

# Is Information Extraction Solved by ChatGPT?

- 性能方面：
  - 明显低于 SOTA，任务越难，差距越大
  - few-shot in-context learning 很有用，明显提升性能
  - chain-of-thought prompting 无法进一步提升性能
- 稳定性：
  - ChatGPT 很少会输出无关回复
  - 无关上下文和标签频率对性能影响极大
  - 对实体顺序不敏感
- 错误类型
  - 绝大部分错误 (1/3) 是 Unannotated spans

# Prompts

- 任务描述、候选标签、输出格式、示例（思维链）、输入

## An example of NER task prompts

Given the list of entity types ["Organization", "Person", "Location", "Miscellaneous"], read the given sentence and find out all words/phrases that indicate the above types of named entities. Answer in the format ["entity\_type", "entity\_name"] without any explanation. If no entity exists, then just answer "[ ]".

Sentence: "Japan began the defense of their Asian Cup title with a lucky 2-1 win against Syria in a Group C championship match on Friday."

Answer: ["Location", "Japan"], ["Miscellaneous", "Asian Cup"], ["Location", "Syria"]

... (More examples are omitted here.)

Sentence: "In Home Health said it previously recorded a reserve equal to 16 percent of all revenue related to the community liaison costs."

Answer:

---

## Expected Output:

["Organization", "In Home Health"]

# 性能

Task	Dataset	SOTA	Zero-shot		5-shot ICL	5-shot COT	Ratio@SOTA
			max	mean (std)	mean (std)	mean (std)	
NER-Flat	CoNLL03	94.6(Wang et al., 2021)	65.13	60.10 (3.81)	70.53 (1.44)	74.63 (1.08)	63.5%
	FewNERD	67.1(Ding et al., 2021)	34.28	31.56 (2.44)	36.87 (0.71)	46.55 (0.64)	47.0%
NER-Nested	ACE04	88.5(Yang et al., 2023)	29.55	27.80 (3.10)	38.52 (2.51)	40.57 (1.83)	31.4%
	ACE05-Ent	87.5(Yang et al., 2023)	24.77	23.38 (1.92)	36.17 (1.78)	33.98 (0.69)	26.7%
	GENIA	81.5(Yang et al., 2023)	39.43	38.09 (1.65)	48.82 (1.31)	50.89 (1.00)	46.7%
RE-RC	CoNLL04	-	65.82	59.21 (3.85)	55.32 (4.56)	-	-
	NYT-multi	93.5(Zhan et al., 2022)	38.74	30.96 (5.51)	26.88 (2.74)	-	33.1%
	TACRED	75.6(Li et al., 2022)	21.58	19.47 (1.49)	27.84 (3.48)	-	25.8%
	SemEval2010	91.3(Zhao et al., 2021)	42.32	39.27 (2.20)	39.44 (2.55)	-	43.0%
RE-Triplet	CoNLL04	78.8(Lou et al., 2023)	23.04	17.84 (3.43)	24.30 (1.29)	11.09 (4.83)	22.6%
	NYT-multi	86.8(Wang et al., 2023b)	3.79	3.48 (0.24)	12.24 (0.59)	2.33 (1.64)	4.0%
	SemEval2010	73.2(Wang et al., 2023a)	7.65	5.82 (1.29)	12.85 (1.14)	-	8.0%
EE-Trigger	ACE05-Evt	77.1(Wang et al., 2023a)	21.93	17.55 (3.86)	27.33 (1.84)	7.81 (2.19)	22.8%
	ACE05+	72.8(Lin et al., 2020)	23.18	18.22 (4.32)	29.17 (2.51)	9.06 (2.79)	25.0%
	CASIE	72.0(Liu et al., 2023)	13.35	7.24 (4.02)	18.23 (5.74)	3.95 (1.12)	10.1%
	Commodity News	94.0(Lee et al., 2021a)	21.58	17.90 (2.37)	37.79 (3.09)	12.75 (2.57)	19.0%
EE-Argument	ACE05-Evt	73.5(Hsu et al., 2022)	27.23	25.09 (1.54)	31.62 (1.68)	-	34.1%
	ACE05+	73.0(Hsu et al., 2022)	27.98	25.80 (1.82)	32.02 (1.77)	-	35.3%
	CASIE	-	19.22	17.31 (2.28)	27.35 (2.75)	-	-
	Commodity News	-	12.97	12.06 (1.45)	15.08 (3.43)	-	-
EE-Joint	ACE05-Evt	57.3(Liu et al., 2023)	11.00	8.74 (1.47)	13.82 (1.18)	-	15.3%
	ACE05+	56.8(Hsu et al., 2022)	11.55	10.12 (0.91)	13.33 (1.08)	-	17.8%
	CASIE	63.5(Wang et al., 2023a)	15.07	14.24 (1.22)	18.96 (1.47)	-	22.4%
	Commodity News	90.0(Lee et al., 2021a)	10.25	8.46 (1.28)	14.02 (3.03)	-	9.4%

# 评估方式

- ChatGPT 倾向生成更长的片段（模仿人类）

Type	Annotated Spans	Predicted Spans
Entity	PGA Europro Tour	2021 PGA Europro Tour
	University of Michigan	The University of Michigan
	Australia	Western Australia
Event Trigger	war	move toward war
	fighting	commit fighting forces
	killed	marines killed
Aspect Term	USB ports	multiple USB ports
	application	application crash
	cable	extender cable
Opinion Term	fast	super fast
	well worth	well worth it
	not handle	does not handle

Task	Dataset	SOTA	Hard	Soft	$\Delta$ F1 (%)
ABSA-AE	$D_{17-14lap}$	85.3	43.03	51.05	+8.02 (18.6%)
ABSA-OE	$D_{17-14lap}$	84.4	48.45	60.07	+11.62 (24.0%)
ABSA-AOE	$D_{19-14lap}$	82.2	57.60	72.13	+14.53 (25.2%)
ABSA-AESC	$D_{20a-14lap}$	70.1	45.48	53.25	+7.77 (17.1%)
ABSA-Pair	$D_{20a-14lap}$	69.1	31.76	39.53	+7.77 (24.5%)
ABSA-Triplet	$D_{20b-14lap}$	61.7	33.17	43.61	+10.44 (31.5%)
NER-Flat	CoNLL03	94.6	60.10	62.12	+2.02 (3.4%)
NER-Nested	ACE05-Ent	87.5	23.38	33.97	+10.59 (45.3%)
RE-Triplet	CoNLL04	78.8	17.84	24.75	+6.91 (38.7%)
EE-Trigger	ACE05-Evt	77.1	17.55	21.20	+3.65 (20.8%)
EE-Argument	ACE05-Evt	73.5	25.09	34.57	+9.48 (37.8%)
EE-Joint	ACE05-Evt	57.3	8.74	11.90	+3.16 (36.2%)

# 稳定性

- ChatGPT 很少生成无关的回复

Task	Dataset	#Sent.	Avg. #Invalid.	Ratio (%)
ABSA-AE	$D_{17-14lap}$	800	38.4	4.8%
ABSA-OE	$D_{17-14lap}$	800	15.8	2.0%
ABSA-ALSC	$D_{17-14lap}$	800	0.0	0.0%
ABSA-AOE	$D_{19-14lap}$	343	21.8	6.4%
ABSA-AESC	$D_{20a-14lap}$	339	0.4	0.1%
ABSA-Pair	$D_{20a-14lap}$	339	6.0	1.8%
ABSA-Triplet	$D_{20b-14lap}$	328	6.8	2.1%
NER-Flat	CoNLL03	3453	396.0	11.5%
NER-Nested	ACE05-Ent	1060	80	7.5%
RE-RC	SemEval2010	2717	13.8	0.5%
RE-Triplet	CoNLL04	288	72.8	25.3%
EE-Trigger	ACE05-Evt	832	28.0	3.4%
EE-Argument	ACE05-Evt	676	18.2	2.7%
EE-Joint	ACE05-Evt	832	38.4	4.6%

# 稳定性

- 对无关上下文非常敏感，性能会大幅下降

Task	Dataset	No-Irrel.	With-Irrel.	$\Delta$ F1 (%)
ABSA-AE	$D_{17-14lap}$	43.03	25.88	17.15 (39.9%)
ABSA-OE	$D_{17-14lap}$	48.45	27.36	21.09 (43.5%)
ABSA-ALSC	$D_{17-14lap}$	74.56	67.89	6.67 ( 8.9%)
ABSA-AOE	$D_{19-14lap}$	57.60	46.24	11.36 (19.7%)
ABSA-AESC	$D_{20a-14lap}$	45.48	29.88	15.60 (34.3%)
ABSA-Pair	$D_{20a-14lap}$	31.76	16.51	15.25 (48.0%)
ABSA-Triplet	$D_{20b-14lap}$	33.17	19.63	13.54 (40.7%)
NER-Flat	CoNLL03	60.10	46.09	14.01 (23.3%)
NER-Nested	ACE05-Ent	23.38	19.65	3.73 (16.0%)
RE-RC	SemEval2010	39.27	37.94	1.33 (3.4%)
RE-Triplet	CoNLL04	17.84	11.40	6.44 (36.1%)
EE-Trigger	ACE05-Evt	17.55	12.01	5.54 (31.6%)
EE-Argument	ACE05-Evt	25.09	20.42	4.67 (18.6%)
EE-Joint	ACE05-Evt	8.74	5.25	3.49 (39.9%)

# 稳定性

- ChatGPT 同样受长尾问题的困扰

Task	Dataset	Head	Tail	Ratio (%)
NER-Flat	FewNERD	34.59	20.82	60.2%
NER-Nested	ACE05-Ent	26.57	7.79	29.3%
RE-RC	NYT-multi	34.71	3.74	10.8%
RE-Triplet	NYT-multi	4.34	0.21	4.8%
EE-Trigger	ACE05-Evt	19.10	14.49	75.9%
EE-Joint	ACE05-Evt	9.48	5.98	63.1%

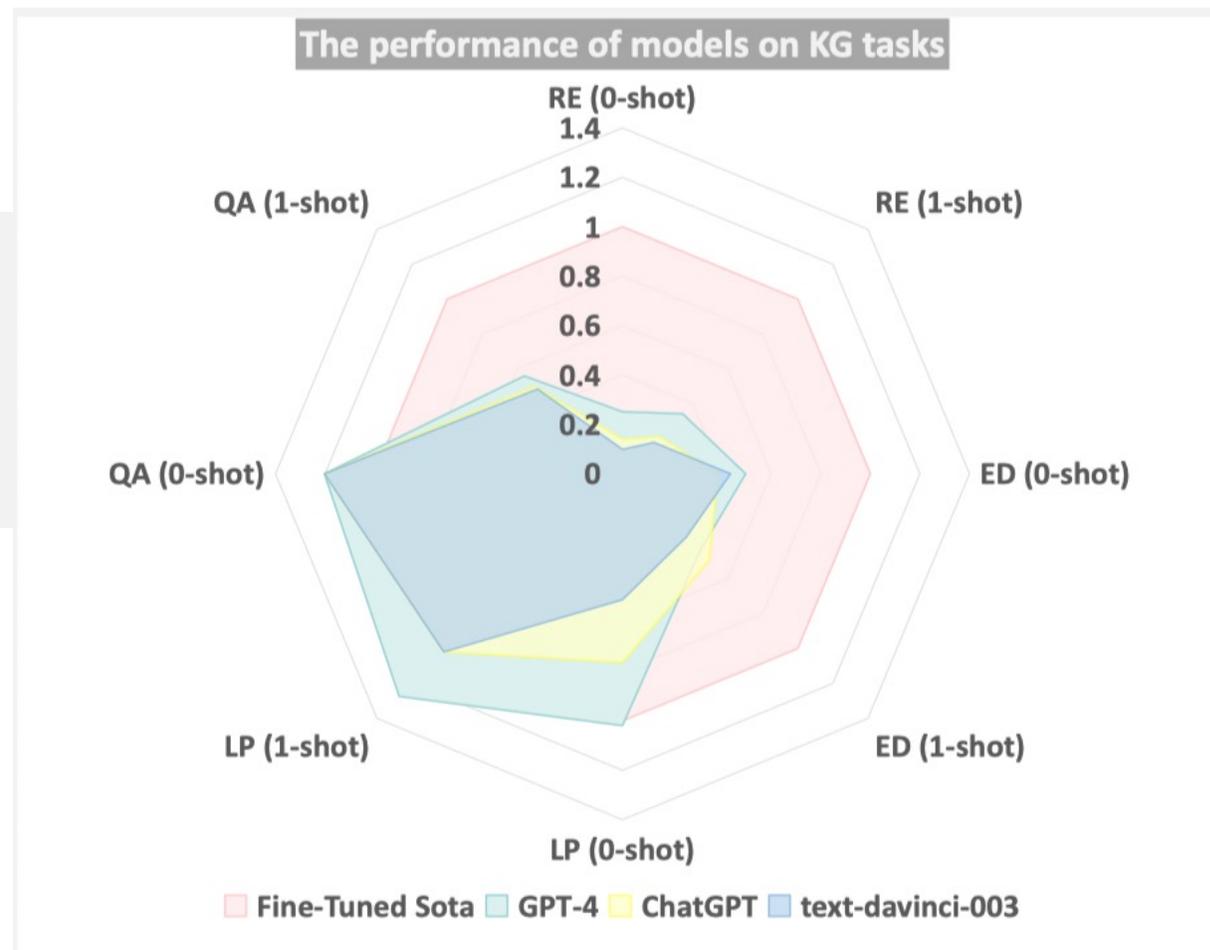
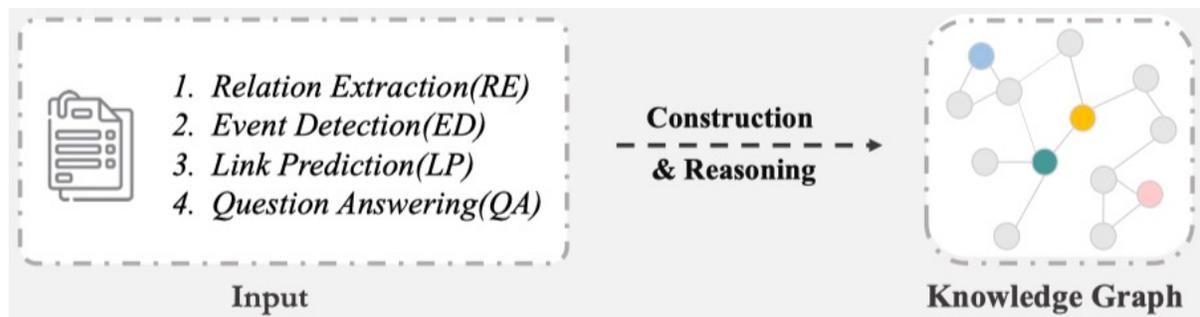
# 错误分析

- 主要的错误是：
  - Unannotated spans
  - Incorrect types
  - Missing spans

Error Type	#Error.	Ratio (%)
Missing spans	2979	15.4%
Unmentioned spans	284	1.5%
Unannotated spans	6361	32.9%
Incorrect span offsets	1744	9.0%
Undefined types	883	4.6%
Incorrect types	4296	22.2%
Other	2801	14.4%
<b>Total</b>	<b>19348</b>	<b>100%</b>

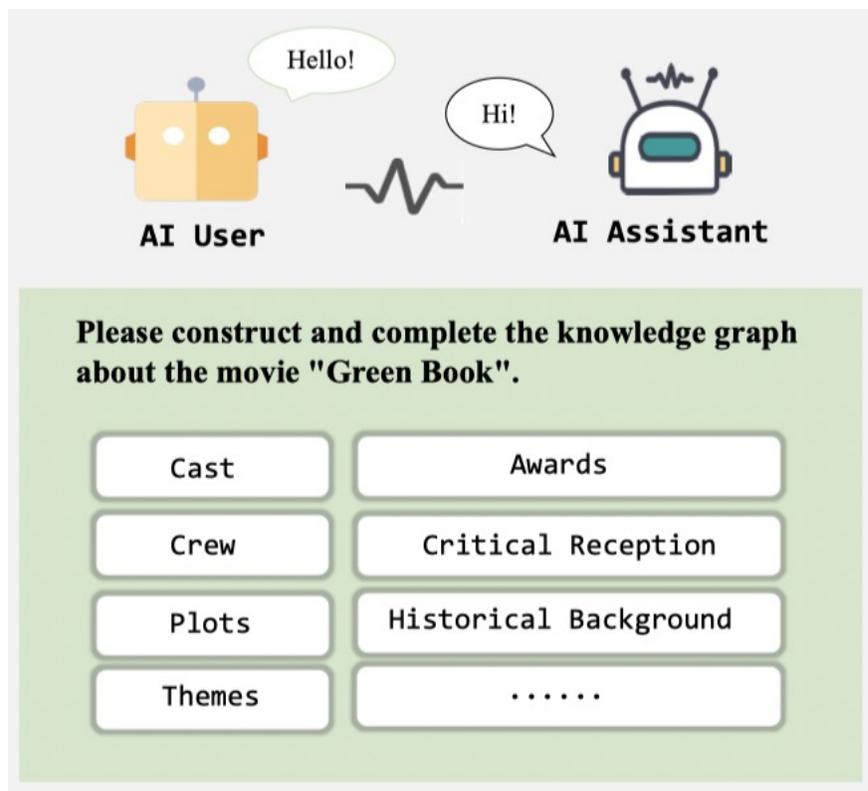
# KG Construction vs. Reasoning

- 抽取不行，推理很行



# 多个 AI 效果更好

- role-playing method in CAMEL (Li et al., 2023).
- AI 助手扮演顾问，AI 用户扮演 KG 领域专家



# 直接问 GPT-4

GPT-4 only



## **Title: Green Book**

- **Release Date:** November 16, 2018
- **Genre:** Biography, Comedy, Drama
- **Director:** Peter Farrelly
- **Screenwriters:** Nick Vallelonga, Brian Hayes Currie, Peter Farrelly
- **Producers:** Jim Burke, Charles B. Wessler, Brian Hayes Currie, Peter Farrelly, Nick Vallelonga
- **Production Companies:** Participant Media, DreamWorks Pictures, Innisfree Pictures, Cinetic Media, Alibaba Pictures
- **Actors/Characters:**
  - Viggo Mortensen as Tony Lip
  - Mahershala Ali as Dr. Don Shirley
  - Linda Cardellini as Dolores
- **Cinematographer:** Sean Porter
- **Music:** Kris Bowers
- **Plot Summary:** The story of Tony Lip, a bouncer from an Italian-American neighborhood in the Bronx, who is hired to drive Dr. Don Shirley, a world-class African-American pianist, on a concert tour from Manhattan to the Deep South. They must rely on "The Green Book" to guide them to the few establishments that were then safe for African-Americans. Confronted with racism, danger as well as unexpected humanity and humor, they are forced to set aside differences to survive and thrive on the journey of a lifetime.
- **Awards:** Won 3 Oscars, including Best Picture, Best Supporting Actor (Mahershala Ali), Best Original Screenplay.



AI User

Knowledge Graph Domain Expert



AI Assistant

Consultant

**Title:** Green Book

**1. Basic Information**

- **Release Year:** 2018
- **Director:** Peter Farrelly
- **Producer:** Jim Burke, Charles B. Wessler, Brian Currie, Peter Farrelly, Nick Vallelonga
- **Writer:** Nick Vallelonga, Brian Currie, Peter Farrelly
- **Genre:** Comedy-drama, Road movie, Biographical film
- **Plot Summary:** Green Book is a 2018 American biographical comedy-drama road film directed by Peter Farrelly. Set in 1962, the film is inspired by the true story of a tour of the Deep South by African American classical and jazz pianist Don Shirley (Mahershala Ali) and Italian American bouncer Frank "Tony Lip" Vallelonga (Viggo Mortensen) who served as Shirley's driver and bodyguard.

**2. Cast**

- Node:** Mahershala Ali
- **Name:** Mahershala Ali
- **Role:** Don Shirley
- **Awards:** Academy Award for Best Supporting Actor, Golden Globe Award for Best Supporting Actor – Motion Picture, Screen Actors Guild Award for Outstanding Performance by a Male Actor in a Supporting Role

- Node:** Viggo Mortensen
- **Name:** Viggo Mortensen
- **Role:** Frank "Tony Lip" Vallelonga
- **Awards:** Academy Award for Best Actor (nominated), Golden Globe Award for Best Actor – Motion Picture Musical or Comedy (nominated), Screen Actors Guild Award for Outstanding Performance by a Male Actor in a Leading Role (nominated)

- Node:** Linda Cardellini
- **Name:** Linda Cardellini
- **Role:** Dolores Vallelonga
- **Awards:** None

- Node:** Dimiter D. Marinov
- **Name:** Dimiter D. Marinov
- **Role:** Oleg
- **Awards:** None

- Node:** Mike Hatton
- **Name:** Mike Hatton
- **Role:** George
- **Awards:** None

**3. Crew**

- Node:** Peter Farrelly
- **Name:** Peter Farrelly
- **Role:** Director, Writer, Producer
- **Awards:** Academy Award for Best Picture, Golden Globe Award for Best Motion Picture – Musical or Comedy, BAFTA Award for Best Original Screenplay

- Node:** Jim Burke
- **Name:** Jim Burke
- **Role:** Producer
- **Awards:** Academy Award for Best Picture, Golden Globe Award for Best Motion Picture – Musical or Comedy

- Node:** Charles B. Wessler
- **Name:** Charles B. Wessler
- **Role:** Producer
- **Awards:** Academy Award for Best Picture, Golden Globe Award for Best Motion Picture – Musical or Comedy

- Node:** Brian Currie
- **Name:** Brian Currie
- **Role:** Writer, Producer
- **Awards:** Academy Award for Best Picture, BAFTA Award for Best Original Screenplay

- Node:** Nick Vallelonga
- **Name:** Nick Vallelonga
- **Role:** Writer, Producer
- **Awards:** Academy Award for Best Picture, Golden Globe Award for Best Screenplay – Motion Picture, BAFTA Award for Best Original Screenplay

**4. Awards**

- Node:** Awards and Recognition for Green Book
- **Academy Awards:** Best Picture, Best Supporting Actor (Mahershala Ali), Best Original Screenplay (Nick Vallelonga, Brian Currie, Peter Farrelly)
- **Golden Globe Awards:** Best Motion Picture – Musical or Comedy, Best Supporting Actor – Motion Picture (Mahershala Ali), Best Screenplay – Motion Picture (Nick Vallelonga, Brian Currie, Peter Farrelly)

.....