



1

Build Your Own Chatbot

LYU2203 Chau Tak Ho (1155143501) Wang Weixiao (1155141608) Supervised by Prof. Michael R. Lyu



Agenda



Summary of 1st Term



Objective of 2nd Term



Experiment & Analysis on ChatGPT

04

In-context Learning



Prompt Design for Historical Bias



Conclusion

Summary of 1st Term

Objective

- Closed-domain Question-Answering
 - Process a question \rightarrow output an answer
 - Focus on a specific domain
 - Advantages :
 - Accurate
 - Efficient

Architecture

- A question decomposition model
 - decompose the compositional questions
- An extractive QA model
 - find the answers to the questions in the documents
- A ranking model

•

 rank all candidate answers and decide the most possible one



Test result

Dataset	Decomposition	Context	Answer
NLP	21 of 28	37 of 41	34 of 41
Bank Service	$5 ext{ of } 5$	89 of 98	86 of 98

Test with 2 datasets extracted manually by documents from different field:

- NLP
- Bank Service

We test whether it can:

- Decompose the sentence correctly.
- Find the context that contains the answer.
- Output the correct format of the answer, especially for those containing terms or special phrases.

Objective of 2nd Term

Objective

Goal: Study chatbots

How do we learn chatbot?

- Learn the theories related to chatbots and build one ourselves (Term 1)
- Analyze and improve existing chatbots (Term 2)
 - Our focus: ChatGPT

ChatGPT

- AI chatbot built on top of OpenAI's GPT-3.5 and GPT-4 families of large language models [1]
- Can generate high-quality, human-like responses for various NLP tasks
- Gains immense popularity due to its ability to produce highquality responses
- Studies revealed its tendency to produce factual and reasoning errors
 - Our focus: QA capability

03

Experiment & Analysis on ChatGPT

Experiment Setup

• Evaluate ChatGPT on five QA benchmarks: SQuAD [2], HotpotQA [3], Common-senseQA [4], TruthfulQA [5], SuperGLUE [6], to test different aspects of reasoning capabilities.

Introductions on Datasets

- SQuAD: a reading comprehension dataset that need to extract the answer from the document.
- HotpotQA: a muti-hop QA dataset that need to generate an answer among many paragraphs.
- SuperGLUE: it contain many tasks. We select 3 of them that related to QA:
 - BoolQ: yes/no question about a paragraph
 - MultiRC: predict whether the possible answer is correct or not.
 - ReCoRD: select one from possible entities to fill the blank.
- CommonsenseQA: select the most reasonable answer to questions about common sense.
- TruthfulQA: select the correct answers from many misleading answers. We only use its MC1 and MC2 dataset, which are single selection and multiple selection.

Experiment Design

- Steps:
 - Randomly pick 100 samples from each datasets
 - Design appropriate prompts to make it better understand the tasks.
 - Summarize failed cases to find its weakness and problems.
 - Analyse the factors which cause these problems and try to avoid them.

Result

Dataset	Metric	SOTA	ChatGPT
SQuAD2.0	EM	90.578	39.0
	F1	92.978	45.35
HotpotQA (Distractor set)	EM	67.46	40.00
	F1	80.52	55.30
SuperGLUE (BoolQ)	Acc	92.4	73.0
SuperGLUE (MultiRC)	EM	65.8	61.90
	F1	89.6	88.421
SuperGLUE (ReCoRD)	EM	95.9	78.00
	F1	96.4	79.73
CommonsenseQA	Acc	78.2	80.0
TruthfulQA(MC1)	Acc	-	69.0
TruthfulQA(MC2)		-	73.9

Observation

- ChatGPT may output different answers to the same question
- ChatGPT always tend to output a long and detailed answers without restrictions
- ChatGPT could remember chat history but may forget it after many turns
- History may affect its output

Improving strategy

- In-context Learning
- Prompt Design for Historical Bias

In-context Learning

In-context Learning

Step [7]:

- Design a demonstration context containing a few examples
- 2. Take the demonstration and a query as the input
- 3. LLMs are responsible for making predictions



In-context Learning

Problem to be focused on:

- Poor performance on unanswerable questions in SQuAD 2.0
- Unable to
 identity unanswerable questions

	EM	F1	Manually
Overall	39.0	45.35	62/100
50 unanswerable	0	0	14/50
50 answerable	78.0	90.70	48/50

Experiment Setup

Step:

- 1. Give an instruction on the problem
- 2. Randomly select 2 samples in training set to be demonstration examples
- 3. Explain the answer in the samples and repeat the instruction again
- 4. Input the question and use "Edit" for inputting next question

I will give you a paragraph and a question. If there is no information about the question in the paragraph, the question is unanswerable. If it is the case, respond with * to indicate unanswerable question.

Okay, I'm ready to help you. Please go ahead.

Example 1:

Paragraph: Spectre (2015) is the twenty-fourth James Bond film produced by Eon Productions. It features Daniel Craig in his fourth performance as James Bond, and Christoph Waltz as Ernst Stavro Blofeld, with the film marking the character's re-introduction into the series. It was directed by Sam Mendes as his second James Bond film following Skyfall, and was written by John Logan, Neal Purvis, Robert Wade and Jez Butterworth. It is distributed by Metro-Goldwyn-Mayer and Columbia Pictures. With a budget around \$245 million, it is the most expensive Bond film and one of the most expensive films ever made. Question: What is the name of the thirty-fourth James Bond film? Answer: *

Example 2:

Paragraph: Spectre (2015) is the twenty-fourth James Bond film produced by Eon Productions. It features Daniel Craig in his fourth performance as James Bond, and Christoph Waltz as Ernst Stavro Blofeld, with the film marking the character's re-introduction into the series. It was directed by Sam Mendes as his second James Bond film following Skyfall, and was written by John Logan, Neal Purvis, Robert Wade and Jez Butterworth. It is distributed by Metro-Goldwyn-Mayer and Columbia Pictures. With a budget around \$245 million, it is the most expensive Bond film and one of the most expensive films ever made. Question: In what year was the thirty-fourth James Bond film produced? Answer: *

In the two examples, the answer to both questions is empty because the question is unanswerable. When the paragraph does not provide information about the question or the question has a false premise, the question is unanswerable. Therefore, when you encounter an unanswerable question, please output "Answer: *" to indicate an unanswerable question.

That's correct. When there is no information in the given paragraph to answer a question or the question is based on a false premise, the answer will be "*".

Result

	Answer include *	Manually
Original	-	14/50
ICL	20/50	24/50
SOTA	-	45/50

Further Exploration

Step:

- 1. Give an instruction on the problem
- 2. Randomly select 2 samples in training set to be demonstration examples
- 3. Explain the answer in the samples and repeat the instruction again
- 4. Input the question and use "Edit" for inputting next question

I will give you a paragraph and a question. If there is no information about the question in the paragraph, the question is unanswerable. If it is the case, respond with * to indicate unanswerable question.

Alright, I'm ready to help! Please provide me with the paragraph and question.

Result

	Answer include *	Manually
Original	-	14/50
ICL	20/50	24/50
ICL (without demonstration examples)	22/50	28/50
SOTA	-	45/50

Possible Reasons

- ChatGPT outputs different answers
- Demonstration examples make ChatGPT forget the instruction
- Design of demonstration examples
 - Number of samples is not enough
 - Better prompt can be used

05

Prompt Design for Historical Bias

What is Historical bias?

• Histories can impact the responses generated by ChatGPT, leading to potential bias.

Dataset used: gsm8k[8]

- contains 8.5K high quality, linguistically diverse grade school math word problems
- most problems are multi-step mathematical reasoning questions created by humans
- test how history impact mathematical reasoning ability of ChatGPT



Ask questions under different histories

History1		History2
Questions		Questions
	2 independent conversation windows	

Recap: ChatGPT may generate different response under the same input

w	What will you experience after doing housework for a long time?select the most appropriate choice. D:get tired A:backache E:satisfaction B:tiredness C:sneezing	
\$	A: backache.	

W What will you experience after doing housework for a long time?select the most appropriate choice. E:backache D:satisfaction B:sneezing A:tiredness C:get tired

The most appropriate choice would be: A: tiredness.

09

09

B

Recap: ChatGPT may generate different response under the same input

- Find a question that ChatGPT may generate correct or wrong answers. Asking questions under to get a "Correct Case" and a "Wrong Case"
- Using "edit" options to keep all followed questions next to the "Correct Case" or "Wrong Case". (So that it won't forget the history)

Edit questions: example

Josh decides to try flipping a house. He buys a house for \$80,000 and then puts in \$50,000 in repairs. This increased the value of the house by 150%. How much profit did he make?

If the repairs increased the value of the house by 150%, then the value of the house after $ilde{D}$ sepairs is 250% (100% original value + 150% increase) of the original value.

To calculate the value of the house after repairs, we can use the following formula:

Value after repairs = Original value + Increase Value after repairs = \$80,000 + (\$80,000 * 150%) Value after repairs = \$80,000 + \$120,000 Value after repairs = \$200,000

Therefore, the house is now worth \$200,000 after the repairs.

Josh invested a total of \$80,000 + \$50,000 = \$130,000 into the house.

To calculate his profit, we need to subtract his initial investment from the final value of the house:

Profit = Final value - Initial investment Profit = \$200,000 - \$130,000 Profit = \$70,000

So, Josh made a profit of \$70,000 from flipping the house.

< 52/52 > 1

Lloyd has an egg farm. His chickens produce 252 eggs per day and he sells them for \$2 per dozen. How much does Lloyd make on eggs per week?



C Regenerate response If Lloyd's chickens produce 252 eggs per uay, unemma week (7 days), they produce 7 × 252 = 0 7

Result

ChatGPT generates more wrong answers after the "Wrong Case" history

History	accuracy
Correct Case	88%
Wrong Case	68%

Further exploration

What if we ask ChatGPT to ignore the history? Example:

< 70/73 >

Ignore the dialog history and answer the question: Dana can run at a rate of speed four times faster than she can walk, but she can skip at a rate of speed that is half as fast as she can run. If she can skip at 3 miles per hour, how many miles can she travel in six hours if she spends one-third of the time running and two-thirds of the time walking?

71/73 >

Dana can run at a rate of speed four times faster than she can walk, but she can skip at a rate of speed that is half as fast as she can run. If she can skip at 3 miles per hour, how many miles can she travel in six hours if she spends one-third of the time running and two-thirds of the time walking?

Result of "Ignore history":

18 of 32 wrong answers becomes correct. Ignore incorrect history could improve its perform ance!

06 Conclusion

Conclusion



Conducted a comprehensive study on chatbots



Built a QA system that is capable of handling questionanswering on professional topics



Studied existing chatbots and evaluated their performance to identify areas for improvement







Thank You!

Reference

[1] "Introducing chatgpt," Introducing ChatGPT. [Online]. Available: https://openai.com/blog/chatgpt.

[2] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for squad," *arXiv.org*, 11-Jun-2018. [Online]. Available: <u>https://arxiv.org/abs/1806.03822</u>.

[3] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning, "Hotpotqa: A dataset for diverse, explainable multi-hop question answering," *ACL Anthology*, 01-Jan-1970. [Online]. Available: <u>https://aclanthology.org/D18-1259/</u>.
[4] A. Talmor, J. Herzig, N. Lourie, and J. Berant, "Commonsenseqa: A question answering challenge targeting commonsense knowledge," *ACL Anthology*. [Online]. Available: <u>https://aclanthology.org/N19-1421/</u>.

[5] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," *arXiv.org*, 08-May-2022. [Online]. Available: <u>https://arxiv.org/abs/2109.07958</u>.

[6] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Superglue: A stickier benchmark for general-purpose language understanding systems," *arXiv.org*, 13-Feb-2020. [Online]. Available:

https://arxiv.org/abs/1905.00537.

[7] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui, "A survey on in-context learning," *arXiv.org*, 08-Feb-2023. [Online]. Available: <u>https://arxiv.org/abs/2301.00234</u>.

[8]] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman, "Training verifiers to solve math word problems,"