大模型的应用

报告人：李英杰

报告时间：2023.06.25

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

1. 生成特性

2. 控制规划

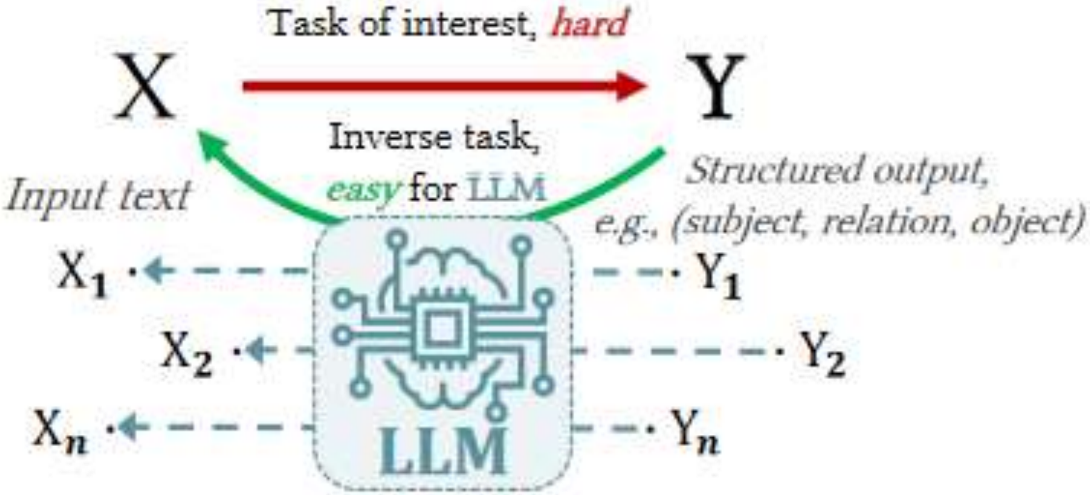
3. 零样本

4. 风险

生成特性——数据增强

Exploiting Asymmetry for Synthetic Training Data Generation: SynthIE and the Case of Information Extraction

Martin Josifoski, Marija Šakota, Maxime Peyrard, Robert West EPFL
{martin.josifoski, marija.sakota, maxime.peyrard, robert.west}@epfl.ch


Task of interest, *hard*
X
Y
Inverse task, *easy* for LLM
Input text
Structured output, e.g., (subject, relation, object)
X_1
X_2
X_n
Y_1
Y_2
Y_n
LLM

动机

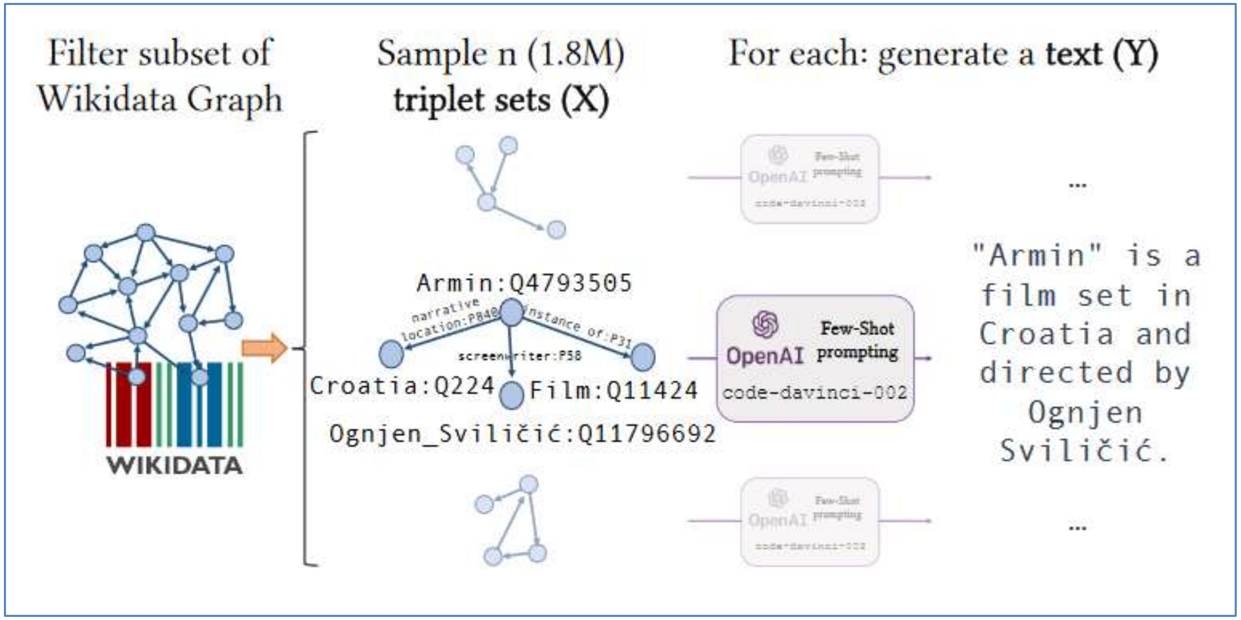
- 对于大模型而言，生成文本比抽取信息要容易，即X→Y和Y→X的不对称性
- 已有数据集的长尾问题、噪声问题

生成特性——数据增强

Exploiting Asymmetry for Synthetic Training Data Generation: SynthIE and the Case of Information Extraction

方法

- 对维基百科的知识图谱进行采样，使实体和关系分布更均匀（解决长尾问题）
- 使用大模型（GPT-3.5）进行实体三元组→句子的生成，修改prompt进行zero-shot或者提供few-samples，人工评测生成质量



Filter subset of Wikidata Graph

Sample n (1.8M) **triplet sets** (X)

For each: generate a **text** (Y)

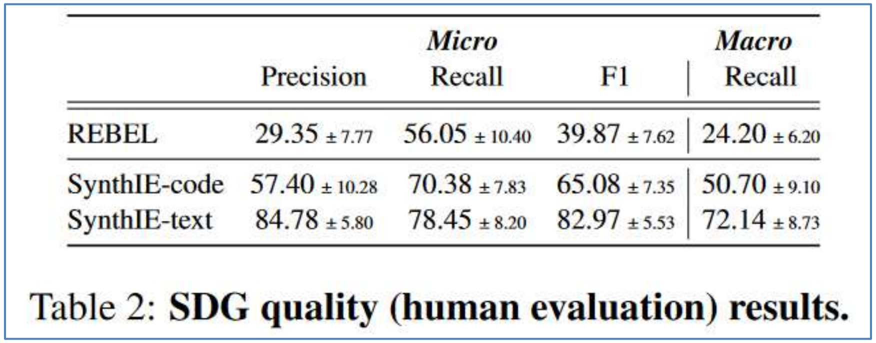
"Armin" is a film set in Croatia and directed by Ognjen Sviličić.

生成特性——数据增强

Exploiting Asymmetry for Synthetic Training Data Generation: SynthIE and the Case of Information Extraction

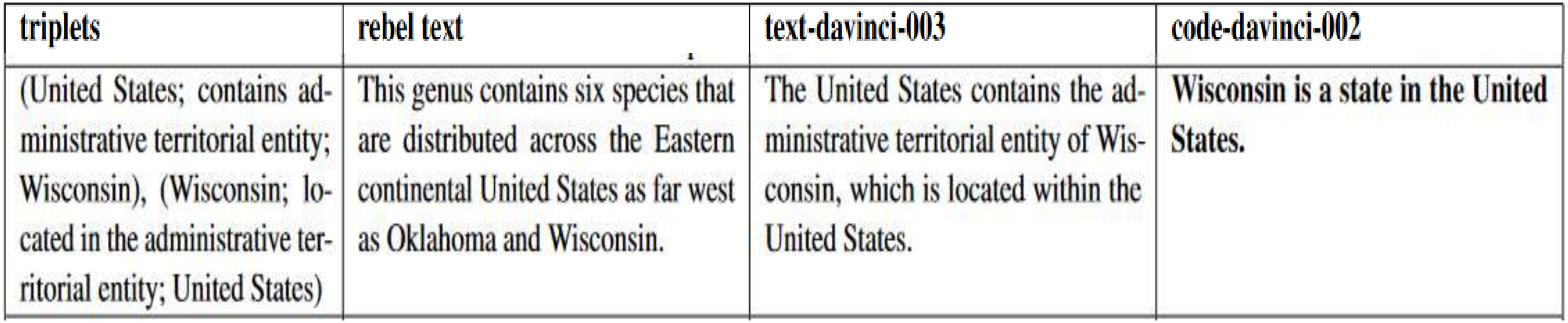
效果（生成数据的部分）

- 对比已有数据集REBEL（半自动构建）
- 评估标准除了召回率和准确率，还有人工的流畅度评价



	Micro			*Macro*
	Precision	Recall	F1	Recall
REBEL	29.35 ± 7.77	56.05 ± 10.40	39.87 ± 7.62	24.20 ± 6.20
SynthIE-code	57.40 ± 10.28	70.38 ± 7.83	65.08 ± 7.35	50.70 ± 9.10
SynthIE-text	84.78 ± 5.80	78.45 ± 8.20	82.97 ± 5.53	72.14 ± 8.73

Table 2: **SDG quality (human evaluation) results.**



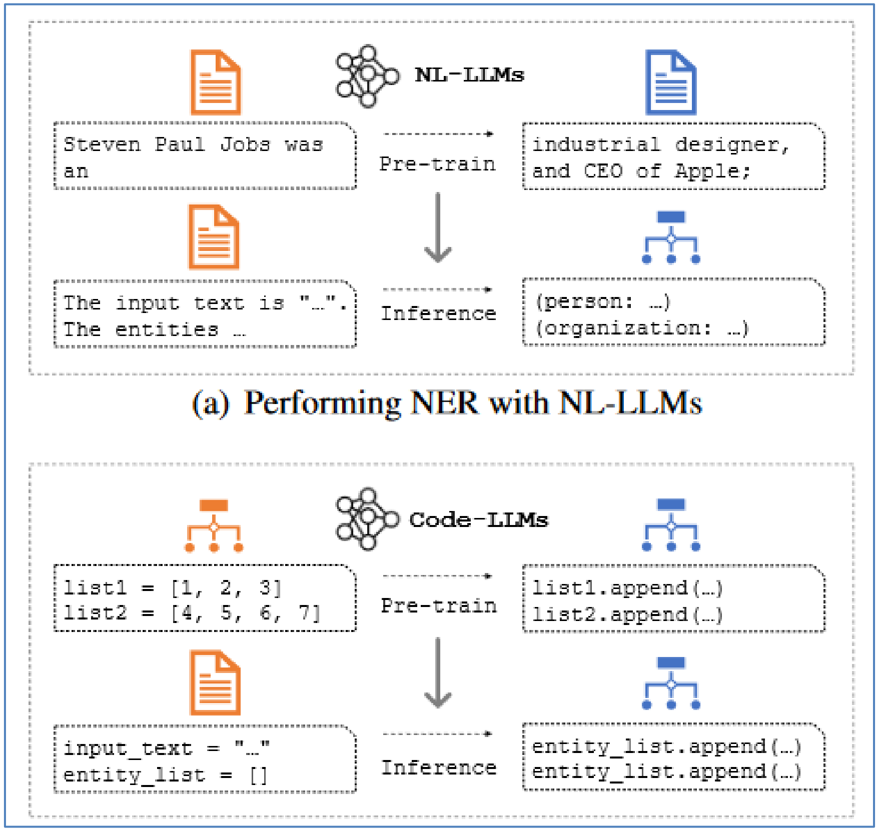
triplets	rebel text	text-davinci-003	code-davinci-002
(United States; contains administrative territorial entity; Wisconsin), (Wisconsin; located in the administrative territorial entity; United States)	This genus contains six species that are distributed across the Eastern continental United States as far west as Oklahoma and Wisconsin.	The United States contains the administrative territorial entity of Wisconsin, which is located within the United States.	Wisconsin is a state in the United States.

生成特性-输出形式与预训练一致

CODEIE: Large Code Generation Models are Better Few-Shot Information Extractors

动机

- 大模型（decoder-only）的预训练为在自然文本或代码（结构化文本）中预测下一个token，而信息抽取任务是自然文本→结构化文本，存在GAP



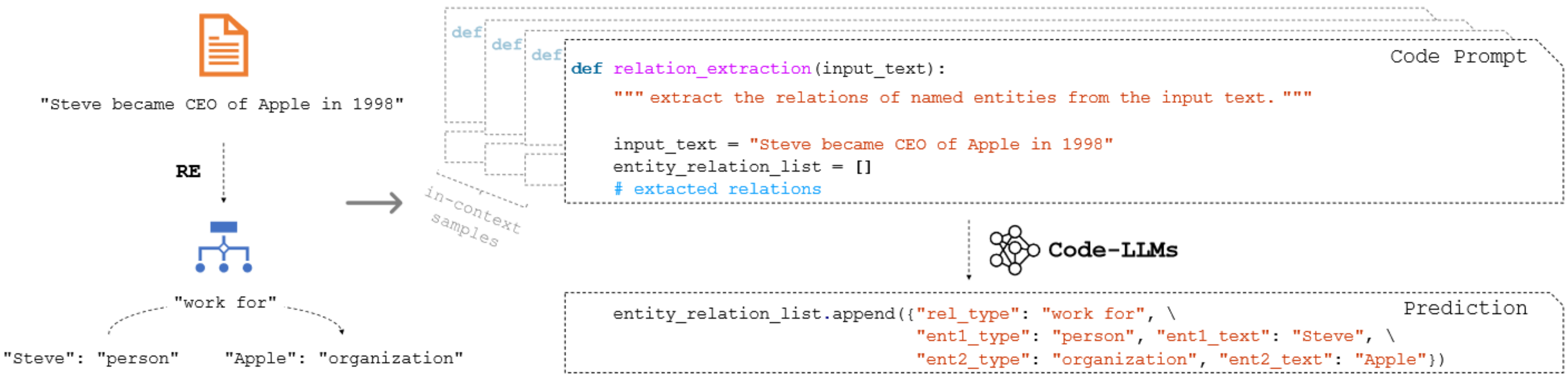
(a) Performing NER with NL-LLMs

生成特性-输出形式与预训练一致

CODEIE: Large Code Generation Models are Better Few-Shot Information Extractors

方法

- 将prompt设计为代码生成的格式（代码形式、注释、函数名、变量名都尽可能贴近任务）



```
def relation_extraction(input_text):
    """ extract the relations of named entities from the input text. """

    input_text = "Steve became CEO of Apple in 1998"
    entity_relation_list = []
    # extacted relations
```

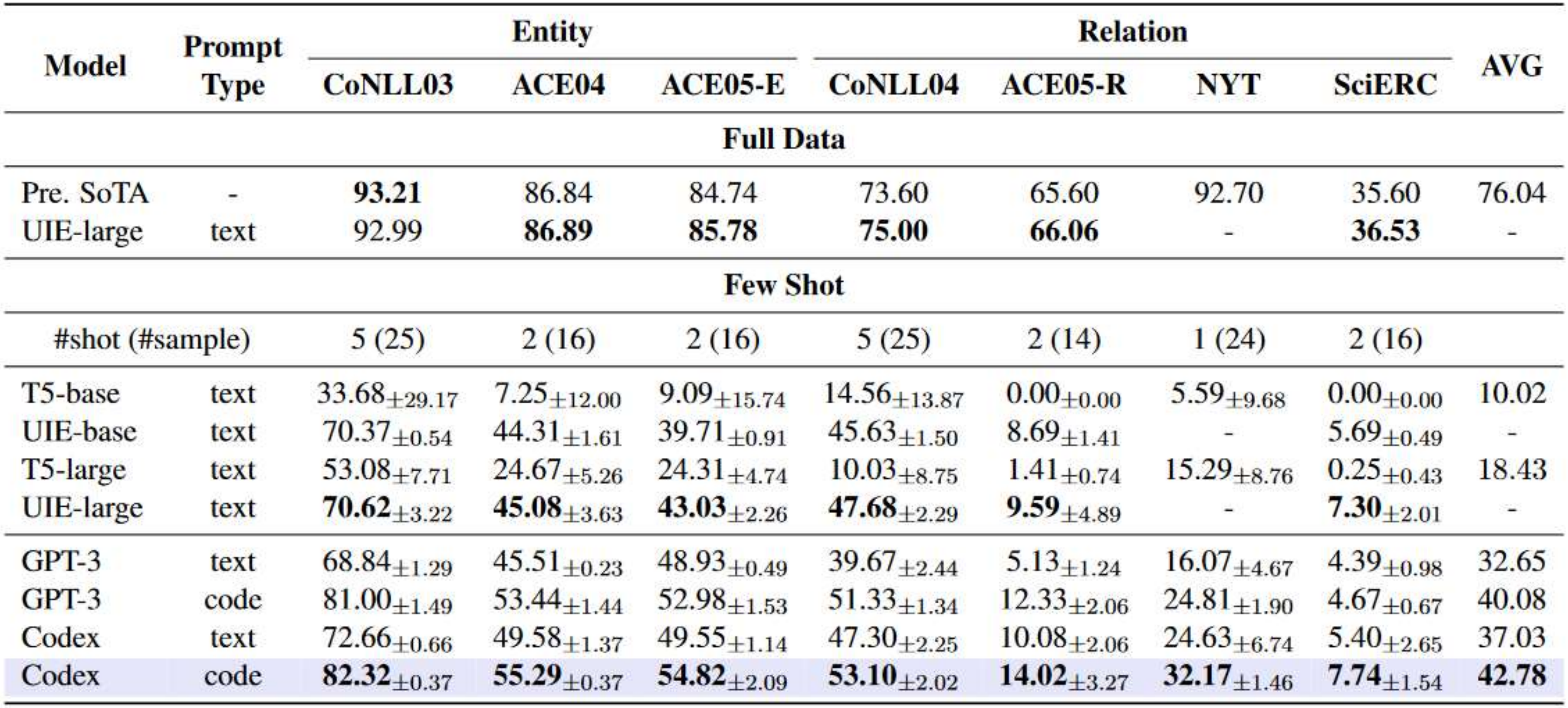
```
    entity_relation_list.append({"rel_type": "work for", \
                                 "ent1_type": "person", "ent1_text": "Steve", \
                                 "ent2_type": "organization", "ent2_text": "Apple"})
```

(b) Converting RE into code generation task

生成特性-输出形式与预训练一致

效果

- 验证了命名实体识别和关系抽取任务上的效果（事件可能更加复杂）
- 在GPT-3和CodeX上的few-shot任务均有明显提升



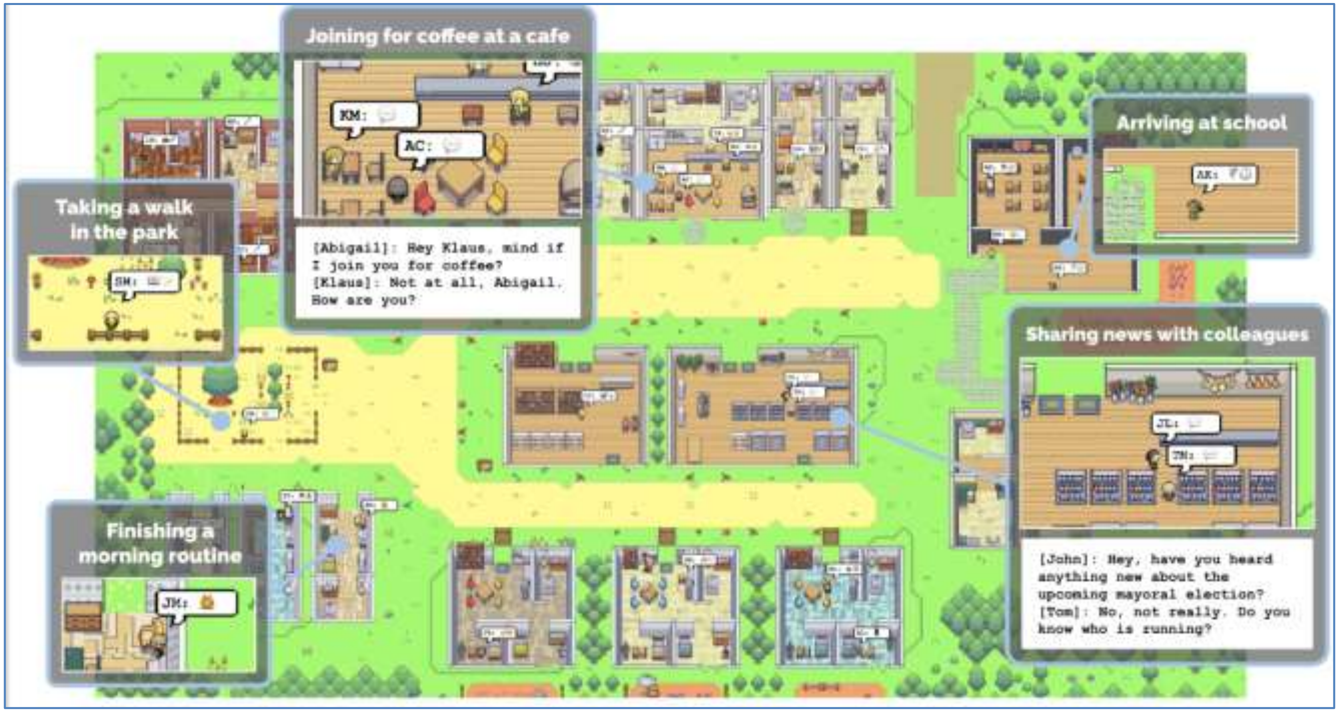
Model	Prompt Type	Entity			Relation				AVG
		CoNLL03	ACE04	ACE05-E	CoNLL04	ACE05-R	NYT	SciERC	
					Full Data				
Pre. SoTA	-	**93.21**	86.84	84.74	73.60	65.60	92.70	35.60	76.04
UIE-large	text	92.99	**86.89**	**85.78**	**75.00**	**66.06**	-	**36.53**	-
					Few Shot				
#shot (#sample)		5 (25)	2 (16)	2 (16)	5 (25)	2 (14)	1 (24)	2 (16)	
T5-base	text	33.68$_{\pm 29.17}$	7.25$_{\pm 12.00}$	9.09$_{\pm 15.74}$	14.56$_{\pm 13.87}$	0.00$_{\pm 0.00}$	5.59$_{\pm 9.68}$	0.00$_{\pm 0.00}$	10.02
UIE-base	text	70.37$_{\pm 0.54}$	44.31$_{\pm 1.61}$	39.71$_{\pm 0.91}$	45.63$_{\pm 1.50}$	8.69$_{\pm 1.41}$	-	5.69$_{\pm 0.49}$	-
T5-large	text	53.08$_{\pm 7.71}$	24.67$_{\pm 5.26}$	24.31$_{\pm 4.74}$	10.03$_{\pm 8.75}$	1.41$_{\pm 0.74}$	15.29$_{\pm 8.76}$	0.25$_{\pm 0.43}$	18.43
UIE-large	text	**70.62**$_{\pm 3.22}$	**45.08**$_{\pm 3.63}$	**43.03**$_{\pm 2.26}$	**47.68**$_{\pm 2.29}$	**9.59**$_{\pm 4.89}$	-	**7.30**$_{\pm 2.01}$	-
GPT-3	text	68.84$_{\pm 1.29}$	45.51$_{\pm 0.23}$	48.93$_{\pm 0.49}$	39.67$_{\pm 2.44}$	5.13$_{\pm 1.24}$	16.07$_{\pm 4.67}$	4.39$_{\pm 0.98}$	32.65
GPT-3	code	81.00$_{\pm 1.49}$	53.44$_{\pm 1.44}$	52.98$_{\pm 1.53}$	51.33$_{\pm 1.34}$	12.33$_{\pm 2.06}$	24.81$_{\pm 1.90}$	4.67$_{\pm 0.67}$	40.08
Codex	text	72.66$_{\pm 0.66}$	49.58$_{\pm 1.37}$	49.55$_{\pm 1.14}$	47.30$_{\pm 2.25}$	10.08$_{\pm 2.06}$	24.63$_{\pm 6.74}$	5.40$_{\pm 2.65}$	37.03
Codex	code	**82.32**$_{\pm 0.37}$	**55.29**$_{\pm 0.37}$	**54.82**$_{\pm 2.09}$	**53.10**$_{\pm 2.02}$	**14.02**$_{\pm 3.27}$	**32.17**$_{\pm 1.46}$	**7.74**$_{\pm 1.54}$	**42.78**

控制规划

Generative Agents: Interactive Simulacra of Human Behavior

简介

- 构建一个可交互的虚拟环境，其中有25个agent，不同于强化学习的“从零开始”根据奖励函数进行学习，使用大模型辅助维持环境中agent与环境，agent与agent的交互


Joining for coffee at a cafe
KM:
AC:
Taking a walk in the park
SM:
Arriving at school
AK:
Sharing news with colleagues
JL:
TM:
Finishing a morning routine
JM:

[Abigail]: Hey Klaus, mind if I join you for coffee?
[Klaus]: Not at all, Abigail. How are you?

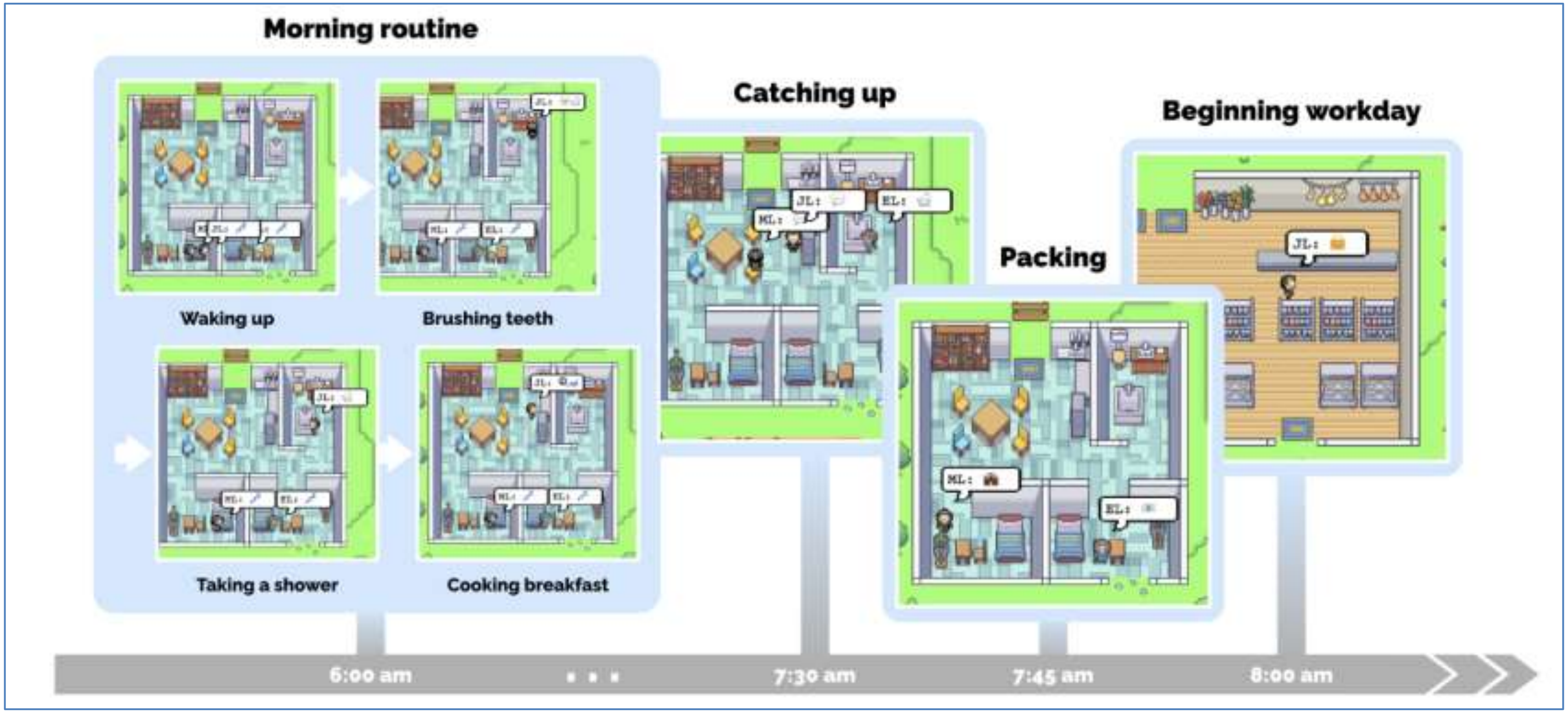
[John]: Hey, have you heard anything new about the upcoming mayoral election?
[Tom]: No, not really. Do you know who is running?

控制规划

Generative Agents: Interactive Simulacra of Human Behavior

简介

- 每个agent根据自己的人物设定、与环境和其他agent的交互，用户输入的命令进行动作

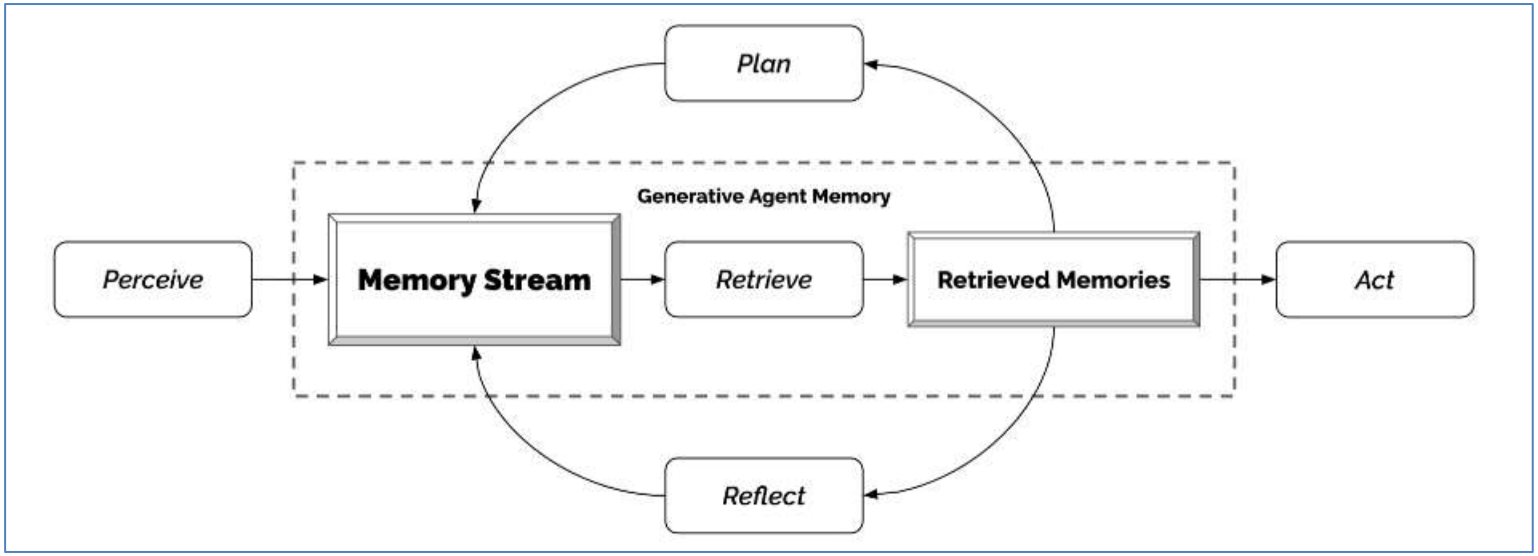


控制规划

Generative Agents: Interactive Simulacra of Human Behavior

架构

- 记忆流：存储长期记忆，使用检索模型检索记忆
- 反思：从记忆中进行总结
- 规划：根据记忆进行规划



控制规划

Generative Agents: Interactive Simulacra of Human Behavior

介绍

- 每个agent有一个自身身份描述的种子记忆
- 架构中所有内容使用自然语言进行描述和推理（即把一个虚拟世界映射到自然语言言文本中)

个体的种子记忆示例：

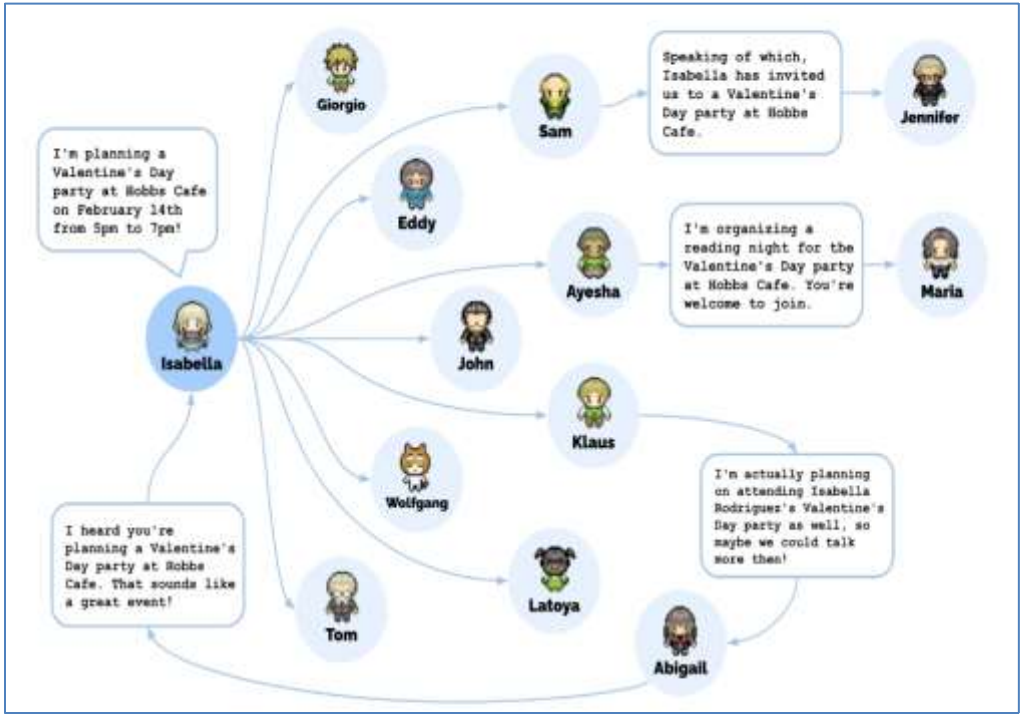
约翰·林（John Lin）是柳树市场和药房的一名药房店主，他喜欢帮助别人。他一直在寻找让顾客更轻松地获得药物的方法；约翰·林（John Lin）与他的妻子梅·林（大学教授）和儿子艾迪·林（Eddy Lin）住在一起，他是一名学习音乐理论的学生。约翰·林非常爱他的家人；约翰·林（John Lin）认识隔壁的老夫妇萨姆·摩尔（Sam Moore）和詹妮弗·摩尔（Jennifer Moore）已经好几年了；约翰·林（John Lin）认为萨姆·摩尔（Sam Moore）是一个善良、善良的人；约翰·林（John Lin）很了解他的邻居山本百合子（Yuriko Yamamoto）。约翰·林 (John Lin) 认识他的邻居塔玛拉·泰勒 (Tamara Taylor) 和卡门·奥尔蒂斯 (Carmen Ortiz)，但之前从未见过他们； John Lin 和 Tom Moreno 是 The Willows Market and Pharmacy 的同事；John Lin 和 Tom Moreno 是朋友，喜欢一起讨论当地政治；约翰·林（John Lin）对莫雷诺一家很了解——丈夫汤姆·莫雷诺（Tom Moreno）和妻子简·莫雷诺（Jane Moreno）。

控制规划

Generative Agents: Interactive Simulacra of Human Behavior

意义

- 使用大模型从观察和互动中得到更抽象的高阶表述（如一个agent经常进行阅读和整理资料，大模型可以从记忆中总结出“痴迷研究”）
- 大模型的能力足够支撑起一个个体的身份属性和互动（需要记忆管理等模块的支持）
- 大模型的探索能力?


I'm planning a Valentine's Day party at Hobbs Cafe on February 14th from 5pm to 7pm!
Giorgio
Sam
Speaking of which, Isabella has invited us to a Valentine's Day party at Hobbs Cafe.
Jennifer
Eddy
Ayesha
I'm organizing a reading night for the Valentine's Day party at Hobbs Cafe. You're welcome to join.
Maria
Isabella
John
Klaus
Wolfgang
I'm actually planning on attending Isabella Rodriguez's Valentine's Day party as well, so maybe we could talk more then!
I heard you're planning a Valentine's Day party at Hobbs Cafe. That sounds like a great event!
Tom
Latoya
Abigail

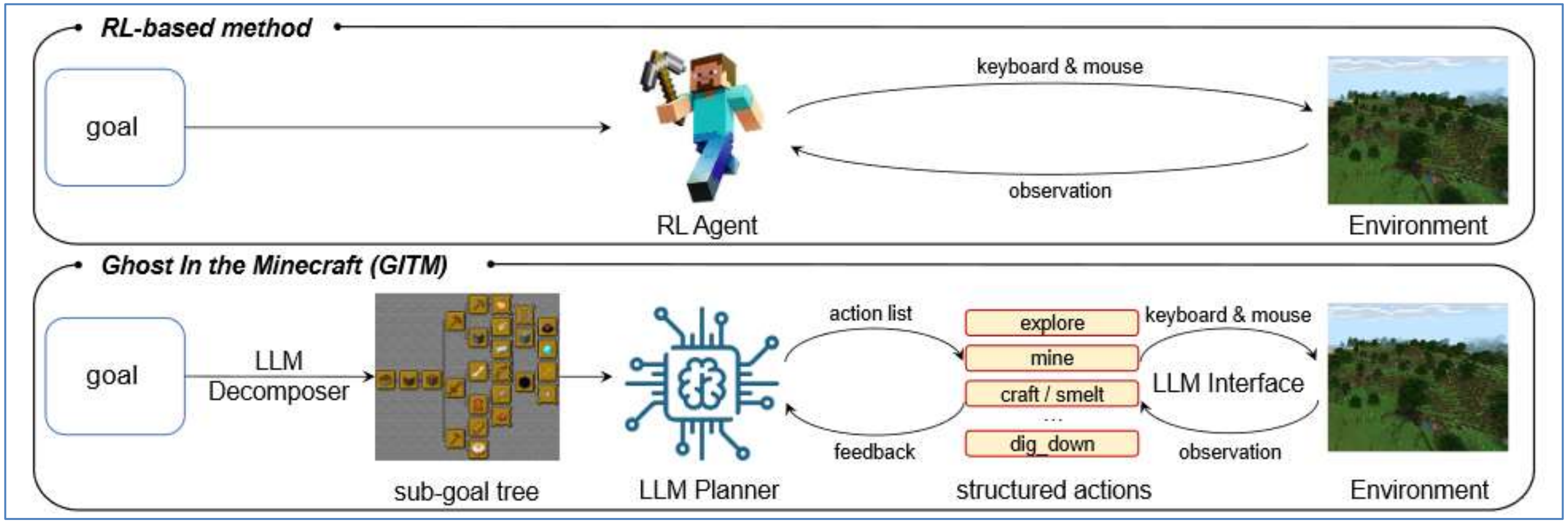
用户输入要求一个agent组织派对后消息的传播链条，共12个agent得到了信息

控制规划

Ghost in the Minecraft: Generally Capable Agents for Open-World Enviroments via Large Language Models with Text-based Knowledge and Memory

简介

- 我的世界——沙盒游戏，可以控制人物进行探索、与环境互动，收集、合成材料等
- 强化学习的范式：设置目标及奖励函数（如获得钻石、合成武器等），奖励稀疏、训练很不稳定

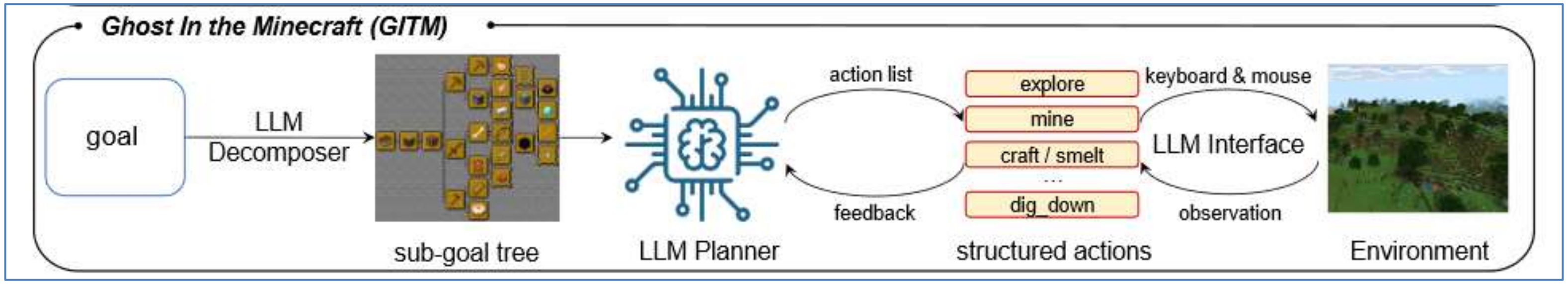


控制规划

Ghost in the Minecraft: Generally Capable Agents for Open-World Enviroments via Large Language Models with Text-based Knowledge and Memory

方法

- 先验知识：大模型本身的预训练获得的能力，网络上关于Minecraft的文字攻略，人工设计的一些任务分解模板、游戏中的动作及其描述
- 使用大模型进行规划和反馈，使用人工设计的一些接口、模板实现与环境的交互反馈

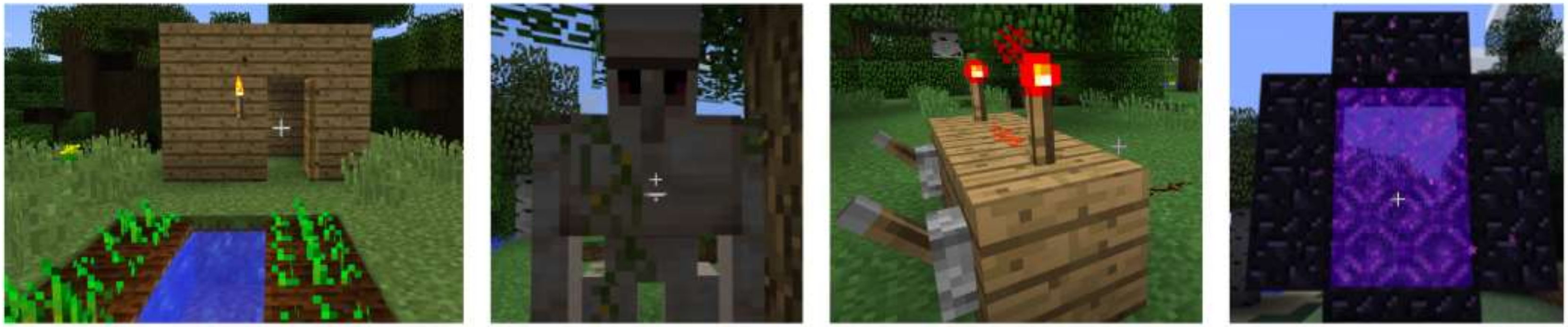


控制规划

Ghost in the Minecraft: Generally Capable Agents for Open-World Enviroments via Large Language Models with Text-based Knowledge and Memory

效果

- 模型解锁了之前相关研究中全部物品的合成路径，之前的方法仅30%
- 实现了在该游戏中进行生存和探索的一些基础任务——庇护所、农田、红石电路（自动化设备）等



(a) Shelter with Farmland (b) Iron Golem (c) Redstone Circuit (d) Nether Portal

控制规划

Ghost in the Minecraft: Generally Capable Agents for Open-World Enviroments via Large Language Models with Text-based Knowledge and Memory

意义

- 同样证明了只要能够将虚拟/现实世界的控制、交互等模式转化为自然语言，大模型就能够发挥很好的规划、总结能力
- 相比强化学习的方法，由于大模型输入输出均为自然语言，能够非常好的应用先验知识，该论文中很多材料的合成路径、寻找方法均是从一些攻略中以文本形式输入大模型的

零样本（少样本）

Is Information Extraction Solved by ChatGPT? An Analysis of Performance, Evaluation Criteria, Robustness and Errors

结论

- ChatGPT和SOTA之间有明显差距
- 任务越难，差距越大
- 在一些简单任务/情况下能够超越SOTA
- ICL上下文学习能够带来一定改进
- CoT思维链不一定能够在ICL基础上带来进一步提升

零样本（少样本）

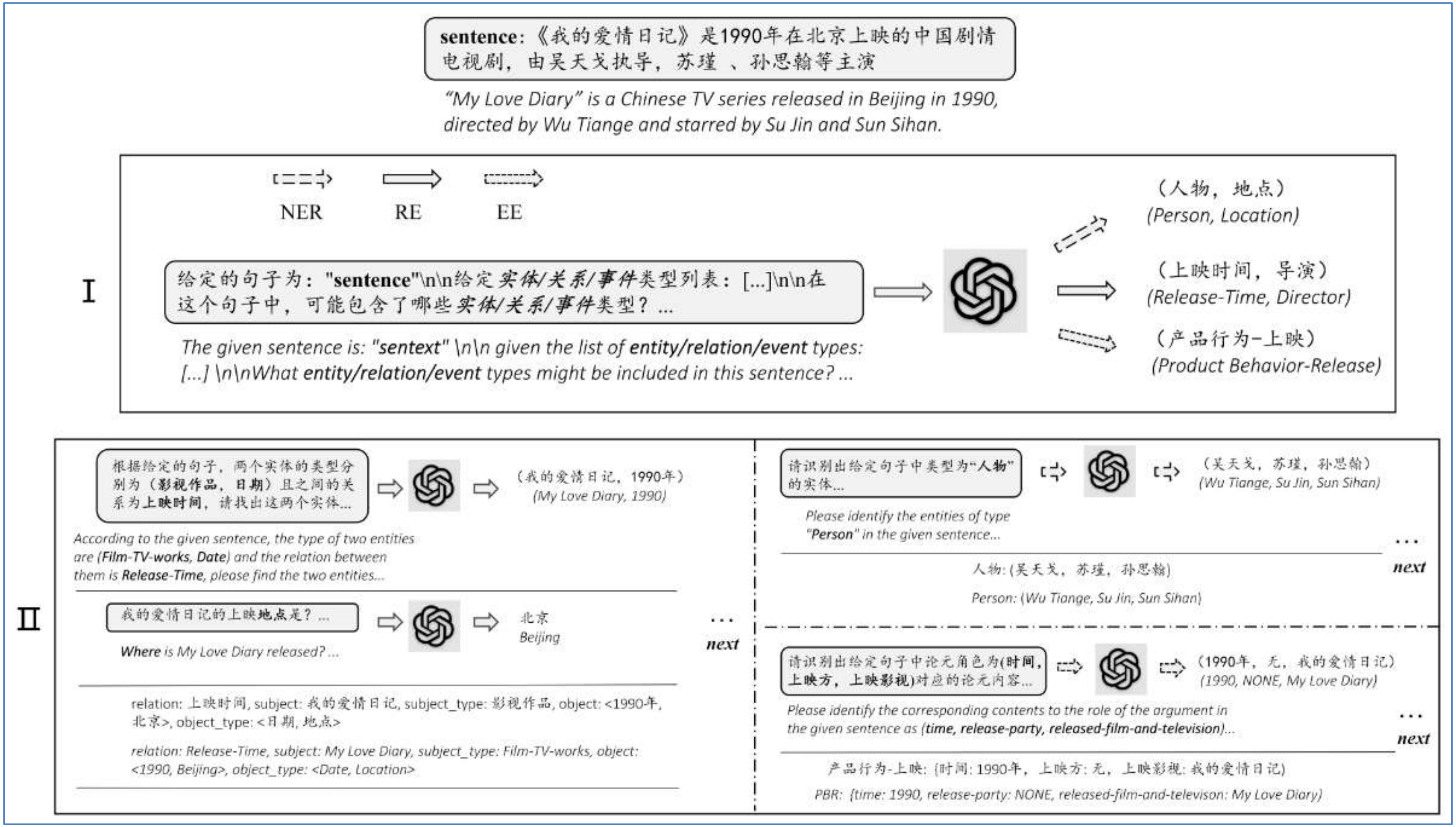
Zero-Shot Information Extraction via Chatting with ChatGPT

方法

- 两段式抽取：（类似思维链方法）
- 1、根据任务构造初始问题，要求大模型找出相关的实体、事件类型等
- 2、根据不同任务和实体、事件类型进行多轮QA，逐步结构化答案

零样本（少样本）

Zero-Shot Information Extraction via Chatting with ChatGPT



sentence:《我的爱情日记》是1990年在北京上映的中国剧情电视剧，由吴天戈执导，苏瑾 、孙思翰等主演

"My Love Diary" is a Chinese TV series released in Beijing in 1990, directed by Wu Tiange and starred by Su Jin and Sun Sihan.

给定的句子为："**sentence**"\n\n给定***实体/关系/事件***类型列表：[...]\n\n在这个句子中，可能包含了哪些***实体/关系/事件***类型？...

*The given sentence is: "**sentext**" \n\n given the list of **entity/relation/event** types: [...] \n\nWhat **entity/relation/event** types might be included in this sentence? ...*

（人物，地点）
(Person, Location)

（上映时间，导演）
(Release-Time, Director)

（产品行为-上映）
(Product Behavior-Release)

根据给定的句子，两个实体的类型分别为（**影视作品，日期**）且之间的关系为**上映时间**，请找出这两个实体...

（我的爱情日记，1990年）
(My Love Diary, 1990)

*According to the given sentence, the type of two entities are (**Film-TV-works, Date**) and the relation between them is **Release-Time**, please find the two entities...*

我的爱情日记的上映**地点**是？...

北京
Beijing

***Where** is My Love Diary released? ...*

relation: 上映时间, subject: 我的爱情日记, subject_type: 影视作品, object: <1990年, 北京>, object_type: <日期, 地点>

relation: Release-Time, subject: My Love Diary, subject_type: Film-TV-works, object: <1990, Beijing>, object_type: <Date, Location>

请识别出给定句子中类型为**"人物"**的实体...

（吴天戈，苏瑾，孙思翰）
(Wu Tiange, Su Jin, Sun Sihan)

*Please identify the entities of type "**Person**" in the given sentence...*

人物:（吴天戈，苏瑾，孙思翰）

Person: (Wu Tiange, Su Jin, Sun Sihan)

请识别出给定句子中论元角色为(**时间，上映方，上映影视**)对应的论元内容...

（1990年，无，我的爱情日记）
(1990, NONE, My Love Diary)

*Please identify the corresponding contents to the role of the argument in the given sentence as (**time, release-party, released-film-and-television**)...*

产品行为-上映:（时间: 1990年，上映方: 无，上映影视: 我的爱情日记）

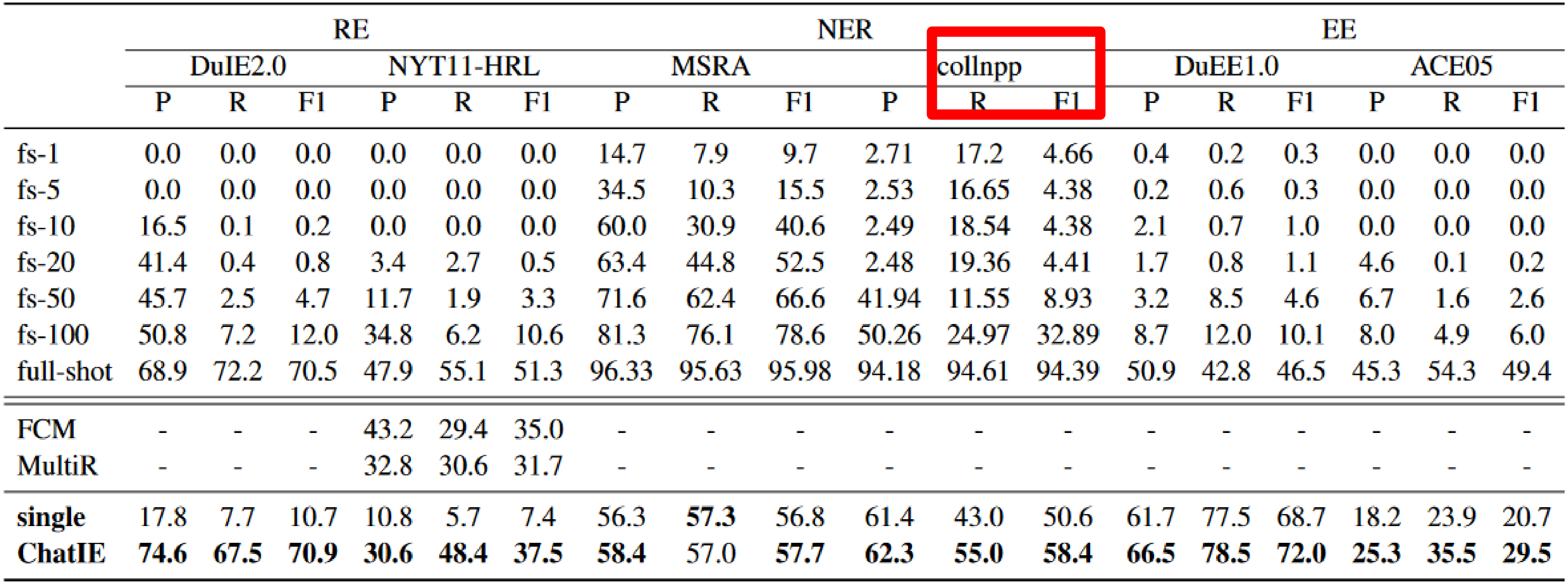
PBR: {time: 1990, release-party: NONE, released-film-and-televison: My Love Diary)

零样本（少样本）

Zero-Shot Information Extraction via Chatting with ChatGPT

结果

- 上面的为已有小样本方法的SOTA，可以看出ChatGPT仅用zero-shot就相比已有小样本方法的巨大提升（？）



	RE						NER						EE					
	DuIE2.0			NYT11-HRL			MSRA			collnpp			DuEE1.0			ACE05		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
fs-1	0.0	0.0	0.0	0.0	0.0	0.0	14.7	7.9	9.7	2.71	17.2	4.66	0.4	0.2	0.3	0.0	0.0	0.0
fs-5	0.0	0.0	0.0	0.0	0.0	0.0	34.5	10.3	15.5	2.53	16.65	4.38	0.2	0.6	0.3	0.0	0.0	0.0
fs-10	16.5	0.1	0.2	0.0	0.0	0.0	60.0	30.9	40.6	2.49	18.54	4.38	2.1	0.7	1.0	0.0	0.0	0.0
fs-20	41.4	0.4	0.8	3.4	2.7	0.5	63.4	44.8	52.5	2.48	19.36	4.41	1.7	0.8	1.1	4.6	0.1	0.2
fs-50	45.7	2.5	4.7	11.7	1.9	3.3	71.6	62.4	66.6	41.94	11.55	8.93	3.2	8.5	4.6	6.7	1.6	2.6
fs-100	50.8	7.2	12.0	34.8	6.2	10.6	81.3	76.1	78.6	50.26	24.97	32.89	8.7	12.0	10.1	8.0	4.9	6.0
full-shot	68.9	72.2	70.5	47.9	55.1	51.3	96.33	95.63	95.98	94.18	94.61	94.39	50.9	42.8	46.5	45.3	54.3	49.4
FCM	-	-	-	43.2	29.4	35.0	-	-	-	-	-	-	-	-	-	-	-	-
MultiR	-	-	-	32.8	30.6	31.7	-	-	-	-	-	-	-	-	-	-	-	-
single	17.8	7.7	10.7	10.8	5.7	7.4	56.3	**57.3**	56.8	61.4	43.0	50.6	61.7	77.5	68.7	18.2	23.9	20.7
ChatIE	**74.6**	**67.5**	**70.9**	**30.6**	**48.4**	**37.5**	**58.4**	57.0	**57.7**	**62.3**	**55.0**	**58.4**	**66.5**	**78.5**	**72.0**	**25.3**	**35.5**	**29.5**

其他应用

各种工具产品(autoGPT, 翻译，论文润色等)

Tree of thought (https://arxiv.org/abs/2305.10601)

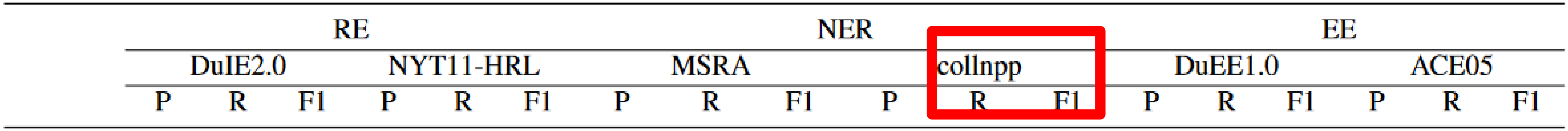
多种大模型和专用模型的组合（Stitchable Neural Networks)

垂直领域（医疗、法律等)

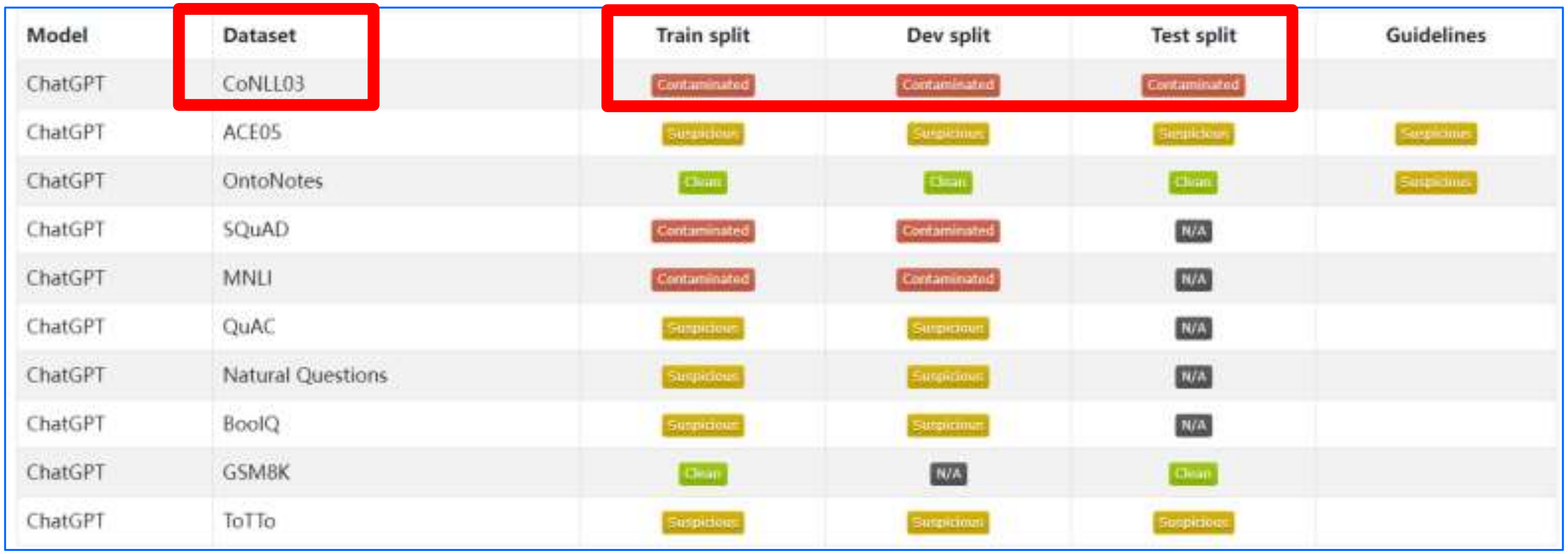
…

风险（科研角度的）

Did ChatGPT cheat on your test?
https://hitz-zentroa.github.io/lm-contamination/



	RE						NER						EE					
	DuIE2.0			NYT11-HRL			MSRA			collnpp			DuEE1.0			ACE05		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1



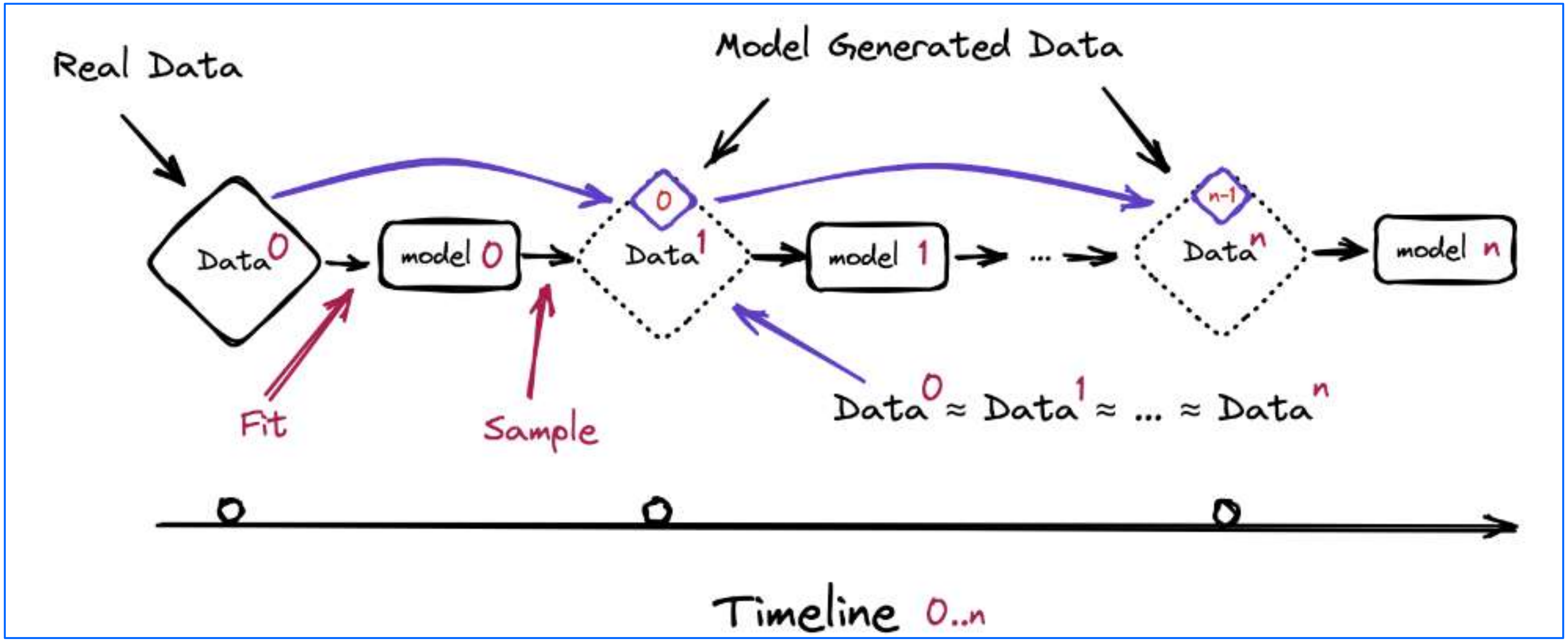
Model	Dataset	Train split	Dev split	Test split	Guidelines
ChatGPT	CoNLL03	Contaminated	Contaminated	Contaminated	
ChatGPT	ACE05	Suspicious	Suspicious	Suspicious	Suspicious
ChatGPT	OntoNotes	Clean	Clean	Clean	Suspicious
ChatGPT	SQuAD	Contaminated	Contaminated	N/A	
ChatGPT	MNLI	Contaminated	Contaminated	N/A	
ChatGPT	QuAC	Suspicious	Suspicious	N/A	
ChatGPT	Natural Questions	Suspicious	Suspicious	N/A	
ChatGPT	BoolQ	Suspicious	Suspicious	N/A	
ChatGPT	GSM8K	Clean	N/A	Clean	
ChatGPT	ToTTo	Suspicious	Suspicious	Suspicious	

- 该数据集很可能已经被泄露了——不给出示例ChatGPT即可生成数据格式和真实数据

风险（科研角度的）

CURSE OF RECURSION: TRAINING ON GENERATED DATA MAKES MODELS FORGET

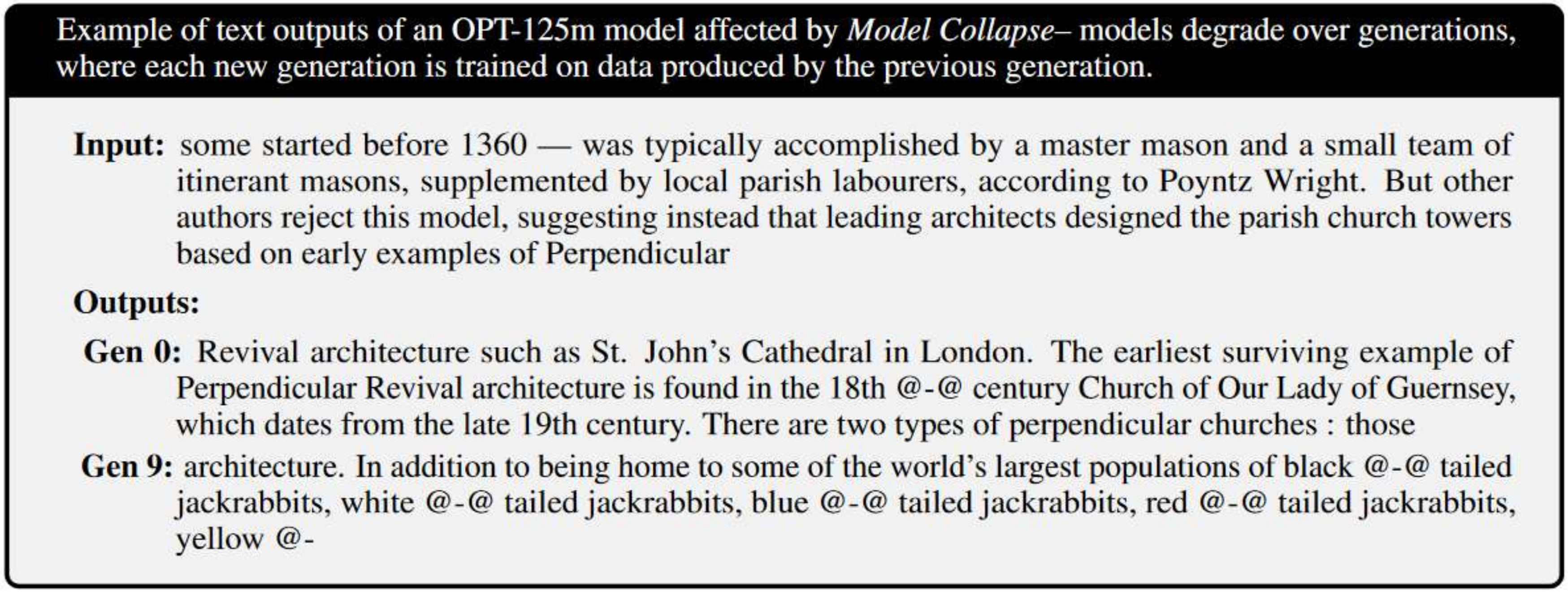
- 理论上生成数据的分布和原始数据是相同的，但是即使在分布不变的情况下使用生成数据训练多轮后仍会导致模型崩溃



风险（科研角度的）

CURSE OF RECURSION: TRAINING ON GENERATED DATA MAKES MODELS FORGET

- 与灾难性遗忘的区别：持续学习中的灾难性遗忘是失去初始分布，模型崩溃仍旧会有原始数据的分布，但是输出会有错误不再可用
- 微调会导致模型有崩溃的初步迹象（困惑度升高）



Example of text outputs of an OPT-125m model affected by *Model Collapse*– models degrade over generations, where each new generation is trained on data produced by the previous generation.

Input: some started before 1360 — was typically accomplished by a master mason and a small team of itinerant masons, supplemented by local parish labourers, according to Poyntz Wright. But other authors reject this model, suggesting instead that leading architects designed the parish church towers based on early examples of Perpendicular

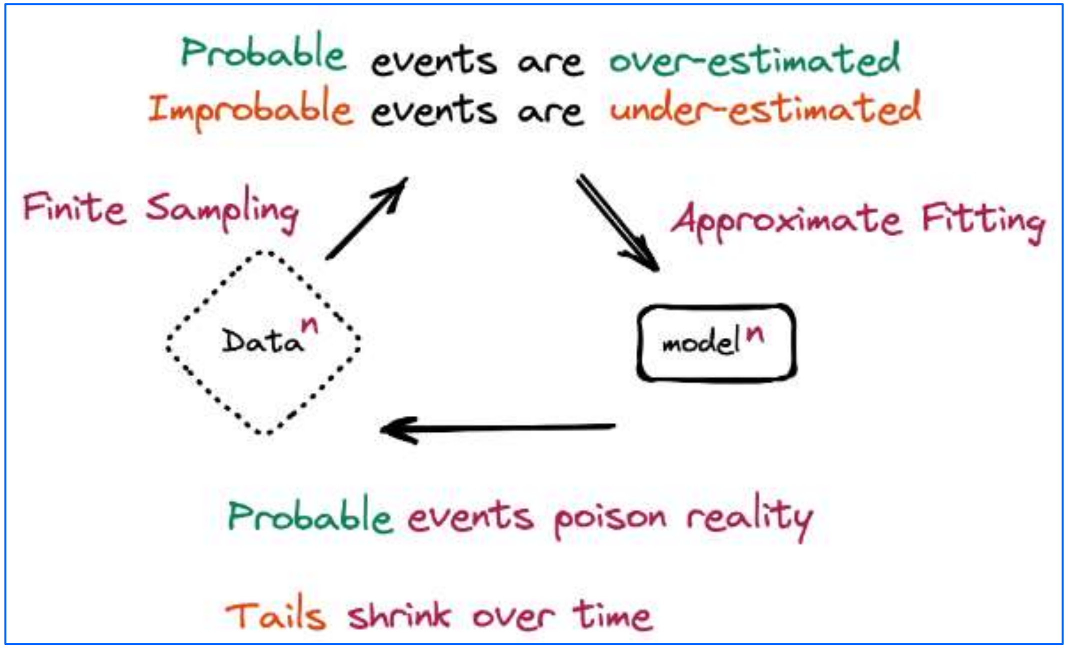
Outputs:

Gen 0: Revival architecture such as St. John's Cathedral in London. The earliest surviving example of Perpendicular Revival architecture is found in the 18th @-@ century Church of Our Lady of Guernsey, which dates from the late 19th century. There are two types of perpendicular churches : those

Gen 9: architecture. In addition to being home to some of the world's largest populations of black @-@ tailed jackrabbits, white @-@ tailed jackrabbits, blue @-@ tailed jackrabbits, red @-@ tailed jackrabbits, yellow @-

风险（科研角度的）

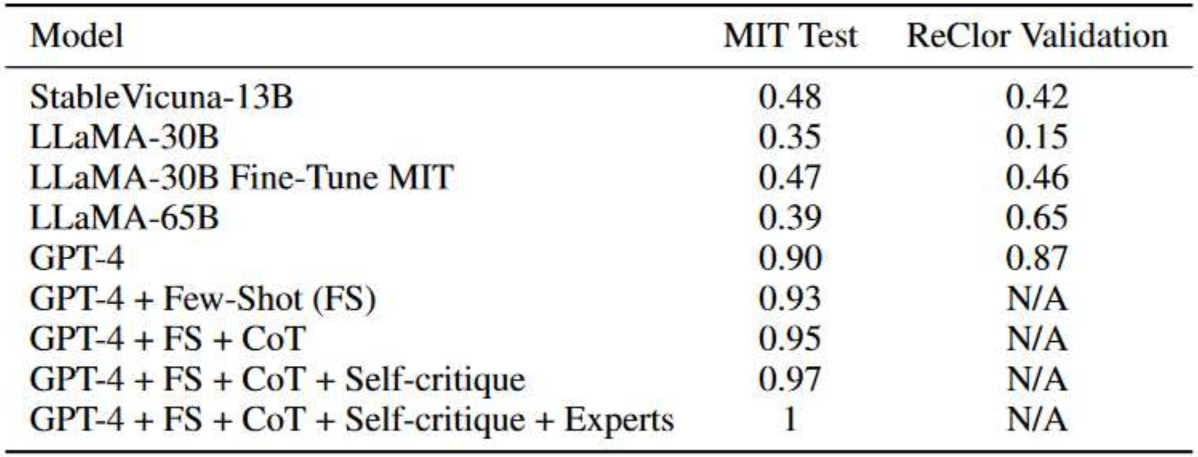
CURSE OF RECURSION:
TRAINING ON GENERATED DATA MAKES MODELS FORGET


Probable events are over-estimated
Improbable events are under-estimated
Finite Sampling
Approximate Fitting
Data^n
model^n
Probable events poison reality
Tails shrink over time

- 随着迭代次数增加，模型遗忘了分布中小概率的部分

风险（科研角度的）

Exploring the MIT Mathematics and EECS Curriculum Using Large Language Models



Model	MIT Test	ReClor Validation
StableVicuna-13B	0.48	0.42
LLaMA-30B	0.35	0.15
LLaMA-30B Fine-Tune MIT	0.47	0.46
LLaMA-65B	0.39	0.65
GPT-4	0.90	0.87
GPT-4 + Few-Shot (FS)	0.93	N/A
GPT-4 + FS + CoT	0.95	N/A
GPT-4 + FS + CoT + Self-critique	0.97	N/A
GPT-4 + FS + CoT + Self-critique + Experts	1	N/A

- 完美的结果，但是实验和评估过程中存在严重问题：
- 数据：无法解决的错误问题，重复或接近重复的问题
- Prompt流程：对于选择题仍采用了多轮prompt直到结果正确的方法
- 评估：GPT-4自身评估不准确

总结

应用大模型的技巧

- 将任务（非NLP任务）转化为自然语言，即非NLP任务→NLP任务
- 使NLP任务形式更加贴近大模型的预训练形式（instruct，结构化文本→代码）
- 复杂任务的拆解、长期记忆的管理
- prompt设计（https://github.com/datawhalechina/prompt-engineering-for-developers）
- …

总结

应用大模型需要注意的问题

- zero-shot的表述要慎重，对于没有开源训练语料的大模型，可能存在数据集泄露的情况（甚至测试集的泄露）
- 任务流程的设计的设计是否存在漏洞（信息泄露、输入长度限制等等）
- 评价的方法，尤其单纯使用GPT-4进行自我评估的方法可能缺乏合理性
- 使用生成数据进行训练可能导致的模型崩溃风险
- ...

参考文献

[1] C. Shen, L. Cheng, Y. You, and L. Bing, “Are Large Language Models Good Evaluators for Abstractive Summarization?” arXiv, May 22, 2023. doi: 10.48550/arXiv.2305.13091.

[2] P. Li et al., “CodeIE: Large Code Generation Models are Better Few-Shot Information Extractors.” arXiv, May 10, 2023. doi: 10.48550/arXiv.2305.05711.

[3] I.-H. Hsu et al., “DEGREE: A Data-Efficient Generation-Based Event Extraction Model,” in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1890–1908. doi: 10.18653/v1/2022.naacl-main.138.

[4] B. Li et al., “Evaluating ChatGPT’s Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness.” arXiv, Apr. 23, 2023. doi: 10.48550/arXiv.2304.11633.

[5] B. Li et al., “Evaluating ChatGPT’s Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness.” arXiv, Apr. 23, 2023. doi: 10.48550/arXiv.2304.11633.

[6] M. Josifoski, M. Sakota, M. Peyrard, and R. West, “Exploiting Asymmetry for Synthetic Training Data Generation: SynthIE and the Case of Information Extraction.” arXiv, Mar. 07, 2023. doi: 10.48550/arXiv.2303.04132.

[7] J. S. Park, J. C. O’Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, “Generative Agents: Interactive Simulacra of Human Behavior.” arXiv, Apr. 06, 2023. doi: 10.48550/arXiv.2304.03442.

[8] X. Wang et al., “InstructUIE: Multi-task Instruction Tuning for Unified Information Extraction.” arXiv, Apr. 17, 2023. doi: 10.48550/arXiv.2304.08085.

[9] R. Han, T. Peng, C. Yang, B. Wang, L. Liu, and X. Wan, “Is Information Extraction Solved by ChatGPT? An Analysis of Performance, Evaluation Criteria, Robustness and Errors.” arXiv, May 23, 2023. doi: 10.48550/arXiv.2305.14450.

[10] Y. Ma, Y. Cao, Y. Hong, and A. Sun, “Large Language Model Is Not a Good Few-shot Information Extractor, but a Good Reranker for Hard Samples!” arXiv, Mar. 15, 2023. doi: 10.48550/arXiv.2303.08559.

[11] Z. Pan, J. Cai, and B. Zhuang, “Stitchable Neural Networks.” arXiv, Mar. 28, 2023. doi: 10.48550/arXiv.2302.06586.

[12] Y. Liu et al., “Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models.” arXiv, Apr. 08, 2023. doi: 10.48550/arXiv.2304.01852.

[13] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, and R. Anderson, “The Curse of Recursion: Training on Generated Data Makes Models Forget.” arXiv, May 31, 2023. doi: 10.48550/arXiv.2305.17493.

[14] S. Yao et al., “Tree of Thoughts: Deliberate Problem Solving with Large Language Models.” arXiv, May 17, 2023. Accessed: Jun. 20, 2023. [Online]. Available: http://arxiv.org/abs/2305.10601

[15] X. Wei et al., “Zero-Shot Information Extraction via Chatting with ChatGPT.” arXiv, Feb. 20, 2023. doi: 10.48550/arXiv.2302.10205.

敬请批评指正

中国科学院信息工程研究所

INSTITUTE OF INFORMATION ENGINEERING,CAS