First meeting.

① Any technical solutions.
Furong said she'll be happy to discuss.

- Federated learning.
(However, it's difficult to set up and also chatGPT is quite stingy with info).

Latent.

some revisions. (on my end ⇒

① Way to perform inference attack is indirect :

① we will query GPT-3 to find out if some datapoints were in its training data.

[Of course we are not expecting a completely honest response.
PROBLEM : There is no actual training dataset available ⇒ figure out false positives, true positives, true negatives, and false negatives.

So: DATA GENERATION STEP.

① Generate a dataset out of the various responses we get from our query.

② Also assemble "likely" data sources for the actual training of chat-GPT:.

→ Wikipedia articles, news articles, other publicly available text sources. (labeled as : in_training_data)

→ Text from unlikely sources like niche domains, recent events (gpt's data is up to 2021), or languages not well represented - (not_in_training data)

2. ATTACKER MODEL DESIGN

(preprocess the dataset we develop and extract the important features).

→ Use it to train a machine learning model.

(Something similar → Here, we can use GPT-2) because the dataset of GPT-2 is very likely the same or inclusive for GPT-3.

Step 3.

Test the performance of attacker model in terms of its accuracy in predicting that a data-point was in chat GPT's training dataset.

Some features we can look at:
- Response length
- Sentiment score
- Response Consistency
- N-gram frequency
- Topic similarity.

} for our analyses of the results.

→ Use performance metrics on our attacker model → precision, recall, and F1 score.

All steps listed above should be completed by our progress report time [Is this feasible?]

- Then the final phase would be:
- trying to improve the results by performing:

① differential privacy. and other useful methods.