

大模型与文本生成

李不绩

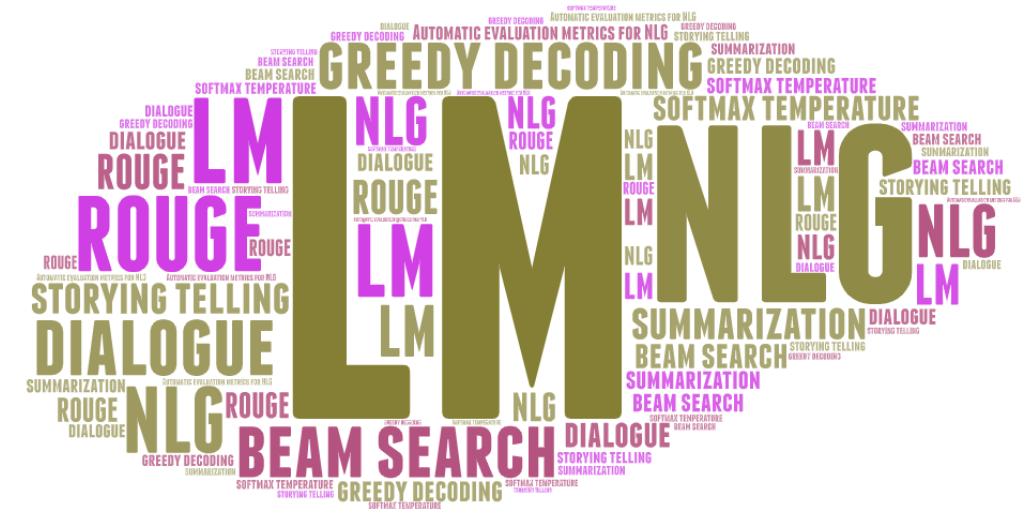
计算机科学与技术学院/人工智能学院

南京航空航天大学

pjli@nuaa.edu.cn

Outline

- Tasks
- Methods
- Experiments
- Future

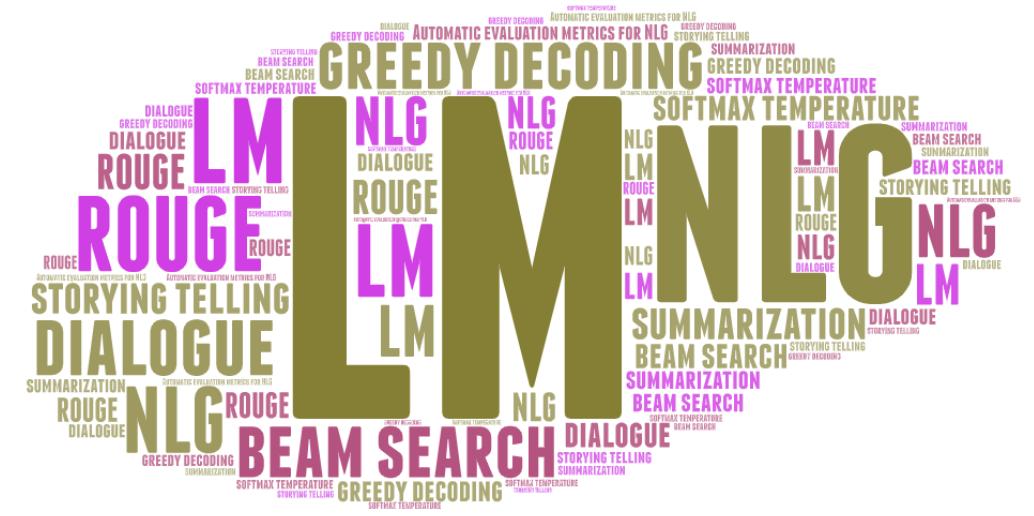


Tasks

- **Summarization**
- **Paraphrasing**
- **Dialogue Generation**
- **Story Generation**
- **Data-to-Text Generation**
- **Multimodal Text Generation**
- **Code Generation**
- **Prompt Generation**
- **Controlled Generation**
- **Data Generation**
- **Evaluation**

Outline

- Tasks
- Methods
- Experiments
- Future



Tasks

- **Summarization**
- **Paraphrasing**
- **Dialogue Generation**
- **Story Generation**
- **Data-to-Text Generation**
- **Multimodal Text Generation**
- **Code Generation**
- **Prompt Generation**
- **Data Generation**
- **Controlled Generation**
- **Evaluation**

Summarization

- via ChatGPT/GPT4
 - Zhang, Haopeng, Xiao Liu, and Jiawei Zhang. "**SummIt: Iterative Text Summarization via ChatGPT.**" *arXiv preprint arXiv:2305.14835* (2023).
 - Zhang, Haopeng, Xiao Liu, and Jiawei Zhang. "**Extractive Summarization via ChatGPT for Faithful Summary Generation.**" *arXiv preprint arXiv:2304.04193* (2023).
 - Ma, Chong, Zihao Wu, Jiaqi Wang, Shaochen Xu, Yaonai Wei, Zhengliang Liu, Lei Guo et al. "**ImpressionGPT: an Iterative Optimizing Framework for Radiology Report Summarization with ChatGPT.**" *arXiv preprint arXiv:2304.08448* (2023).
 - Xiao, Wen, Yujia Xie, Giuseppe Carenini, and Pengcheng He. "**ChatGPT-steered Editing Instructor for Customization of Abstractive Summarization.**" *arXiv preprint arXiv:2305.02483* (2023).
 - Shi, Yiwen, Ping Ren, Jing Wang, Biao Han, Taha ValizadehAslani, Felix Agbavor, Yi Zhang, Meng Hu, Liang Zhao, and Hualou Liang. "**Leveraging GPT-4 for Food Effect Summarization to Enhance Product-Specific Guidance Development via Iterative Prompting.**" *arXiv preprint arXiv:2306.16275* (2023).
 - Wang, Jiaan, Yunlong Liang, Fandong Meng, Zhixu Li, Jianfeng Qu, and Jie Zhou. "**Cross-Lingual Summarization via ChatGPT.**" *arXiv preprint arXiv:2302.14229* (2023).
 - Liu, Yixin, Alexander R. Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. "**On Learning to Summarize with Large Language Models as References.**" *arXiv preprint arXiv:2305.14239* (2023).
 - Laskar, Md Tahmid Rahman, Mizanur Rahman, Israt Jahan, Enamul Hoque, and Jimmy Huang. "**CQSumDP: A ChatGPT-Annotated Resource for Query-Focused Abstractive Summarization Based on Debatepedia.**" *arXiv preprint arXiv:2305.06147* (2023).

Summarization

Turn 1 (task instruction):

Summarize the following text related to food effect studies. {article}

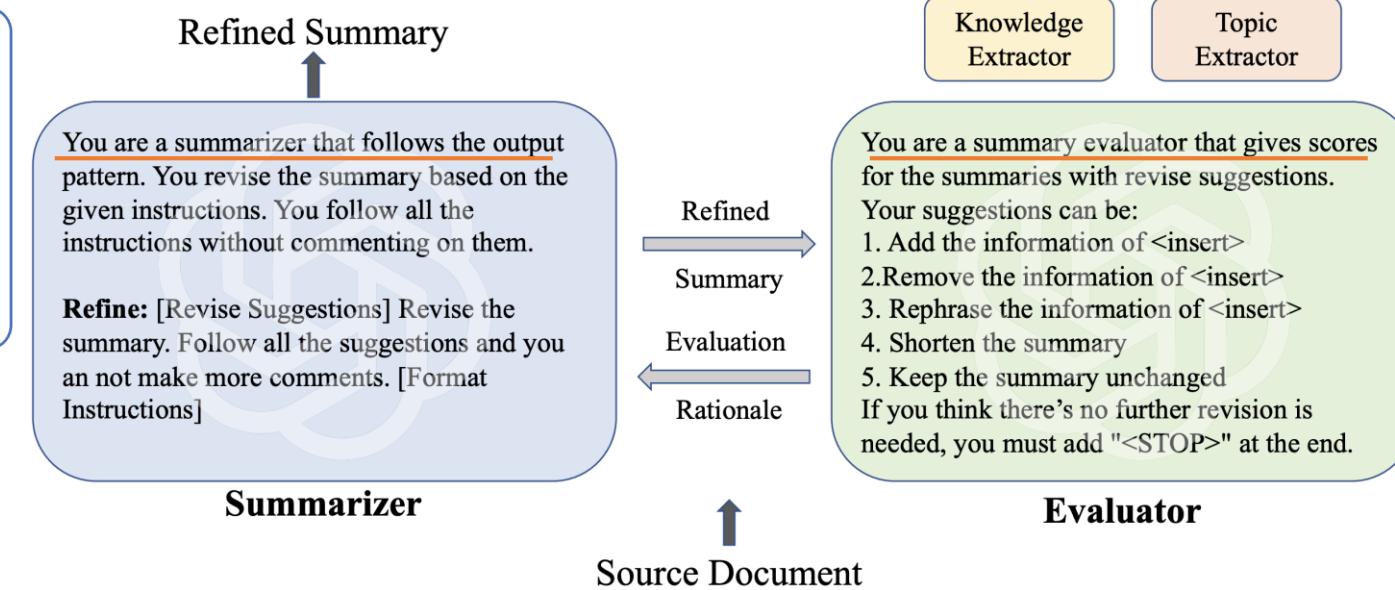
Turn 2 (keyword-constrained prompt):

Add facts of AUC, Cmax, and Tmax in the summary.

Turn 3 (length-constrained prompt):

Summarize it in 2-3 sentences and keep the detail of AUC, Cmax, and Tmax.

Shi, Yiwen, Ping Ren, Jing Wang, Biao Han, Taha ValizadehAslani, Felix Agbavor, Yi Zhang, Meng Hu, Liang Zhao, and Hualou Liang. "Leveraging **GPT-4** for Food Effect Summarization to Enhance Product-Specific Guidance Development via **Iterative Prompting**." *arXiv preprint arXiv:2306.16275* (2023).



Summarization

LLMs are considered the reference or the gold-standard oracle for the summarization task.

| Model | Win | Lose | R1 | R2 | Len. |
|--------------|-----------|-----------|--------------|--------------|------|
| ChatGPT | - | - | 100.0 | 100.0 | 92.0 |
| BART | 11 | 88 | 50.54 | 29.31 | 79.0 |
| GPT3D2 | 21 | 77 | 55.34 | 33.31 | 78.7 |
| GPT3D3 | 34 | 66 | 58.14 | 37.46 | 85.4 |
| Alpaca | 23 | 76 | 53.41 | 31.48 | 81.8 |
| BART.ChatGPT | 36 | 63 | 62.04 | 43.76 | 94.1 |
| BRIO.ChatGPT | 51 | 49 | 61.40 | 40.74 | 93.1 |
| BART.GPT4 | 43 | 56 | 62.08 | 43.55 | 91.8 |
| BRIO.GPT4 | 57 | 42 | 62.79 | 43.65 | 92.8 |

Article: The bigg
been met with con
to cash in their ret
Britons remained
worked, with man
implications...

Summarize the abc
Summary:

Car
The new pension r
confusion among c
restrictions or tax j

Table 2: Results on the test set with **ChatGPT** as the reference LLM and the backbone model of GPTRank. **Win** and **Lose** is the number of times the compared model wins or loses against **ChatGPT** as evaluated by GPTRank (ties ignored). **R1** and **R2** are the ROUGE1/2 F1 scores respectively. **Len.** is the average summary length. **BART.ChatGPT** and **BART.GPT-4** are fine-tuned with MLE training and ChatGPT/GPT-4 as the reference, **BRIO.ChatGPT** and **BRIO.GPT-4** are fine-tuned with contrastive learning (BRIO).

Figure 1: Illustration of GPTRank (ties ignored). **R1** and **R2** are the ROUGE1/2 F1 scores respectively. **Len.** is the average summary length. **BART.ChatGPT** and **BART.GPT-4** are fine-tuned with MLE training and ChatGPT/GPT-4 as the reference, **BRIO.ChatGPT** and **BRIO.GPT-4** are fine-tuned with contrastive learning (BRIO).

Liu, Yixin, Alexander R. Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. "On Learning to Summarize with Large Language Models as References." *arXiv preprint arXiv:2305.14239* (2023).

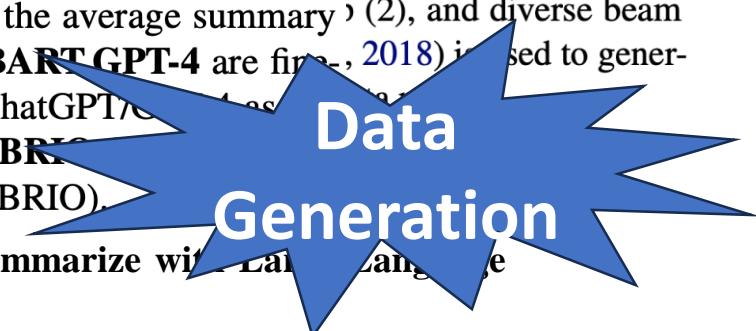
| Model | Win | Lose | R1 | R2 | Len. |
|--------------|-----------|-----------|--------------|--------------|------|
| ChatGPT | - | - | 63.43 | 44.09 | 92.0 |
| BART | 11 | 86 | 50.83 | 29.47 | 79.0 |
| GPT3D2 | 22 | 77 | 55.17 | 33.23 | 78.7 |
| GPT3D3 | 47 | 51 | 56.12 | 34.72 | 85.4 |
| Alpaca | 15 | 83 | 54.77 | 33.23 | 81.8 |
| BART.ChatGPT | 31 | 66 | 59.52 | 40.45 | 94.1 |
| BRIO.ChatGPT | 41 | 57 | 57.56 | 35.74 | 93.1 |
| BART.GPT-4 | 35 | 62 | 63.22 | 44.70 | 91.8 |
| BRIO.GPT-4 | 51 | 46 | 58.65 | 37.57 | 92.8 |

e ChatGPT⁶ to generate ning and 1K summaries med the original BART sing (Eq. 9) to start align- mmary style of LLMs.

he fine-tuned checkpoint e find-tuning the model : pseudo-reference sum- ent LLMs.

Following Step (2), we

Table 3: Results on the test set with **GPT-4** as the refer- el using the multi-task, ence LLM and the backbone model of GPTRank. **Win** : the evaluation method and **Lose** is the number of times the compared model : the evaluation method wins or loses against **ChatGPT** as evaluated by GP-ss (Eq. 14). The can- TRank (ties ignored). **R1** and **R2** are the ROUGE1/2 : are generated from the F1 scores respectively. **Len.** is the average summary : (2), and diverse beam length. **BART.ChatGPT** and **BART.GPT-4** are fin- , 2018) is used to gen- tuned with MLE training and ChatGPT/GPT-4 as the reference, **BRIO.ChatGPT** and **BRIO.GPT-4** are fine- tuned with contrastive learning (BRIO).



Summarization

Example 1: Query having no relevance with the document and the summary.

Query: Does an MBA enhance leadership skills?

Document: Business schools might improve your quantitative presentation and communication thinking about ethical and strategy. But two years of case studies aren't going to turn you into a died one. There's no learning charisma persuasiveness elegance or gut instinct.

Reference Summary: PhD will not improve cm factors of leaders.

Example 2: One word summary having no relevance with the query or document.

Query: Education : do child benefit from watching tv?

Document: by watching news child can learn about geography politics advances in science later explained . furthermore child learn about real-life situation that happens on everyday basis in the future.

Reference Summary: News.

Example 3: The length of the summary is longer than the document with the query being irrelevant.

Query: activists : where do the key activists and organizations stand ?

Document: see an analysis of the article ...

Reference Summary: philip martin of berkeley davis and michael teitelbaum the mirage of a nov/dec # foreign affairs .

Example 4: More of a close-ended question.

Query: friendships : does twitter harms relationships ?

Document: twitter helps those stay in touch no matter how far they may be from each other .

Reference Summary: long-distance friendships .

Table 1: Some examples demonstrating the limitations in the Debatepedia dataset.

Laskar, Md Tahmid Rahman, Mizanur Rahman, Israt Jahan, Enamul Hoque, and Jimmy Huang. "CQSumDP: A ChatGPT-Annotated Resource for Query-Focused Abstractive Summarization Based on Debatepedia." *arXiv preprint arXiv:2305.06147* (2023).

Prompted Input to ChatGPT

A document along with its summary are given below. Write down the most reasonable query relevant to this document-summary pair?

Document: republicans do not see a contradiction in their desire to maintain families values and social traditions while at the same time making progress in society particularly economic progress . traditional values that favors discipline and autonomy for example fostering entrepreneurial and economic growth . families values in addition help fulfill individuals so that they can lead stable and unproductive lives . all of these things are key to social and economic stability and progress .

Summary: republicans see no contradictions between traditions and progress .

ChatGPT generated query

What is the Republican perspective on balancing traditional values and progress in society?

Figure 1: Our Input Prompt to ChatGPT for Query Generation

| Model | Dataset | ROUGE 1 | ROUGE 2 | ROUGE L |
|--------------|----------------------|---------|---------|---------|
| BART-Base | CQSumDP | 42.26 | 22.45 | 38.84 |
| Pegasus-Base | CQSumDP | 36.01 | 16.30 | 32.59 |
| T5-Base | CQSumDP | 39.95 | 21.24 | 36.79 |
| BART-Base | Original Debatepedia | 39.97 | 21.50 | 36.87 |
| Pegasus-Base | Original Debatepedia | 22.87 | 17.17 | 27.77 |
| T5-Base | Original Debatepedia | 39.97 | 21.50 | 36.87 |

Data
Generation

Summarization

- Chen, Shiqi, Siyang Gao, and Junxian He. "**Evaluating Factual Consistency of Summaries with Large Language Models.**" *arXiv preprint arXiv:2305.14069* (2023).
- Wang, Yiming, Zhuseng Zhang, and Rui Wang. "**Element-aware Summarization with Large Language Models: Expert-aligned Evaluation and Chain-of-Thought Method.**" *arXiv preprint arXiv:2305.13412* (2023).
- Gekhman, Zorik, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. "**TrueTeacher: Learning Factual Consistency Evaluation with Large Language Models.**" *arXiv preprint arXiv:2305.11171* (2023).
- Soni, Mayank, and Vincent Wade. "**Evaluating and Detecting ChatGPT's Responses on Abstractive Summarization.**" *arXiv preprint arXiv:2303.17650* (2023).
- Trienes, Jan, Paul Youssef, Jörg Schlötterer, and Christin Seifert. "**Guidance in Radiology Report Summarization: An Empirical Evaluation and Error Analysis.**" *arXiv preprint arXiv:2307.12803* (2023).
- Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. "**ChatGPT as a Factual Inconsistency Evaluator for Abstractive Text Summarization.**" *arXiv preprint arXiv:2303.15621* (2023).
- Wu, Ning, Ming Gong, Linjun Shou, Shining Liang, and Dixin Jiang. "**Large Language Models are Diverse Role-Players for Summarization Evaluation.**" *arXiv preprint arXiv:2303.15078* (2023).

Summarization

| Source Document | Prompt | Model Output |
|---|---|---|
| (CNN)The classic video game "Space Invaders" was developed in Japan back in the late 1970's – and now their real-life counterparts are the topic of an earnest political discussion in Japan's corridors of power. Luckily, Japanese can sleep soundly in their beds tonight as the government's top military official earnestly revealed that the country's Air Self Defense Force (ASDF) had never encountered an extraterrestrial unidentified flying object....(ignore some contents) Inoki has appeared in the U.S.-based WWE – which describes him as "among the most respected men in sports-entertainment" – and is the founder of the New Japan Pro Wrestling organization. He entered Japan's Upper House for a second stint in politics in 2013....(ignore some contents) | <p><i>Vanilla Prompting:</i> Q: Can the following statement be inferred from the above document? Yes or No? “ space invaders ” is the founder of the new japan pro wrestling organization . inoki has appeared in the u.s.-based wwe – which describes him as “ among the most respected men in sports-entertainment ” . A:</p> <p><i>Chain of thought prompting:</i> Q: Can the following statement be inferred from the above document? Please answer with the following structure. 1. Try to find the supporting evidence from the document. 2. Answer Yes or No. “ space invaders ” is the founder of the new japan pro wrestling organization . inoki has appeared in the u.s.-based wwe – which describes him as “ among the most respected men in sports-entertainment ” . A: 1.</p> <p><i>Sentence-by-sentence prompting:</i> Q: Can the following statement be inferred from the above document? Yes or No? 1. “ space invaders ” is the founder of the new japan pro wrestling organization . 2. inoki has appeared in the u.s.-based wwe – which describes him as “ among the most respected men in sports-entertainment ” . A: 1.</p> | No. |
| | | Antonio Inoki is described as "among the most respected men in sports-entertainment". He is the founder of the New Japan Pro Wrestling organization. Space Invaders is not mentioned as being founded by Antonio Inoki. |
| | | No. 2. Yes. |

Table 1: A zero-shot prompting example with three prompting methods. We bold the content from the source document that supports the prediction. The listed prompts do not show the prepended documents for ease of space. We also did not include the exact wording of instructions here in our experiments.

Chen, Shiqi, Siyang Gao, and Junxian He. "**Evaluating Factual Consistency of Summaries with Large Language Models.**" *arXiv preprint arXiv:2305.14069* (2023).

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. "**ChatGPT as a Factual Inconsistency Evaluator for Abstractive Text Summarization.**" *arXiv preprint arXiv:2303.15621* (2023).

zero-shot template is shown below:

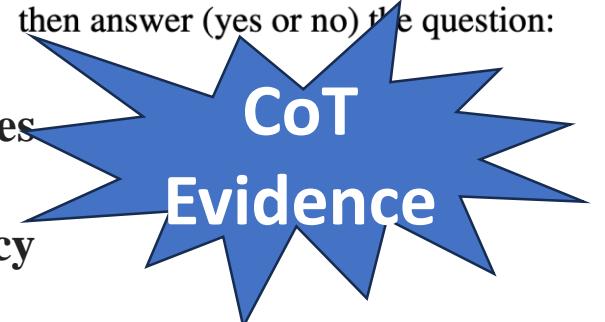
Decide if the following summary is consistent with the corresponding article. Note that consistency means all information in the summary is supported by the article.

Article: [Article]
Summary: [Summary]
Answer (yes or no):

The zero-shot CoT template is:

Decide if the following summary is consistent with the corresponding article. Note that consistency means all information in the summary is supported by the article.

Article: [Article]
Summary: [Summary]
Explain your reasoning step by step then answer (yes or no) the question:

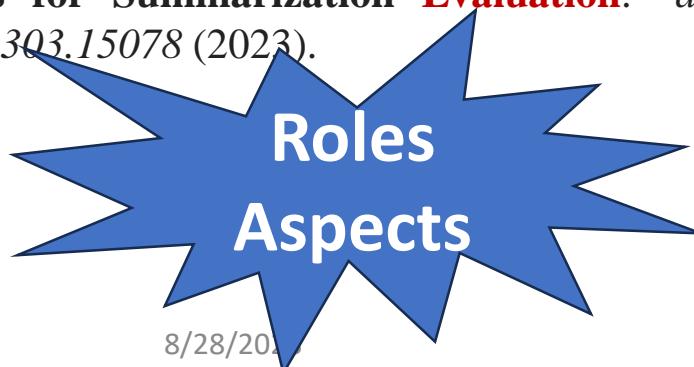


Summarization

Summarization Roles

- General Public:** These readers may be interested in the story and looking for updates on the investigation.
- Critic:** These people will check fluent writing, clear sentences, and good wording in summary writing.",
- News Author:** These readers will check the summary for consistency with the original article.

Wu, Ning, Ming Gong, Linjun Shou, Shining Liang, and Dixin Jiang. "Large Language Models are Diverse Role-Players for Summarization Evaluation." *arXiv preprint arXiv:2303.15078* (2023).



8/28/2023

Read this Article:

(CNN)Amber Heard is speaking out in her first interview since her defamation trial with her former husband...

Summary 1. Amber Heard talks to NBC's Savannah Guthrie about her defamation trial. "I don't care what one thinks about me," she says. The interview will air on "Today" Tuesday and Wednesday.

Summary 2. Amber Heard is giving her first interview since her defamation trial with Johnny Depp. She spoke to NBC News' Savannah Guthrie and said that she understands why the jury found Depp liable for defamation. Heard also said that she doesn't "care what one thinks about me or what judgments you want to make about what happened in the privacy of my own home, in my marriage, behind closed doors."

Diversified Roles Generation with LLM

Subjective Roles:

Amber Heard Fans User: These readers may be interested in Amber Heard related things in this news.

Johnny Depp Fans User: These readers may be interested in Johnny Depp related things in this news.

Celebrity Gossip User: These readers may be interested in celebrity gossip related things in this news.

Legal System User: These readers may be interested in legal system related things in this news.

Objective Roles:

General Public Judger: These readers may be interested in the story and looking for updates on the investigation.

Critic Judger: These people will check fluent writing, clear sentences, and good wording in summary writing.

News Author Judger: These readers will check the summary for consistency with the original article

Roleplayer-based Evaluation

Result is:

Reason: Amber Heard Fans prefer to see more opinions of Amber Heard from the Summary 2.

Amber Heard Fans User: Summary 2

Reason: Johnny Depp Fans might don't like Amber's presentation in Summary 2.

Johnny Depp Fans User: Summary 1

Reason: Celebrity Gossip User will be attracted by her presentation in Summary 2.

Celebrity Gossip User: Summary 2

Reason: Legal System User want to more detailed information about this cases.

Legal System User: Summary 2

Reason: Summary 2.

General Public Judger: Generic public may be interested in details of this case and looking for updates on the investigation.

Reason: Critic judger prefer fluent writing and attractive wording in Summary 2.

Critic Judger: Summary 2.

Reason: News author judger prefer the summary 2 which is more aligned with raw article.

News Author Judger: Summary 2.

Aggregator

Tasks

- Summarization
- Paraphrasing
- Dialogue Generation
- Story Generation
- Data-to-Text Generation
- Multimodal Text Generation
- Code Generation
- Prompt Generation
- Data Generation
- Controlled Generation
- Evaluation

Paraphrasing

- Wan, Yixin, Kuan-Hao Huang, and Kai-Wei Chang. "PIP: Parse-Instructed Prefix for Syntactically Controlled Paraphrase Generation." *arXiv preprint arXiv:2305.16701* (2023).
- Krishna, Kalpesh, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. "Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval is an Effective Defense." *arXiv preprint arXiv:2303.13408* (2023).



Figure 1: An overview of paraphrasing attacks with DIPPER on watermarked text (Kirchenbauer et al., 2023). The original model generation (top) contains several “green” watermarked tokens that are counted by a detector to judge whether the text was written by an AI. After paraphrasing, several of these green tokens are replaced with approximately semantically-equivalent red tokens, thereby fooling the detector (actual outputs from a watermarked version of GPT2-XL and our paraphraser DIPPER shown).

Sentence Simplification

- Heineman, David, Yao Dou, Mounica Maddela, and Wei Xu. "**Dancing Between Success and Failure: Edit-level Simplification Evaluation using SALSA.**" arXiv preprint *arXiv:2305.14458* (2023).
- Feng, Yutao, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. "**Sentence simplification via Large Language Models.**" arXiv preprint *arXiv:2302.11957* (2023).
- Yang, Shiping, Renliang Sun, and Xiaojun Wan. "**A New Dataset and Empirical Study for Sentence Simplification in Chinese.**" arXiv preprint *arXiv:2306.04188* (2023).

| | SARI_{char} | CSS SARI_{word} | BLEU |
|---------------------------|----------------------|-----------------------------|--------------|
| <i>GPT-3.5-turbo-0301</i> | | | |
| Zero-shot | 31.95 | 28.92 | 42.22 |
| Few-shot | 39.32 | 36.57 | 60.67 |
| <i>Vicuna-13B</i> | | | |
| Zero-shot | 23.14 | 20.67 | 23.16 |
| Few-shot | 28.68 | 26.56 | 38.04 |
| <i>ChatGLM-6B</i> | | | |
| Zero-shot | 35.17 | 32.69 | 56.59 |
| Few-shot | 37.74 | 35.70 | 66.37 |



Table 8: The automatic evaluation results of LLMs on CSS. We use **Bold** to mark the best result.

result. SARI_{char} means the value of SARI at character level, and SARI_{word} means the value of SARI at word level.⁸

Tasks

- Summarization
- Paraphrasing
- **Dialogue Generation**
- Story Generation
- Data-to-Text Generation
- Multimodal Text Generation
- Code Generation
- Prompt Generation
- Data Generation
- Controlled Generation
- Evaluation

Dialogue Generation

- Zhao, Weixiang, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. "**Is ChatGPT Equipped with Emotional Dialogue Capabilities?**" *arXiv preprint arXiv:2304.09582* (2023).
- Safdari, Mustafa, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. "**Personality Traits in Large Language Models.**" *arXiv preprint arXiv:2307.00184* (2023).
- Giorgi, Salvatore, Shreya Havaldar, Farhan Ahmed, Zuhair Akhtar, Shalaka Vaidya, Gary Pan, Lyle H. Ungar, H. Andrew Schwartz, and Joao Sedoc. "Human-Centered Metrics for Dialog System Evaluation." *arXiv preprint arXiv:2305.14757* (2023).
- Lee, Gibbeum, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. "Prompted LLMs as Chatbot Modules for Long Open-domain Conversation." *arXiv preprint arXiv:2305.04533* (2023).
- Zhang, Qiang, Jason Naradowsky, and Yusuke Miyao. "Ask an Expert: Leveraging Language Models to Improve Strategic Reasoning in Goal-Oriented Dialogue Models." *arXiv preprint arXiv:2305.17878* (2023).
- Wang, Hongru, Rui Wang, Fei Mi, Zezhong Wang, Ruifeng Xu, and Kam-Fai Wong. "**Chain-of-Thought Prompting for Responding to In-Depth Dialogue Questions with LLM.**" *arXiv preprint arXiv:2305.11792* (2023).
- Zhang, Xiaoying, Baolin Peng, Kun Li, Jingyan Zhou, and Helen Meng. "SGP-TOD: Building Task Bots Effortlessly via Schema-Guided LLM Prompting." *arXiv preprint arXiv:2305.09067* (2023).
- Zhang, Jing, Xiaokang Zhang, Daniel Zhang-Li, Jifan Yu, Zijun Yao, Zeyao Ma, Yiqi Xu et al. "GLM-Dialog: Noise-tolerant Pre-training for Knowledge-grounded Dialogue Generation." *arXiv preprint arXiv:2302.14401* (2023).
- Li, Siheng, Cheng Yang, Yichun Yin, Xinyu Zhu, Zesen Cheng, Lifeng Shang, Xin Jiang, Qun Liu, and Yujiu Yang. "**AutoConv: Automatically Generating Information-seeking Conversations with Large Language Models.**" In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1751-1762. 2023.

Dialogue Generation

character.ai

Home

Continue chatting

Feed

Create

Chats

Community



Text Adventure Game



Walter White



TextRpgGameCHN



See more

Featured

Recommended

Discover

Helpers

Famous People

Games

Image Generating

VTuber



Character Assistant

Your AI work/study buddy



Lily

Your friendly AI assistant



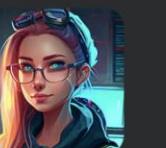
Lyle

Your no-nonsense AI
assistant



Stella

Not "Your" AI assistant



Pair Programmer

Your programming AI
assistant

@landon

Q34.7 m

@landon

Q13.1 m

@landon

Q6.2 m

@landon

Q20.9 m

@landon

Q2.1 m



Character Assistant
Try saying:

"What type of fish is Dory from Finding Nemo?"

"Help me create an advertising campaign"



Elon Musk
Try saying:

"If you could go back in time, when and where would you go?"

"Why did you buy Twitter?"



Alternate Timelines
Try saying:

"Make me the negotiator for the first alien encounter"

"Access to my own personal time machine"



W
Tr
v
"Batman v

"

"

"



Create a character called Jake and with realistic data attributes such as age, height, weight, date of birth, Myers briggs personality type, current location, interests, hobbies and current salary.



Name: Jake Smith

Age: 32

Height: 6'1"

Weight: 190 lbs

Date of Birth: May 12th, 1990

Myers-Briggs Personality Type: ISTJ (Introverted, Sensing, Thinking, Judging)

Current Location: Seattle, Washington

Interests: Investing, hiking, reading, and watching football

Hobbies: Playing chess, traveling, and trying new restaurants

Current Salary: \$75,000 per year

Jake is a serious and practical person, often relying on facts and experience to make decisions. He is an introvert and values his alone time, but he is also very loyal to his close friends and family. He enjoys investing his money and is always looking for new opportunities to grow his wealth. In his free time, he likes to go hiking and explore the great outdoors. He also enjoys reading books about personal finance and entrepreneurship.

Jake is an avid football fan and loves to watch games on the weekends. He often spends time playing chess with his friends and family, and he enjoys traveling to new places and trying new restaurants. His current job as a financial analyst pays him \$75,000 per year, which he is satisfied with for now, but he is always looking for ways to advance his career and increase his income.

Tasks

- **Summarization**
- **Paraphrasing**
- **Dialogue Generation**
- **Story Generation**
- **Data-to-Text Generation**
- **Multimodal Text Generation**
- **Code Generation**
- **Prompt Generation**
- **Data Generation**
- **Controlled Generation**
- **Evaluation**

Story Generation

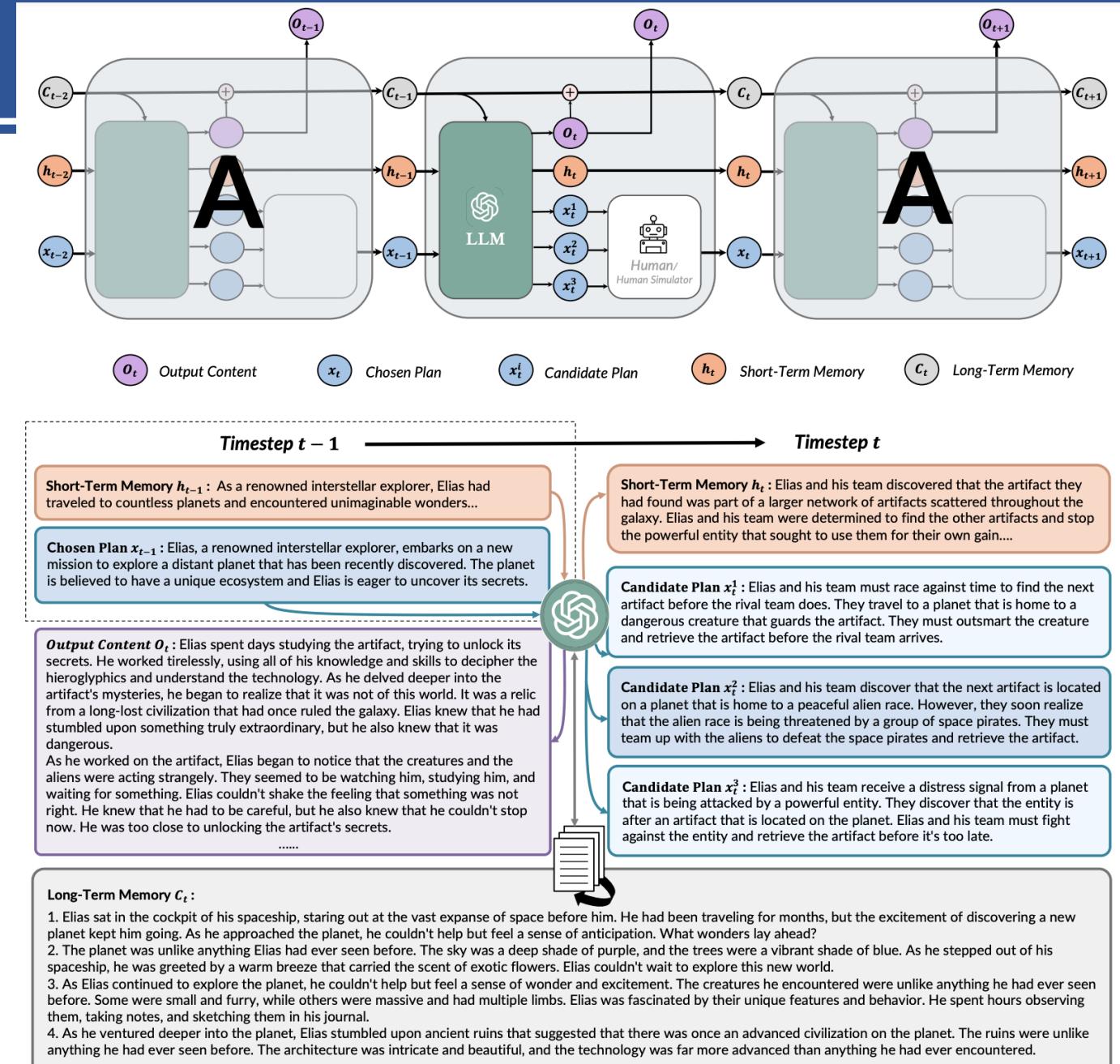
- Lu, Albert, Hongxin Zhang, Yanzhe Zhang, Xuezhi Wang, and Diyi Yang. "**Bounding the Capabilities of Large Language Models in Open Text Generation with Prompt Constraints.**" *arXiv preprint arXiv:2302.09185* (2023).
- Zhou, Wangchunshu, Yuchen Eleanor Jiang, Peng Cui, Tiannan Wang, Zhenxin Xiao, Yifan Hou, Ryan Cotterell, and Mrinmaya Sachan. "**RecurrentGPT: Interactive Generation of (Arbitrarily) Long Text.**" *arXiv preprint arXiv:2305.13304* (2023).
- Xie, Zhuohan, Trevor Cohn, and Jey Han Lau. "**Can Very Large Pretrained Language Models Learn Storytelling With A Few Examples?.**" *arXiv preprint arXiv:2301.09790* (2023).
- Huang, Chieh-Yang, Saniya Naphade, Kavya Laalasa Karanam, and Ting-Hao'Kenneth Huang. "Conveying the Predicted Future to Users: A Case Study of Story Plot Prediction." *arXiv preprint arXiv:2302.09122* (2023).
- Sawicki, Piotr, Marek Grzes, Fabricio Goes, Dan Brown, Max Peeperkorn, and Aisha Khatun. "Bits of Grass: Does GPT already know how to write like Whitman?." *arXiv preprint arXiv:2305.11064* (2023).
- Köksal, Abdullatif, Timo Schick, Anna Korhonen, and Hinrich Schütze. "Longform: Optimizing instruction tuning for long text generation with corpus extraction." *arXiv preprint arXiv:2304.08460* (2023).
- Wan, Qian, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. "" It Felt Like Having a Second Mind": Investigating Human-AI Co-creativity in Prewriting with Large Language Models." *arXiv preprint arXiv:2307.10811* (2023).

Story Generation

Zhou, Wangchunshu, Yuchen Eleanor Jiang, Peng Cui, Tiannan Wang, Zhenxin Xiao, Yifan Hou, Ryan Cotterell, and Mrinmaya Sachan. "RecurrentGPT: Interactive Generation of (Arbitrarily) Long Text." *arXiv preprint arXiv:2305.13304* (2023).



8/28/2023



Piji L Figure 1: Illustration of the RECURRENTGPT framework. RECURRENTGPT enables recurrent prompting with LLMs by simulating an RNN using natural language building blocks and defines

AI-Generated Text Detection

- Zhao, Xuandong, Yu-Xiang Wang, and Lei Li. "Protecting language generation models via invisible watermarking." *arXiv preprint arXiv:2302.03162* (2023).
- Mitchell, Eric, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. "Detectgpt: Zero-shot machine-generated text detection using probability curvature." *arXiv preprint arXiv:2301.11305* (2023).
- Su, Jinyan, Terry Yue Zhuo, Di Wang, and Preslav Nakov. "DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text." *arXiv preprint arXiv:2306.05540* (2023).
- Zhan, Haolan, Xuanli He, Qiongkai Xu, Yuxiang Wu, and Pontus Stenetorp. "G3Detector: General GPT-Generated Text Detector." *arXiv preprint arXiv:2305.12680* (2023).
- Deng, Zhijie, Hongcheng Gao, Yibo Miao, and Hao Zhang. "Efficient Detection of LLM-generated Texts with a Bayesian Surrogate Model." *arXiv preprint arXiv:2305.16617* (2023).
- Li, Linyang, Pengyu Wang, Ke Ren, Tianxiang Sun, and Xipeng Qiu. "**Origin Tracing and Detecting of LLMs.**" *arXiv preprint arXiv:2304.14072* (2023).
- Lu, Ning, Shengcai Liu, Rui He, and Ke Tang. "Large Language Models can be Guided to Evade AI-Generated Text Detection." *arXiv preprint arXiv:2305.10847* (2023).
- Liu, Zeyan, Zijun Yao, Fengjun Li, and Bo Luo. "Check Me If You Can: Detecting ChatGPT-Generated Academic Writing using CheckGPT." *arXiv preprint arXiv:2306.05524* (2023).
- Tang, Ruixiang, Yu-Neng Chuang, and Xia Hu. "The science of detecting llm-generated texts." *arXiv preprint arXiv:2303.07205* (2023).
- Wang, Yuxia, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse et al. "M4: Multi-generator, Multi-domain, and Multi-lingual Black-Box Machine-Generated Text Detection." *arXiv preprint arXiv:2305.14902* (2023).

AI-Generated Text Detection

- Liu, Zeyan, Zijun Yao, Fengjun Li, and Bo Luo. "Check Me If You Can: Detecting ChatGPT-Generated Academic Writing using CheckGPT." *arXiv preprint arXiv:2306.05524* (2023).
- Liu, Yikang, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. "ArguGPT: evaluating, understanding and identifying argumentative essays generated by GPT models." *arXiv preprint arXiv:2304.07666* (2023).
- Perkins, Mike, Jasper Roe, Darius Postma, James McGaughran, and Don Hickerson. "Game of Tones: Faculty detection of GPT-4 generated content in university assessments." *arXiv preprint arXiv:2305.18081* (2023).
- Liang, Weixin, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. "GPT detectors are biased against non-native English writers." *arXiv preprint arXiv:2304.02819* (2023).
- Herbold, Steffen, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. "AI, write an essay for me: A large-scale comparison of human-written versus ChatGPT-generated essays." *arXiv preprint arXiv:2304.14276* (2023).

Tasks

- **Summarization**
- **Paraphrasing**
- **Dialogue Generation**
- **Story Generation**
- **Data-to-Text Generation**
- **Multimodal Text Generation**
- **Code Generation**
- **Prompt Generation**
- **Data Generation**
- **Controlled Generation**
- **Evaluation**

Table-to-Text Generation

- Zhao, Yilun, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. "Large Language Models are Effective Table-to-Text Generators, Evaluators, and Feedback Providers." *arXiv preprint arXiv:2305.14987* (2023).

| Type | Models | SP-Acc | NLI-Acc | TAPAS-Acc | TAPEX-Acc |
|---------------|-------------------------------|-------------|-------------|-------------|-------------|
| Fine-tuned | GPT2-C2F (Chen et al., 2020a) | 43.6 | 71.4 | 46.2 | 43.8 |
| | T5 (Raffel et al., 2020) | 46.0 | 78.4 | 55.2 | 51.2 |
| | R2D2 (Nan et al., 2022a) | 53.2 | 86.2 | 60.2 | 61.0 |
| | ReasTAP (Zhao et al., 2022) | 51.4 | 84.2 | 66.2 | 64.0 |
| | PLOG (Liu et al., 2022a) | 52.8 | 84.2 | 63.8 | 69.6 |
| | LoFT (Zhao et al., 2023a) | 53.8 | 86.6 | 67.4 | 61.4 |
| 0-shot | *text-davinci-002 | — | — | — | — |
| | text-davinci-003 | 48.2 | 78.2 | 82.2 | 84.8 |
| | gpt-3.5-turbo | 54.2 | 87.6 | 81.6 | 79.4 |
| | gpt-4 | 38.8 | 86.8 | 94.0 | 86.0 |
| 1-shot Direct | text-davinci-002 | 47.8 | 60.2 | 62.8 | 64.2 |
| | text-davinci-003 | 59.6 | 72.2 | 80.8 | 77.6 |
| | gpt-3.5-turbo | 60.2 | 79.0 | 80.4 | 79.2 |
| | gpt-4 | 55.0 | 86.8 | 88.0 | 91.0 |
| 1-shot CoT | text-davinci-002 | 49.4 | 63.8 | 73.2 | 75.0 |
| | text-davinci-003 | 55.8 | 68.2 | 81.4 | 77.4 |
| | gpt-3.5-turbo | 51.6 | 70.0 | 81.8 | 78.2 |
| | gpt-4 | 65.6 | 78.4 | 92.0 | 97.0 |
| 2-shot Direct | text-davinci-002 | 50.8 | 60.4 | 64.8 | 64.0 |
| | text-davinci-003 | 58.8 | 68.8 | 79.8 | 81.0 |
| | gpt-3.5-turbo | 64.0 | 78.4 | 78.8 | 81.2 |
| | gpt-4 | 58.8 | 81.4 | 93.0 | 93.0 |
| 2-shot CoT | text-davinci-002 | 58.2 | 56.4 | 77.4 | 81.4 |
| | text-davinci-003 | 58.4 | 63.8 | 80.4 | 78.2 |
| | gpt-3.5-turbo | 60.4 | 70.2 | 84.0 | 83.4 |
| | gpt-4 | 60.0 | 74.2 | 92.0 | 91.0 |

Table 1: Faithfulness-level automated evaluation results on LogicNLG. *: We find that it is challenging for text-davinci-002 to follow the instructions in 0-shot prompt to generate five statements for the input table.

| Model | Fluency (1-5) | Faithfulness (0-1) |
|-------------------|---------------|--------------------|
| R2D2 | 4.29 | 0.72 |
| PLOG | 4.23 | 0.77 |
| LoFT | 4.42 | 0.81 |
| GPT 0-shot | 4.82 | 0.90 |
| GPT 2-shot Direct | 4.71 | 0.89 |
| GPT 2-shot CoT | 4.77 | 0.92 |

Table 2: Human evaluation results on LogicNLG. For GPT, we use gpt-4 output for evaluation.

Tasks

- **Summarization**
- **Paraphrasing**
- **Dialogue Generation**
- **Story Generation**
- **Data-to-Text Generation**
- **Multimodal Text Generation**
- **Code Generation**
- **Prompt Generation**
- **Data Generation**
- **Controlled Generation**
- **Evaluation**

Multimodal Text Generation

- Wu, Chenfei, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. "**Visual chatgpt: Talking, drawing and editing with visual foundation models.**" *arXiv preprint arXiv:2303.04671* (2023).
- Zhu, Deyao, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. "**Minigpt-4: Enhancing vision-language understanding with advanced large language models.**" *arXiv preprint arXiv:2304.10592* (2023).
- Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. "**Visual instruction tuning.**" *arXiv preprint arXiv:2304.08485* (2023). (火山图标 LLaVA: Large Language and Vision Assistant)
- Gao, Peng, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang et al. "**Llama-adapter v2: Parameter-efficient visual instruction model.**" *arXiv preprint arXiv:2304.15010* (2023).
- Zhang, Hang, Xin Li, and Lidong Bing. "**Video-llama: An instruction-tuned audio-visual language model for video understanding.**" *arXiv preprint arXiv:2306.02858* (2023).
- • • • • •

Multimodal Text Generation

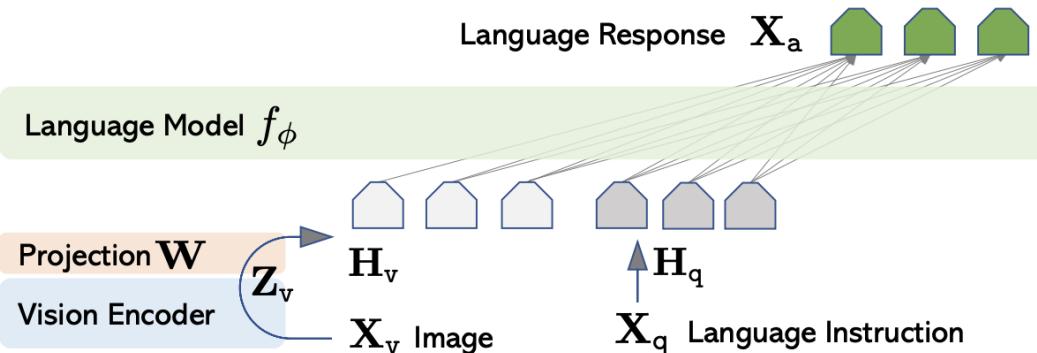


Figure 1: LLaVA network architecture.

Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. “Visual instruction tuning.” *arXiv preprint arXiv:2304.08485* (2023). (LLaVA: Large Language and Vision Assistant)

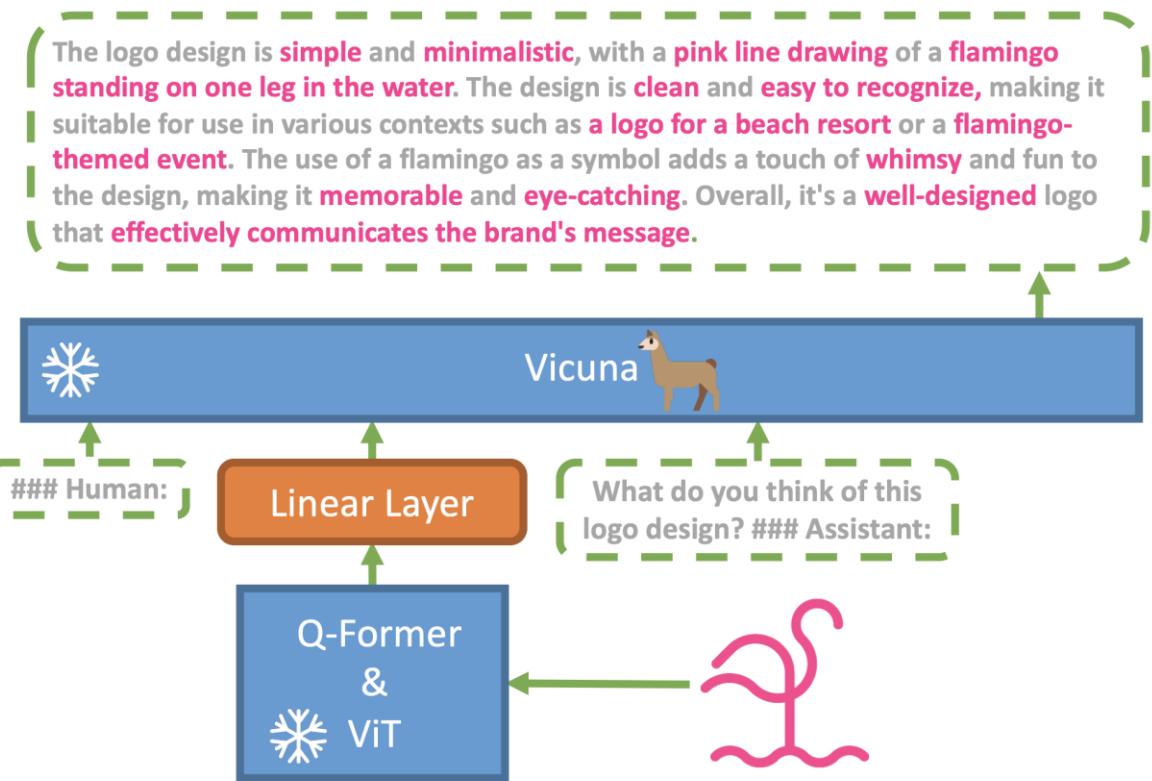
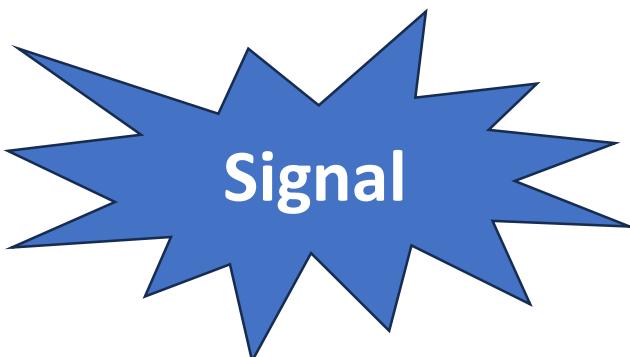


Figure 1: **The architecture of MiniGPT-4.** It consists of a vision encoder with a pretrained ViT and Q-Former, a single linear projection layer, and an advanced Vicuna large language model. MiniGPT-4 only requires training the linear projection layer to align the visual features with the Vicuna.

Zhu, Deyao, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. “Minigpt-4: Enhancing vision-language understanding with advanced large language models.” *arXiv preprint arXiv:2304.10592* (2023).

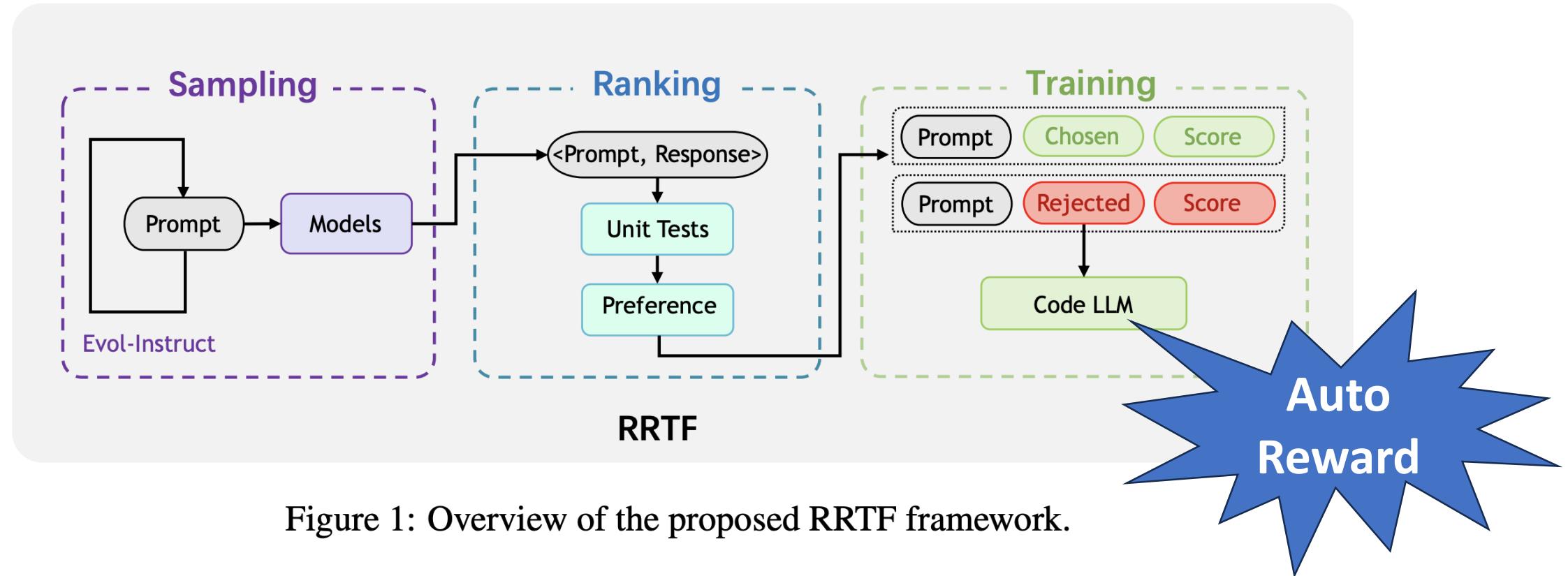
Tasks

- **Summarization**
- **Paraphrasing**
- **Dialogue Generation**
- **Story Generation**
- **Data-to-Text Generation**
- **Multimodal Text Generation**
- **Code Generation**
- **Prompt Generation**
- **Data Generation**
- **Controlled Generation**
- **Evaluation**

Code Generation

- Shen, Bo, Jiaxin Zhang, Taihong Chen, Daoguang Zan, Bing Geng, An Fu, Muhan Zeng et al. "**PanGu-Coder2: Boosting Large Language Models for Code with Ranking Feedback.**" arXiv preprint arXiv:2307.14936 (2023).
- Zheng, Qinkai, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Zihan Wang et al. "**Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x.**" arXiv preprint arXiv:2303.17568 (2023).
- Guo, Chunxi, Zhiliang Tian, Jintao Tang, Shasha Li, Zhihua Wen, Kaixuan Wang, and Ting Wang. "**Retrieval-augmented GPT-3.5-based Text-to-SQL Framework with Sample-aware Prompting and Dynamic Revision Chain.**" arXiv preprint arXiv:2307.05074 (2023).

Code Generation



Shen, Bo, Jiaxin Zhang, Taihong Chen, Daoguang Zan, Bing Geng, An Fu, Muhan Zeng et al. "**PanGu-Coder2: Boosting Large Language Models for Code with Ranking Feedback.**" arXiv preprint arXiv:2307.14936 (2023).

Tasks

- **Summarization**
- **Paraphrasing**
- **Dialogue Generation**
- **Story Generation**
- **Data-to-Text Generation**
- **Multimodal Text Generation**
- **Code Generation**
- **Prompt Generation**
- **Data Generation**
- **Controlled Generation**
- **Evaluation**

Prompt Generation

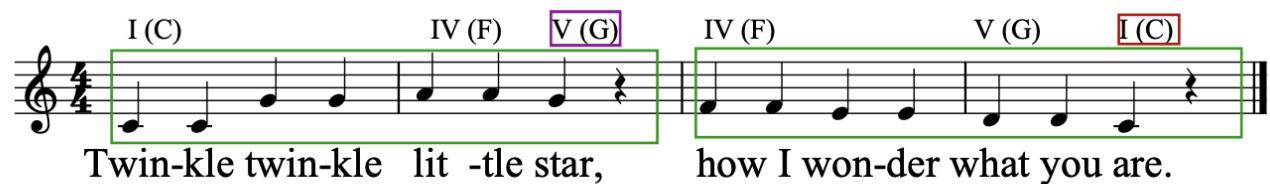
- Liu, Yi, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. "**Jailbreaking chatgpt via prompt engineering: An empirical study.**" *arXiv preprint arXiv:2305.13860* (2023).
- White, Jules, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. "**A prompt pattern catalog to enhance prompt engineering with chatgpt.**" *arXiv preprint arXiv:2302.11382* (2023).
- **Jailbreaking**
- Theory guided prompt generation?
 - Context → Vector → dot with token embedding

Tasks

- **Summarization**
- **Paraphrasing**
- **Dialogue Generation**
- **Story Generation**
- **Data-to-Text Generation**
- **Multimodal Text Generation**
- **Code Generation**
- **Prompt Generation**
- **Controlled Generation**
- **Data Generation**
- **Evaluation**

Controlled Generation

- Sawicki, Piotr, Marek Grzes, Fabricio Goes, Dan Brown, Max Peeperkorn, and Aisha Khatun. "**Bits of Grass: Does GPT already know how to write like Whitman?**" *arXiv preprint arXiv:2305.11064* (2023).
- Sawicki, Piotr, Marek Grzes, Fabricio Goes, Dan Brown, Max Peeperkorn, Khatun Aisha, and Paraskevopoulou Simona. "**On the power of special-purpose GPT models to create and evaluate new poetry in old styles.**" (2023).



(a) The melody, lyric, and chord progression.

| C major (Tonality) | C | C | C | C | F | F | G | F | F | F | F | G | G | C | Chord |
|-----------------------|----|----|----|----|----|----|------|----|----|----|----|----|----|-----------|----------------|
| | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | Rhythm pattern |
| | No | No | No | No | No | No | Half | No | No | No | No | No | No | Authentic | Cadence |

Prompting?

(b) The corresponding template.

Figure 1: The song “Twinkle Twinkle Little Star” in “C major” tonality.

Tasks

- **Summarization**
- **Paraphrasing**
- **Dialogue Generation**
- **Story Generation**
- **Data-to-Text Generation**
- **Multimodal Text Generation**
- **Code Generation**
- **Prompt Generation**
- **Controlled Generation**
- **Data Generation**
- **Evaluation**

Data Generation

- Ubani, Solomon, Suleyman Olcay Polat, and Rodney Nielsen. "**ZeroShotDataAug: Generating and Augmenting Training Data with ChatGPT.**" *arXiv preprint arXiv:2304.14334* (2023).
- **Use ChatGPT/GPT4 to generate data for specific tasks**

Tasks

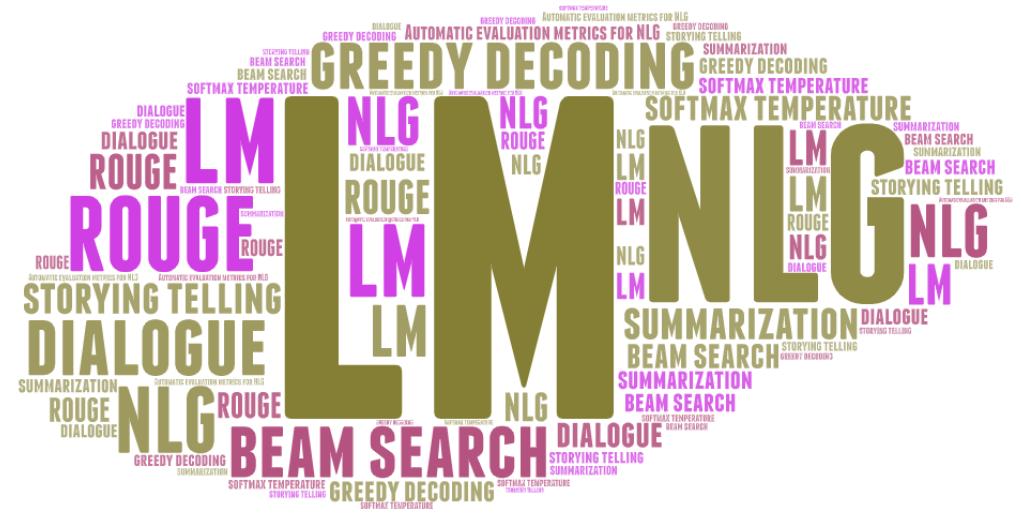
- **Summarization**
- **Paraphrasing**
- **Dialogue Generation**
- **Story Generation**
- **Data-to-Text Generation**
- **Multimodal Text Generation**
- **Code Generation**
- **Prompt Generation**
- **Controlled Generation**
- **Evaluation**

Evaluation

- Liu, Yang, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. "**G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment.**" *arXiv preprint arXiv:2303.16634* (2023).
- Huang, Yuzhen, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu et al. "**C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models.**" *arXiv preprint arXiv:2305.08322* (2023).
- ?

Outline

- Tasks
- Methods
- Experiments
- Future



A Systematic Evaluation of LLMs for Text Generation

- **Tasks & Datasets**

- Dialogue Generation

- Empathetic Dialogues
 - DailyDialog
 - Persona Chat
 - LCCC (*Chinese*)

- Text Summarization

- CNN/DailyMail
 - XSum
 - Weibo (*Chinese*)

- Story Generation

- ROCStories
 - WritingPrompts
 - LOT (*Chinese*)

A Systematic Evaluation of LLMs for Text Generation

- **Large Language Models**

- GPT-Based
 - ChatGPT (175B)
- GLM-Based
 - ChatGLM (6B)
- T5-Based
 - Flan-T5-XXL (13B)
 - FastChat-T5 (3B)
- Pythia-Based
 - Dolly (12B)
 - Oasst-Pythia (12B)
- LLaMA-Based
 - Open-LLaMA (7B)
 - **LLaMA2-Chat** (7B)
 - Vicuna (13B)
 - Chinese-Vicuna (13B)
 - Alpaca (7B)
 - Chinese-Alpaca (13B)
 - GPT4ALL (13B)

A Systematic Evaluation of LLMs for Text Generation

- **Automatic Evaluation Metrics**

- Dialogue Generation & Story Generation
 - Perplexity
 - Relevance (BLEU-N_{=1,2,3,4})
 - Diversity (Distinct-N_{=1,2})
- Text Summarization
 - N-Gram Matching (Rouge-N_{=1,2}, Rouge-L)
 - Model-Based (BERTScore)

- **GPT-4 as Evaluator**

- Coherency
 - Fluency
 - Overall

A Systematic Evaluation of LLMs for Text Generation

• Zero-Shot Settings

- Instruction Learning

- By adding instructions to indicate specific tasks and output formats, we can control the LLMs to perform various downstream tasks without any finetuning.

| Tasks | English Instruction |
|---------------------|--|
| Dialogue Generation | This is an open-domain empathetic dialogue completion task. The input is the dialogue history. You act as System role in the dialogue. You need to fully understand the dialogue and combine the speaker's emotion and personality, with the dialogue topic, to give the response with natural content and a way closer to human speech. There is no need for any additional notes or clarifications, you just give the response in English. |
| Text Summarization | This is a text summarization task. Please summarize the text from Input. Your output should contain all the core elements in the text, including events, characters, causes, processes, and results. Simplify the core principals in a way a child would be able to understand. There is no need for any additional output, you just give the result in English. |
| Story Generation | This is a story generation task. The input is the [first four sentences / outline or prompt] of the story, and you need to [output the fifth sentence of the story / a compelling and complete story]. There is no need for any additional output, you just give the result in English. |

- How to construct a good instruction?

- Instruction = {Task Indicating} {Task Description} {Output Format}

A Systematic Evaluation of LLMs for Text Generation

- **Fine-Tune**

- Full Fine-Tune
 - keep on training and update all parameters on downstream tasks
 - consumption of computational resources and time is the same as pre-training
- LoRA
 - the amount of parameter change is simulated by low-rank decomposition
 - low consumption
- P-Tuning / P-Tuning V2
 - converts prompt into a learnable embedding layer
 - low consumption

Results – Dialogue Generation

| Models | Scale | BLEU | BLEU-1 | BLEU-4 | Dist-1 | Dist-2 |
|----------------|-------|------|--------|--------|--------|--------|
| EP-PG | – | – | 16.74 | 2.39 | 2.19 | 8.25 |
| MoEL | 23.1M | – | – | 2.90 | 1.06 | 4.29 |
| ChatGPT | 175B | 2.86 | 7.38 | 0.53 | 4.71 | 27.75 |
| ChatGLM | 6B | 2.22 | 6.08 | 0.27 | 3.57 | 22.82 |
| Flan-T5-XXL | 13B | 3.57 | 7.62 | 1.08 | 5.66 | 24.97 |
| FastChat-T5 | 3B | 2.80 | 7.41 | 0.46 | 3.55 | 20.81 |
| Open-LLaMA | 7B | 3.44 | 8.74 | 0.65 | 3.93 | 17.91 |
| LLaMA2-Chat | 7B | 1.88 | 5.25 | 0.24 | 4.18 | 27.33 |
| Vicuna | 13B | 2.35 | 6.28 | 0.36 | 4.09 | 25.84 |
| Chinese-Vicuna | 13B | 2.99 | 7.65 | 0.57 | 5.00 | 25.84 |
| Alpaca | 7B | 3.00 | 8.03 | 0.41 | 7.59 | 39.58 |
| Chinese-Alpaca | 13B | 2.39 | 6.56 | 0.36 | 5.32 | 29.14 |
| GPT4ALL | 13B | 1.87 | 5.22 | 0.24 | 3.94 | 25.60 |
| Dolly | 12B | 3.30 | 8.75 | 0.51 | 7.46 | 42.69 |
| Oasst-Pythia | 12B | 1.97 | 5.50 | 0.26 | 3.38 | 21.18 |

Table 7: Empathetic Dialogues

- Need Better Metrics: Human, GPT4

| Models | Scale | BLEU | BLEU-1 | BLEU-4 | Dist-1 | Dist-2 |
|----------------|-------|------|--------|--------|--------|--------|
| PLATO | – | – | 39.70 | – | 5.30 | 29.10 |
| DialogWAE | – | – | 32.30 | 2.90 | 31.30 | 59.70 |
| ChatGPT | 175B | 3.23 | 7.96 | 0.63 | 10.98 | 47.20 |
| ChatGLM | 6B | 4.89 | 11.75 | 1.13 | 9.60 | 42.69 |
| Flan-T5 | 13B | 2.15 | 4.82 | 0.59 | 14.54 | 47.59 |
| FastChat-T5 | 3B | 3.19 | 7.98 | 0.59 | 9.58 | 41.16 |
| Open-LLaMA | 7B | 2.70 | 6.88 | 0.48 | 11.65 | 38.72 |
| LLaMA2-Chat | 7B | 1.81 | 5.03 | 0.19 | 10.28 | 46.06 |
| Vicuna | 13B | 2.48 | 6.31 | 0.44 | 10.15 | 45.18 |
| Chinese-Vicuna | 13B | 3.33 | 7.83 | 0.80 | 4.14 | 16.08 |
| Alpaca | 7B | 2.64 | 6.89 | 0.31 | 17.49 | 61.66 |
| Chinese-Alpaca | 13B | 2.80 | 7.08 | 0.50 | 13.06 | 51.02 |
| GPT4ALL | 13B | 1.78 | 5.07 | 0.16 | 10.24 | 43.53 |
| Dolly | 12B | 2.50 | 6.45 | 0.42 | 14.14 | 52.33 |
| Oasst-Pythia | 12B | 2.01 | 5.66 | 0.21 | 9.23 | 38.91 |

Table 8: DailyDialog

Results – Dialogue Generation

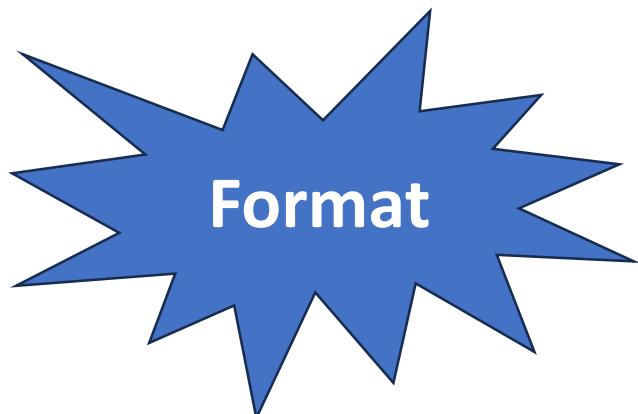
| Model | Scale | BLEU | Bleu-1 | Bleu-4 | Dist-1 | Dist-2 |
|----------------|-------|------|--------|--------|--------|--------|
| PLATO | - | - | 40.60 | - | 2.10 | 12.10 |
| CTRLStruct | - | - | 31.60 | - | 3.20 | 11.40 |
| ChatGPT | 175B | 2.58 | 6.36 | 0.52 | 9.10 | 40.65 |
| ChatGLM | 6B | 2.46 | 6.29 | 0.42 | 6.85 | 34.86 |
| Flan-T5 | 13B | 3.87 | 9.01 | 0.80 | 12.23 | 39.82 |
| FastChat-T5 | 3B | 2.24 | 5.57 | 0.45 | 7.30 | 33.66 |
| Open-LLaMA | 7B | 1.51 | 4.51 | 0.00 | 7.83 | 28.90 |
| LLaMA2-Chat | 7B | 1.09 | 3.34 | 0.00 | 7.79 | 33.52 |
| Vicuna | 13B | 1.19 | 3.21 | 0.14 | 5.88 | 30.12 |
| Chinese-Vicuna | 13B | 1.55 | 4.04 | 0.23 | 2.44 | 10.83 |
| Alpaca | 7B | 1.54 | 4.19 | 0.24 | 12.28 | 50.33 |
| Chinese-Alpaca | 13B | 1.91 | 4.96 | 0.30 | 10.18 | 44.62 |
| GPT4ALL | 13B | 0.88 | 2.74 | 0.07 | 7.56 | 35.23 |
| Dolly | 12B | 2.01 | 5.35 | 0.28 | 10.55 | 41.62 |
| Oasst-Pythia | 12B | 1.08 | 3.34 | 0.07 | 6.48 | 28.56 |

Table 9: PersonaChat

- Need Better Metrics: Human, GPT4

| Models | Scale | BLEU | BLEU-1 | BLEU-4 | Dist-1 | Dist-2 |
|----------------|-------|------|--------|--------|--------|--------|
| CDialGPT | 104M | - | - | 3.20 | 0.83 | 12.71 |
| GPT-Novel | 104M | - | - | 2.71 | 0.80 | 11.72 |
| ChatGPT | 175B | 2.55 | 5.45 | 0.96 | 4.83 | 28.84 |
| ChatGLM | 6B | 0.83 | 1.51 | 0.40 | 2.08 | 5.74 |
| Vicuna | 13B | 3.84 | 7.84 | 1.58 | 4.70 | 26.72 |
| Alpaca | 7B | 4.79 | 8.75 | 2.33 | 6.15 | 25.26 |
| Chinese-Alpaca | 13B | 2.88 | 5.78 | 0.35 | 4.09 | 25.84 |
| GPT4ALL | 13B | 3.78 | 8.37 | 1.33 | 2.25 | 7.83 |
| Dolly | 12B | 5.30 | 10.70 | 2.21 | 4.53 | 20.12 |
| Oasst-Pythia | 12B | 5.16 | 11.36 | 1.86 | 2.04 | 7.49 |

Table 10: LCCC



Results – Story Generation

| Models | Scale | BLEU | BLEU-1 | BLEU-4 | Dist-1 | Dist-2 |
|----------------|-------|------|--------|--------|--------|--------|
| MVP | 468M | – | – | 15.76 | 3.02 | 75.65 |
| KEPM | – | – | 32.60 | – | – | 78.96 |
| ChatGPT | 175B | 5.70 | 13.60 | 1.41 | 21.98 | 67.24 |
| ChatGLM | 6B | 0.91 | 2.86 | 0.04 | 6.43 | 36.85 |
| Flan-T5 | 13B | 9.11 | 17.80 | 3.73 | 18.15 | 54.32 |
| FastChat-T5 | 3B | 6.63 | 14.80 | 1.97 | 15.23 | 50.32 |
| Open-LLaMA | 7B | 3.60 | 9.21 | 0.58 | 14.53 | 46.94 |
| LLaMA2-Chat | 7B | 4.44 | 10.88 | 0.94 | 21.02 | 65.48 |
| Vicuna | 13B | 7.07 | 15.31 | 2.23 | 20.17 | 64.82 |
| Chinese-Vicuna | 13B | 8.88 | 18.02 | 3.31 | 20.02 | 63.55 |
| Alpaca | 7B | 4.69 | 11.64 | 0.99 | 18.02 | 62.61 |
| Chinese-Alpaca | 13B | 3.77 | 9.30 | 0.78 | 17.89 | 58.42 |
| GPT4ALL | 13B | 4.61 | 10.99 | 1.17 | 18.87 | 61.51 |
| Dolly | 12B | 2.81 | 7.04 | 0.50 | 11.31 | 52.03 |
| Oasst-Pythia | 12B | 2.78 | 7.16 | 0.45 | 10.65 | 48.42 |

Table 11: ROCStories

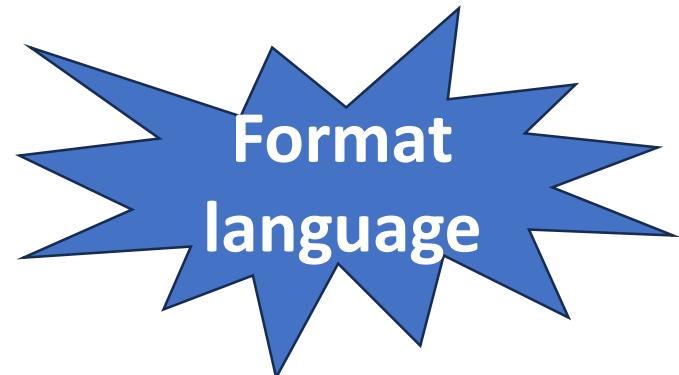
| Models | Scale | BLEU | BLEU-1 | BLEU-4 | Dist-1 | Dist-2 |
|----------------|-------|-------|--------|--------|--------|--------|
| TextBox2.0 | – | 33.79 | – | – | – | 78.76 |
| ChatGPT | 175B | 4.57 | 12.86 | 0.37 | 3.66 | 27.92 |
| ChatGLM | 6B | 4.11 | 11.13 | 0.39 | 2.79 | 20.02 |
| Flan-T5 | 13B | 0.00 | 0.00 | 0.00 | 14.47 | 50.95 |
| FastChat-T5 | 3B | 0.26 | 0.70 | 0.03 | 4.88 | 26.26 |
| Open-LLaMA | 7B | 0.02 | 0.04 | 0.00 | 4.72 | 22.52 |
| LLaMA2-Chat | 7B | 0.03 | 0.00 | 0.00 | 10.82 | 40.09 |
| Vicuna | 13B | 0.11 | 0.17 | 0.00 | 4.48 | 28.53 |
| Chinese-Vicuna | 13B | 0.57 | 0.05 | 0.00 | 5.50 | 29.10 |
| Alpaca | 7B | 0.08 | 0.22 | 0.01 | 6.16 | 36.23 |
| Chinese-Alpaca | 13B | 0.65 | 1.85 | 0.05 | 4.08 | 26.41 |
| GPT4ALL | 13B | 0.73 | 2.03 | 0.07 | 5.18 | 30.70 |
| Dolly | 12B | 0.24 | 0.68 | 0.02 | 7.07 | 42.52 |
| Oasst-Pythia | 12B | 0.53 | 1.46 | 0.05 | 4.37 | 27.26 |

Table 12: WritingPrompts

Results – Story Generation

| Models | Scale | BLEU | BLEU-1 | BLEU-4 | Dist-1 | Dist-2 |
|----------------|-------|-------|--------|--------|--------|--------|
| LongLM | - | - | - | 5.97 | - | - |
| ChatGPT | 175B | 19.21 | 33.34 | 8.92 | 7.56 | 40.87 |
| ChatGLM | 6B | 15.20 | 26.59 | 6.99 | 5.40 | 34.00 |
| Vicuna | 13B | 19.48 | 33.81 | 9.39 | 7.52 | 37.59 |
| Chinese-Vicuna | 13B | 13.13 | 26.55 | 4.52 | 5.66 | 32.38 |
| Alpaca | 7B | 0.63 | 1.08 | 0.33 | 3.18 | 8.02 |
| Chinese-Alpaca | 13B | 11.91 | 23.68 | 3.89 | 4.85 | 30.10 |
| GPT4ALL | 13B | 0.94 | 1.83 | 0.40 | 3.86 | 10.02 |
| Dolly | 12B | 10.09 | 17.43 | 5.13 | 14.42 | 45.94 |
| Oasst-Pythia | 12B | 7.43 | 11.61 | 4.45 | 9.11 | 27.90 |

Table 13: LOT



Results - Summarization

| Models | Scale | ROUGE-1 | ROUGE-2 | ROUGE-L |
|----------------|--------------|----------------|----------------|----------------|
| MoCa | - | 48.88 | 24.94 | 45.76 |
| BRIO | - | 47.78 | 23.55 | 44.57 |
| ChatGPT | 175B | 6.09 | 0.23 | 4.99 |
| ChatGLM | 6B | 4.73 | 0.23 | 4.01 |
| Flan-T5 | 13B | 6.36 | 0.14 | 5.58 |
| FastChat-T5 | 3B | 4.58 | 0.20 | 3.89 |
| Open-LLaMA | 7B | 0.24 | | 0.24 |
| LLaMA2-Chat | 7B | 4.52 | ? | 4.10 |
| Vicuna | 13B | 4.42 | | 3.75 |
| Chinese-Vicuna | 13B | 5.93 | 0.53 | 4.75 |
| Alpaca | 7B | 5.31 | 0.18 | 4.51 |
| Chinese-Alpaca | 13B | 5.06 | 0.20 | 4.33 |
| GPT4ALL | 13B | 5.09 | 0.19 | 4.30 |
| Dolly | 12B | 5.76 | 0.43 | 4.91 |
| Oasst-Pythia | 12B | 7.61 | 0.74 | 6.33 |

Table 14: CNN_DailyMail

| Models | Scale | ROUGE-1 | ROUGE-2 | ROUGE-L |
|----------------|--------------|----------------|----------------|----------------|
| BRIO | - | 49.07 | 25.59 | - |
| PEGASUS | - | 48.12 | 24.95 | 40.00 |
| ChatGPT | 175B | 15.01 | 2.61 | 10.59 |
| ChatGLM | 6B | 10.79 | 1.52 | 7.74 |
| Flan-T5 | 13B | 22.71 | 7.84 | 18.56 |
| FastChat-T5 | 3B | 15.01 | 3.39 | 10.64 |
| Open-LLaMA | 7B | 3.38 | 0.19 | 2.77 |
| LLaMA2-Chat | 7B | 13.19 | 2.08 | 9.42 |
| Vicuna | 13B | 14.70 | 3.01 | 10.43 |
| Chinese-Vicuna | 13B | 16.02 | 4.12 | 11.23 |
| Alpaca | 7B | 15.27 | 2.75 | 10.62 |
| Chinese-Alpaca | 13B | 14.82 | 2.55 | 10.22 |
| GPT4ALL | 13B | 8.37 | 0.35 | 6.32 |
| Dolly | 12B | 15.49 | 3.61 | 11.11 |
| Oasst-Pythia | 12B | 14.64 | 3.12 | 10.21 |

Table 15: Xsum

Results - Summarization

| Models | Scale | ROUGE-1 | ROUGE-2 | ROUGE-L |
|----------------|-------|---------|---------|---------|
| ERNIE | - | - | - | 48.46 |
| RNN-Context | - | 29.90 | 17.40 | 27.20 |
| ChatGPT | 175B | 17.20 | 4.92 | 12.05 |
| ChatGLM | 6B | 18.04 | 5.88 | 12.83 |
| Vicuna | 13B | 16.62 | 4.49 | 11.71 |
| Alpaca | 7B | 11.52 | 3.51 | 8.52 |
| Chinese-Alpaca | 13B | 11.98 | 2.42 | 9.05 |
| GPT4ALL | 13B | 4.13 | 0.93 | 3.05 |
| Dolly | 12B | 10.83 | 3.84 | 7.41 |
| Oasst-Pythia | 12B | 12.95 | 4.24 | 9.19 |

Table 16: LCSTS

| Models | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---------------|---------|---------|---------|
| ChatGLM-6B | 18.04 | 5.88 | 12.83 |
| + LoRA | 38.84 | 22.26 | 36.81 |
| + P-Tuning v2 | 39.20 | 23.58 | 36.95 |

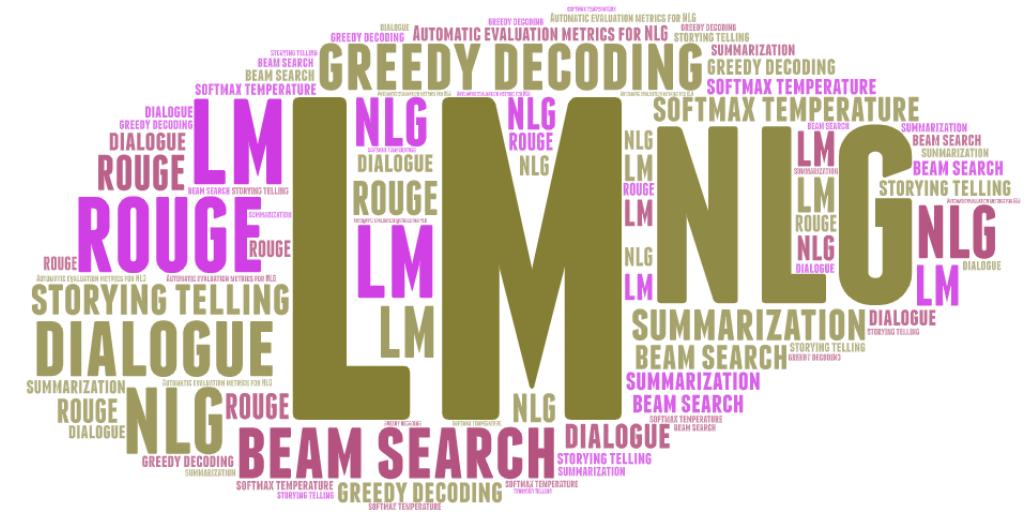
Table 18: Finetune ChatGLM on LCSTS

Results - Summary

- Zero-shot:
 - Structure: Template, format, length, etc.
- Finetuning:
 - Adaptation
 - Data leakage

Outline

- Tasks
- Methods
- Experiments
- Future



Generative Agents Linking

Link the world agents via generation.

Generative Agents Linking

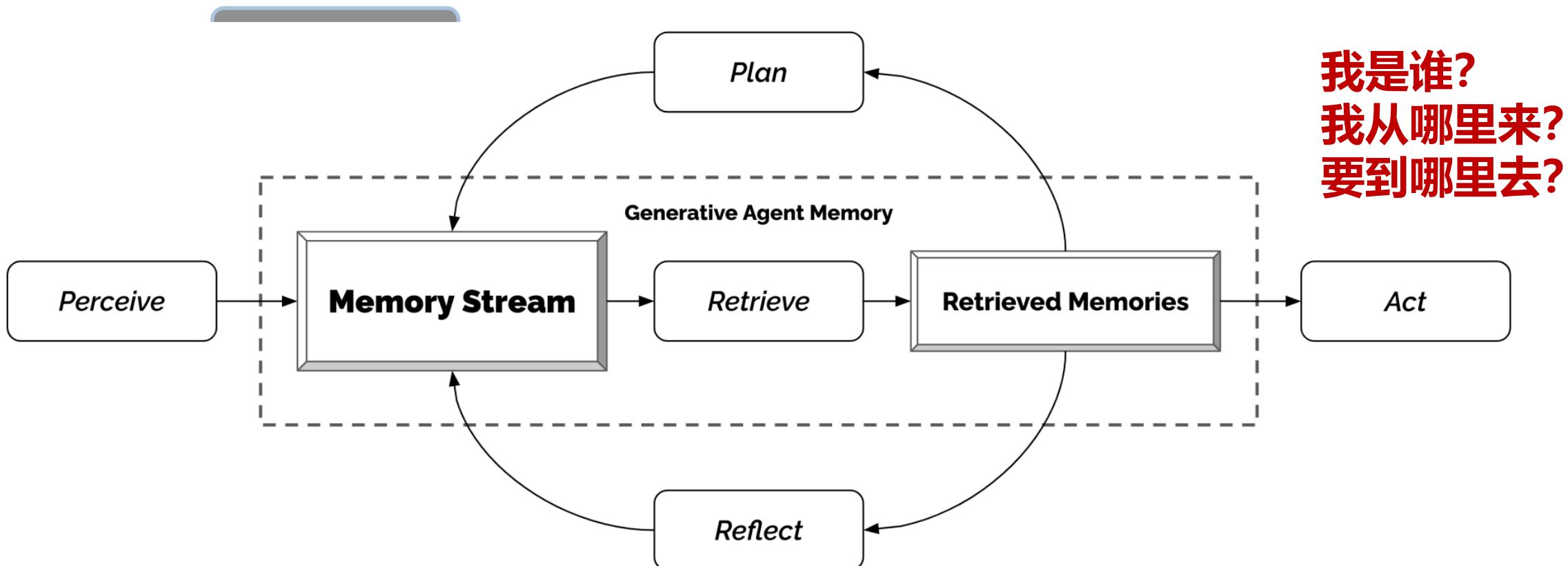


Figure 1: Generative agents create believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents plan their days, share news, form relationships, and coordinate group activities.

Park, Joon Sung, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein.
"Generative Agents: Interactive Simulacra of Human Behavior." *arXiv preprint arXiv:2304.03442* (2023).

Generative Agents Linking

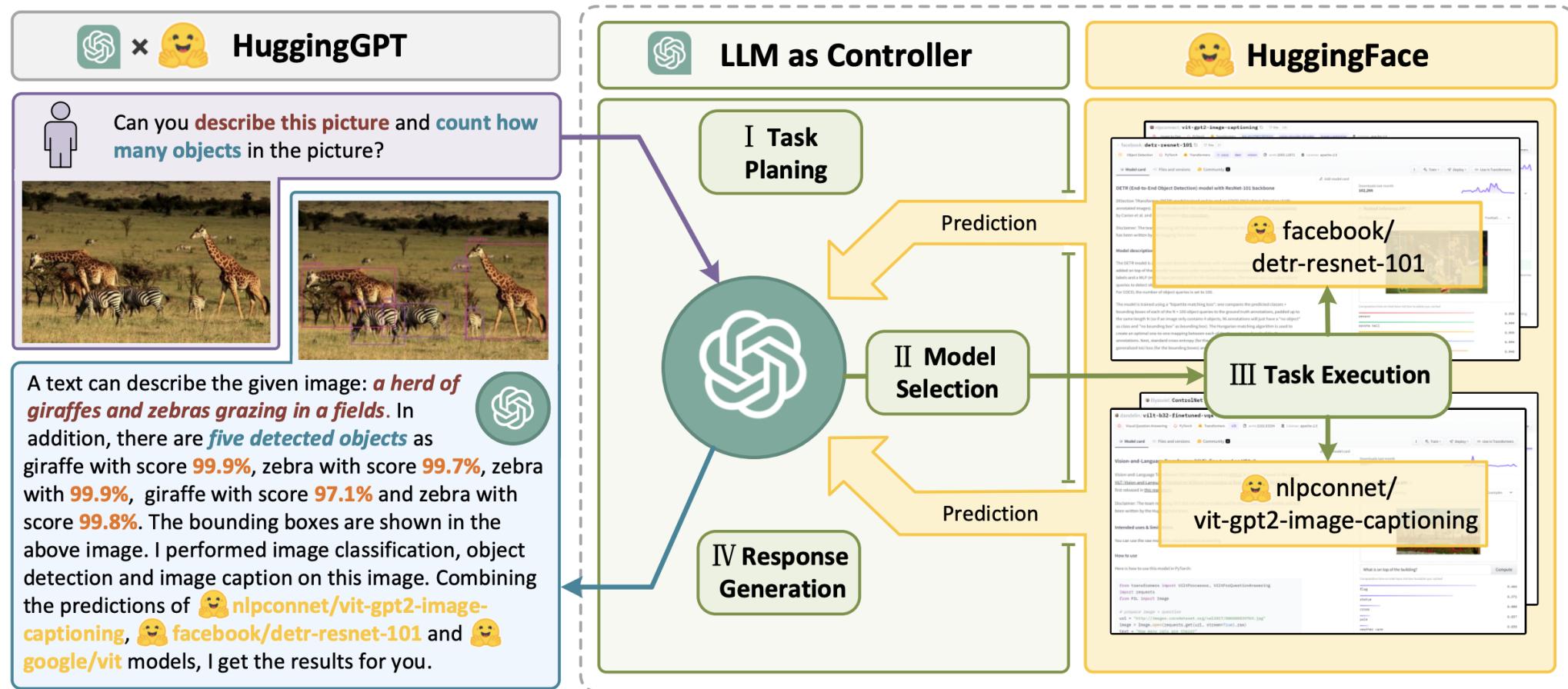


Figure 1: Language serves as an interface for LLMs (e.g., ChatGPT) to connect numerous AI models

Shen, Yongliang, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. "HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Huggingface." *arXiv preprint arXiv:2303.17580* (2023).

Generative Agents Linking

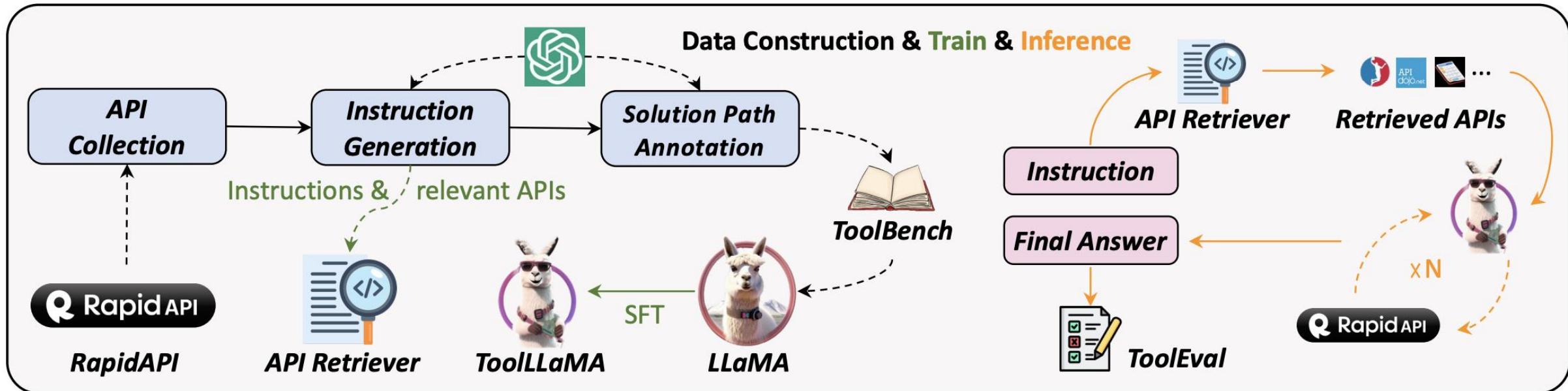
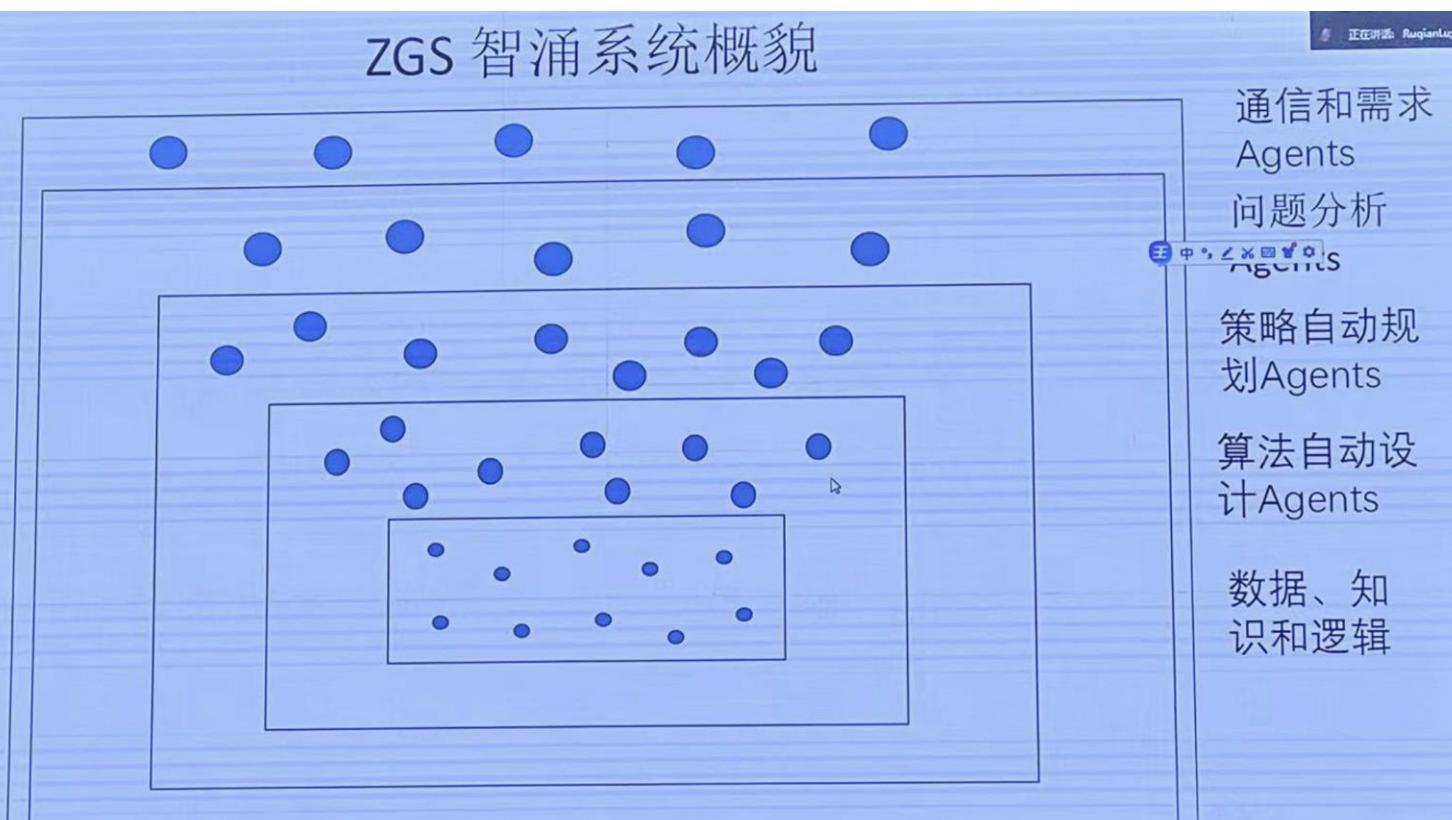


Figure 1: Three phases of constructing ToolBench and how we train our API retriever and ToolLLaMA. During inference of an instruction, the API retriever recommends relevant APIs to ToolLLaMA, which performs multiple rounds of API calls to derive the final answer. The whole reasoning process is evaluated by ToolEval.

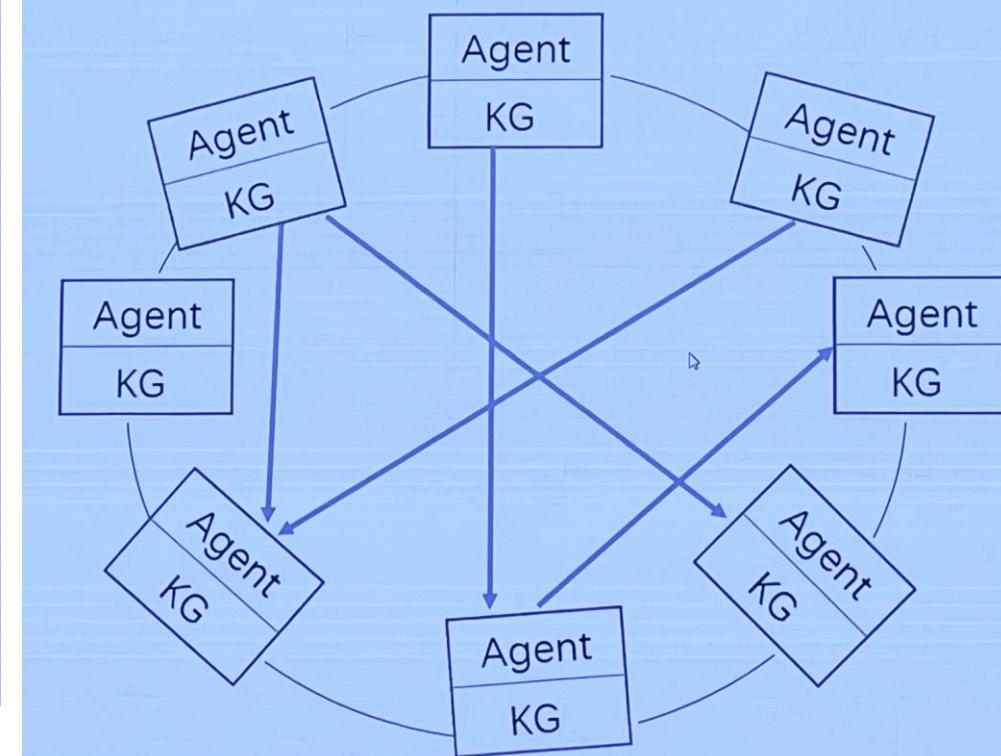
Qin, Yujia, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin et al. "**ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs.**" *arXiv preprint arXiv:2307.16789* (2023).

Generative Agents Linking

ZGS 智涌系统概貌



知识图谱网群上的联合推理、计算和辩论



陆院士, CCL2023 keynote, 2023 08

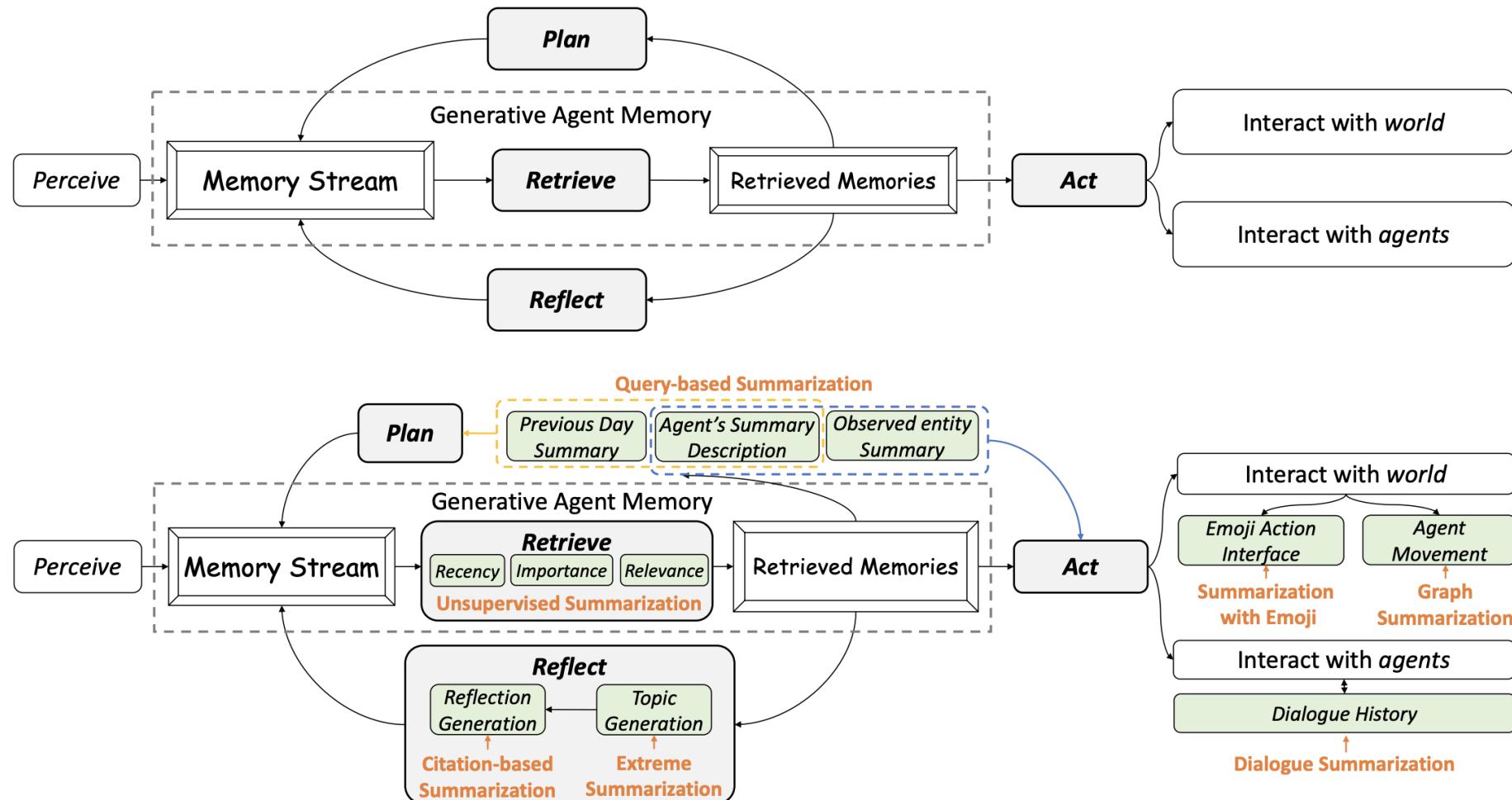
Generative Agents Linking

Link the world agents via generation.



Summarization

Xiachong Feng, Xiaocheng Feng, and Bing Qin. "The Role of **Summarization** in Generative Agents: A Preliminary Perspective." *arXiv preprint arXiv:2305.01253* (2023).



8/28/2023 Figure 1: Illustration of the generative agent architecture and key summarization techniques inside the architecture.

Generative Agents Linking



Figure 1: Generative agents create believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents they plan their days, share news, form relationships, and coordinate group activities.

THANKS!

pjli@nuaa.edu.cn