, accuracy, level of details of their responses. Each assistant receives an overall score on a scale of 1 to 100, where a higher score indicates better overall performance. Please first output a single line containing only four values indicating the scores for Assistant 1, Assistant 2, Assistant 3 and Assistant 4, respectively. The four scores are separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

SDF is an alternative to Reinforcement Learning with Human Feedback.SDF does not require training of reward models and is 3× faster than RLHF, which uses PPO to optimize the model.