ChatGPT Evaluation on Sentence Level Relations: A Focus on Temporal,Causal, and Discourse Relations

对ChatGPT句子层次关系的评估（时间关系、因果关系和篇章关系）

时间关系

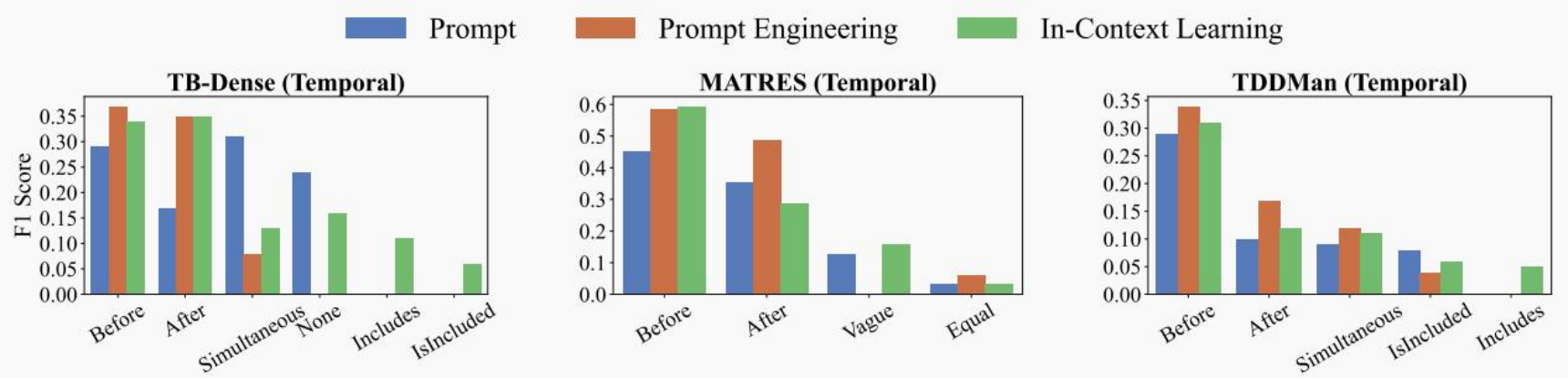
MATRES				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
Prompt	Sentence: "It had a multiplying effect.", "We were pleased that England and New Zealand knew about it, and we thought that's where it would stop." event1: had event2: pleased Question: What is the temporal relation between event1 and event2 in the sentence? A. AFTER B. BEFORE C. EQUAL D. VAGUE Answer:	AFTER	EQUAL	F
Prompt Engineering	Determine the temporal order from "had" to "pleased" in the following sentence: ""It had a multiplying effect.", "We were pleased that England and New Zealand knew about it, and we thought that's where it would stop. "". Only answer one word from AFTER, BEFORE, EQUAL, VAGUE. Answer:	EQUAL	EQUAL	T
In-Context Learning	Determine the temporal order from "give" to "tried" in the following sentence: "It will give the rest of the world the view that Cuba is like any other nation, something the US has, of course, tried to persuade the world that it is not.". Only answer one word from AFTER, BEFORE, EQUAL, VAGUE. Answer: AFTER Determine the temporal order from "invited" to "come" in the following sentence: "Fidel Castro invited John Paul to come for a reason.". Only answer one word from AFTER, BEFORE, EQUAL, VAGUE. Answer: BEFORE Determine the temporal order from "earned" to "rose" in the following sentence: "In the nine months, EDS earned $315.8 million, or $2.62 a share, up 13 % from $280.7 million, or $2.30 a share.". Only answer one word from AFTER, BEFORE, EQUAL, VAGUE. Answer: EQUAL Determine the temporal order from "created" to "become" in the following sentence: "Ms. Atimadi says the war has created a nation of widows. Women have become the sole support of their families.". Only answer one word from AFTER, BEFORE, EQUAL, VAGUE. Answer: VAGUE Determine the temporal order from "had" to "pleased" in the following sentence: ""It had a multiplying effect.", "We were pleased that England and New Zealand knew about it, and we thought that's where it would stop. "". Only answer one word from AFTER, BEFORE, EQUAL, VAGUE. Answer:	BEFORE	EQUAL	F

1、将该任务制定为一个多项选择题问题
2、提醒ChatGPT首先注意时间顺序和这两个事件
3、上下文学习：手动选择输入输出范例

ChatGPT Evaluation on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations

时间关系

Method	TB-Dense	MATRES	TDDMan
Random	15.0	25.8	17.3
BERT-base	62.2	77.2	37.5
Fine-tuned SOTA	68.7	84.0	45.5
$\text{ChatGPT}_{\text{Prompt}}$	23.3	35.0	14.1
$\text{ChatGPT}_{\text{PE}}$	27.0	47.9	16.8
$\text{ChatGPT}_{\text{ICL}}$	25.0	44.9	14.7



1、三个数据集上都落后于SOTA超过30%，ChatGPT不擅长识别两个事件之间的时间关系
2、与Prompt比，PE提高，上下文学习没有提高；多数据集性能非常不稳定
3、上下文学习提高了更难区分的关系的性能，对更容易区分的关系的性能产生了负面影响
4、BEFORE预测结果较好（事件1的序列通常在文本中的事件2之前）
5、在长依赖时间关系提取中，TDDMan上的表现比Random差（TDDMan数据集主要关注长文档）

ChatGPT Evaluation on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations

因果关系

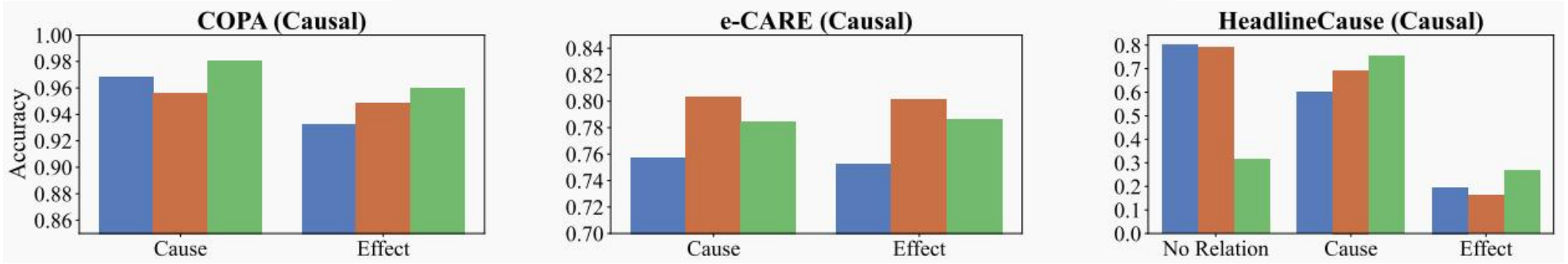
COPA				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
Prompt	The cause of The cashier opened the cash register is: 1. The customer searched his wallet. 2. The customer handed her money. Only answer '1' or '2' only without any other words.	2.	2	T
Prompt Engineering	Given the event The cashier opened the cash register, which choice is more likely to be the cause of this event? 1. The customer searched his wallet. 2. The customer handed her money. Only answer '1' or '2' only without any other words.	2.	2	T
In-Context Learning	Given the event The shirt shrunk, the cause of this event is likely to be I put it in the dryer. Given the event It got dark outside, the effect of this event is likely to be The moon became visible in the sky. Given the event The cashier opened the cash register, which choice is more likely to be the cause of this event? 1. The customer searched his wallet. 2. The customer handed her money. Only answer '1' or '2' only without any other words.	2	2	T

1、简单地向ChatGPT呈现前提及其相应的因果关系选项
2、强调 给定的事件与其选项之间的关系是什么?
3、上下文学习：手动选择输入输出范例

ChatGPT Evaluation on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations

因果关系

Method	COPA	e-CARE	HeadlineCause
Random	50.0	50.0	20.0
Fine-tuned RoBERTa	90.6	70.7	73.5
$\text{ChatGPT}_{\text{Prompt}}$	94.8	74.8	71.4
$\text{ChatGPT}_{\text{PE}}$	95.2	79.6	72.7
$\text{ChatGPT}_{\text{ICL}}$	97.0	78.6	36.2



1、使用PE和上下文学习可以提高ChatGPT的性能，ChatGPT在推理因果关系方面表现出色
2、COPA数据集优于其他两个数据集上的性能；COPA、e-CARE优于微调的RoBERTa
3、上下文学习提高了ChatGPT识别因果关系的能力，也使模型更难区分无关系条目

ChatGPT Evaluation on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations

篇章关系（PDTB-Style篇章关系识别、多源篇章关系识别、对话篇章关系识别、篇章理解应用）

Explicit Discourse Relation Tasks

Strategies	Template input	ChatGPT	Gold	T/F
Top-level Prompt	Argument 1:"When used as background in this way, the music has an appropriate eeriness" Argument 2:"Served up as a solo the music lacks the resonance provided by a context within another medium" Connective between Argument 1 and Argument 2:"however" Question:What is the discourse relation between Argument 1 and Argument 2? A. Comparison B. Contingency C. Expansion D. Temporal Answer:	B. Contingency	A. Comparison	F
Second-level Prompt	Argument 1:"When used as background in this way, the music has an appropriate eeriness" Argument 2:"Served up as a solo the music lacks the resonance provided by a context within another medium" Connective between Argument 1 and Argument 2:"however" Question:What is the discourse relation between Argument 1 and Argument 2? A. Concession B. Contrast C. Cause D. Condition E. Alternative F. Conjunction G. Instantiation H. List I. Restatement J. Asynchronous K. Synchrony Answer:	B. Contrast	B. Contrast	T
Prompt Engineering	Argument 1:"When used as background in this way, the music has an appropriate eeriness" Argument 2:"Served up as a solo the music lacks the resonance provided by a context within another medium" Connective between Argument 1 and Argument 2:"however" Question:What is the discourse relation between Argument 1 and Argument 2? A. Comparison.Concession, nonetheless B. Comparison.Contrast, however C. Contingency.Cause, so D. Contingency.Condition, if E. Expansion.Alternative, instead F. Expansion.Conjunction, also G. Expansion.Instantiation, for example H. Expansion.List, and I. Expansion.Restatement, specifically J. Temporal.Asynchronous, before K. Temporal.Synchrony, when Answer:	B.Comparison. Contrast, however	B. Comparison. Contrast	T

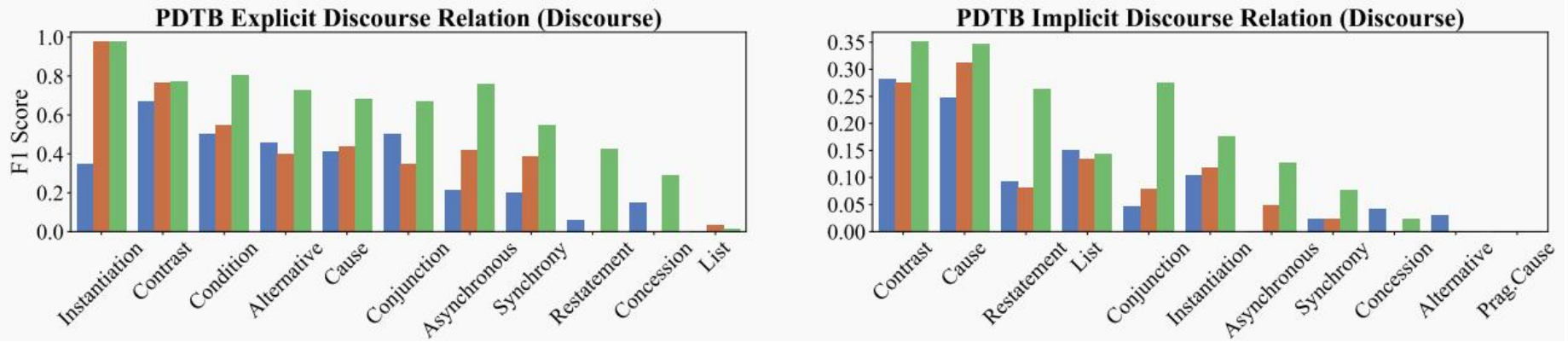
1、将该任务制定为一个多项选择题问题
2、使用PDTB标签依赖性和所选择的连接词引导LLM
3、上下文学习：手动选择输入输出范例

ChatGPT Evaluation on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations

篇章关系（PDTB-Style篇章关系识别、多源篇章关系识别、对话篇章关系识别、篇章理解应用）

Method	Top		Second	
	F1	Acc	F1	Acc
Random	25.12	25.70	7.30	9.19
Zhou et al. (2022a)	93.59	94.78	-	-
Varia et al. (2019)	95.48	96.20	-	-
Chan et al. (2023)	95.64	96.73	-	-
$\text{ChatGPT}_{\text{Prompt}}$	34.94	39.38	31.92	43.26
$\text{ChatGPT}_{\text{PE}}$	69.26	70.21	39.34	50.80
$\text{ChatGPT}_{\text{ICL}}$	84.66	85.97	60.68	63.47

Method	Top		Second	
	F1	Acc	F1	Acc
Random	24.74	25.47	6.48	8.78
Liu et al. (2020)	63.39	69.06	35.25	58.13
Jiang et al. (2022)	65.76	72.52	41.74	61.16
Long and Webber (2022)	69.60	72.18	49.66	61.69
Chan et al. (2023)	70.84	75.65	49.03	64.58
$\text{ChatGPT}_{\text{Prompt}}$	29.85	32.89	9.27	15.59
$\text{ChatGPT}_{\text{PE}}$	33.78	34.94	10.73	20.31
$\text{ChatGPT}_{\text{ICL}}$	36.11	44.18	16.20	24.54



1、ChatGPT可以利用显式连接词的信息识别篇章关系；Contrast,Condition, Instantiation识别表现良好
2、隐式关系仍然是ChatGPT的一项具有挑战性的任务（可能是GPT不能理解篇章关系的抽象意义，不能从文本中把握语言特征）

ChatGPT Evaluation on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations

篇章关系（PDTB-Style篇章关系识别、多源篇章关系识别、对话篇章关系识别、篇章理解应用）

DiscoGeM				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
Prompt	Argument 1:"Allow me to make a few general comments on European solidarity, on the Solidarity Fund and on some events that may provide lessons for the future." Argument 2:"In 2002 I had the experience of leading a country that was struck by terrible floods, together with the Federal Republic of Germany and Austria. It was the scale of that disaster that provided the incentive for the creation of the Solidarity Fund." Question:What is the discourse relation between Argument 1 and Argument 2? (0) arg1-as-denier (1) arg1-as-detail (2) arg1-as-goal (3) arg2-as-denier (4) arg2-as-detail (5) arg2-as-goal (6) arg2-as-instance (7) arg2-as-subst (8) conjunction (9) contrast (10) differentcon (11) disjunction (12) precedence (13) reason (14) result (15) similarity (16) succession (17) synchronous Answer:	(2) arg1-as-goal	(4) arg2-as-detail	F

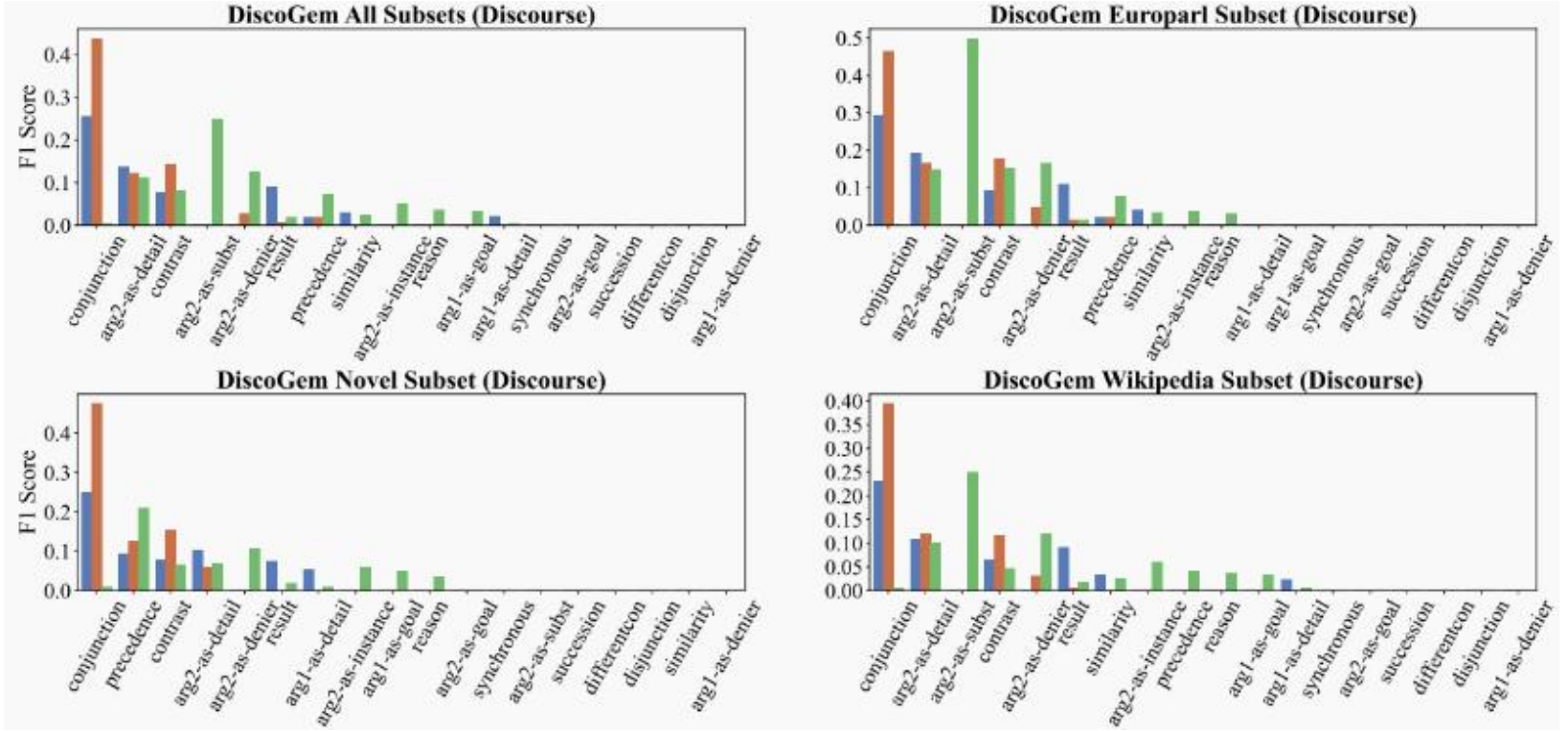
DiscoGeM				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
Prompt Engineering	Argument 1:"However, the Member States are not obliged to replace fixed-term contracts with open-ended contracts assuming that there are other effective measures in place that would prevent or sanction such abuse. The European Court of Justice confirmed this interpretation in its judgment of 4 July 2006 in Case C-212/04 (Adeneler) pertaining to Greek legislation." Argument 2:"The European Court of Justice also stated that interpretation of the relevant national legislation does not fall within its competence. It is entirely for the Greek courts to provide an interpretation of relevant Greek legislation and to determine whether this legislation complies with the requirements of the Directive regarding the existence of effective measures that would prevent and sanction abuse arising from the use of successive fixed-term employment contracts." Question:What is the discourse relation between Argument 1 and Argument 2? (0) arg1-as-denier: despite the fact that (1) argument 1 as detail: in short (2) argument 1 as goal: for that purpose (3) argument 2 as denier: despite this (4) argument 2 as detail: in more detail (5) argument 2 as goal: ensuring that (6) argument 2 as instance: for instance (7) argument 2 as substitution: rather (8) conjunction: in addition (9) contrast: by comparison (10) differentcon: none (11) disjunction: or alternatively (12) precedence: subsequently (13) reason: the reasons is/are that (14) result: consequently (15) similarity: similarly (16) succession: previously (17) synchronous: at that time Answer:	(8) conjunction: in addition	(8) conjunction: in addition	T

1、将该任务制定为一个多项选择题问题 需要直接预测数据集中的原始标签
2、给原始标签添加解释
3、上下文学习：手动选择输入输出范例

ChatGPT Evaluation on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations

篇章关系（PDTB-Style篇章关系识别、多源篇章关系识别、对话篇章关系识别、篇章理解应用）

Method	All		Europarl		Novel		Wiki.	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Random	5.5	3.2	5.5	3.2	5.8	3.1	5.6	3.2
(Liu et al., 2020)	48.7	22.3	53.3	25.9	45.3	23.1	45.6	24.0
ChatGPT$_{Prompt}$	10.8	3.5	13.7	4.2	9.9	3.7	9.4	3.1
ChatGPT$_{PE}$	20.8	4.2	21.6	5.0	25.3	4.8	17.7	3.7
ChatGPT$_{ICL-1}$	3.7	4.5	4.8	6.5	3.1	3.5	3.4	4.2
ChatGPT$_{ICL-3}$	3.3	2.8	3.1	2.4	4.3	4.2	2.9	2.5
ChatGPT$_{ICL-18}$	2.0	2.1	1.2	2.9	3.1	1.7	1.9	2.0



1、ChatGPT的性能显著落后于监督模型（Liu et al.，2020）
2、PE提高ChatGPT的性能（可能是由于引入了标签的介绍，为任务理解提供了额外的信息）
3、ICL随着实例数量的增加，模型的表现比随机更差（能是由于隐式篇章关系可以表达不止一种意义）

ChatGPT Evaluation on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations

篇章关系（PDTB-Style篇章关系识别、多源篇章关系识别、对话篇章关系识别、篇章理解应用）

Dialogue DP-STAC

Strategies	Template input	ChatGPT	Gold
Prompt (w/o desc.)	Here is a multi-party dialogue: Utterance 0: (Speaker A) sorry raef- Utterance 1: (Speaker A) at least i forgot to play it Utterance 2: (Speaker A) before that 6 was rolled Utterance 3: (Speaker B) well at least people should realize your advantage now Q: Predict all the possible discourse relations between utterances and their types line by line (e.g., 'Utterance 0 and utternace 1: (0) Utterance 0 and utterance 3: (1)') Candidate types are listed below: Choose from: (0) Comment (1) Clarification question (2) Question-answer pair (3) Continuation (4) Acknowledgement (5) Question and elaboration (6) Result (7) Elaboration (8) Explanation (9) Correction (10) Contrast (11) Conditional (12) Background (13) Narration (14) Alternation (15) Parallel	Utterance 0 and utterance 1: (2) Utterance 1 and utterance 2: (0) Utterance 2 and utterance 3: (9) Utterance 3 and utterance 4: (0) Utterance 4 and utterance 5: (5) Utterance 5 and utterance 6: (0) Utterance 6 and utterance 7: (7) Utterance 7 and utterance 8: (0) Utterance 8 and utterance 9: (3) Utterance 9 and utterance 10: (14)	Utterance 0 and utterance 1: (8) Utterance 1 and utterance 2: (13) Utterance 1 and utterance 3: (0)

1、是否为标签提供描述（w/desc）-PE
2、没有（w/o desc）-prompt
3、上下文学习：手动选择输入输出范例--是否为标签提供描述

Dialogue DP-STAC

Strategies	Template input	ChatGPT	Gold
Prompt (w/ desc.)	Here is a multi-party dialogue: Utterance 0: (Speaker A) sorry raef- Utterance 1: (Speaker A) at least i forgot to play it Utterance 2: (Speaker A) before that 6 was rolled Utterance 3: (Speaker B) well at least people should realize your advantage now Q: Predict all the possible discourse relations between utterances and their types line by line (e.g., 'Utterance 0 and utternace 1: (0) Utterance 0 and utterance 3: (1)') Candidate types are listed below: Choose from: (0) Comment: Utterance y comments utterance x. (1) Clarification question: Utterance y clarifies utterance x. (2) Question-answer pair: Utterance x is a question and utterance y is the answer of utterance x. (3) Continuation: Utterance y is the continuation of utterance x. (4) Acknowledgement: Utterance y acknowledges utterance x. (5) Question and elaboration: Utterance x is a question and utterance y tries to elaborate utterance x. (6) Result: Utterance y is the effect brought about by the situation described in utterance x. (7) Elaboration: Utterance y elaborates utterance x. (8) Explanation: Utterance y is the explanation of utterance x. (9) Correction: Utterance y corrects utterance x. (10) Contrast: Utterance x and utterance y share a predicate or property and a difference on shared property. (11) Conditional: Utterance x is the condition of utterance y or utterance y is the condition of utterance x. (12) Background: Utterance y is the background of utterance x. (13) Narration: Utterance y is the narration of utterance x. (14) Alternation: Utterance x and utterance y denote alternative situations. (15) Parallel: Utterance y and utterance x are parallel and present almost the same meaning.	Utterance 0 and utterance 1: (2) Utterance 0 and utterance 3: (1) Utterance 1 and utterance 5: (0) Utterance 2 and utterance 3: (4) Utterance 4 and utterance 5: (0) Utterance 6 and utterance 7: (4) Utterance 8 and utterance 9: (0) Utterance 9 and utterance 10: (9)	Utterance 0 and utterance 1: (8) Utterance 1 and utterance 2: (13) Utterance 1 and utterance 3: (0)

ChatGPT Evaluation on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations

篇章关系（PDTB-Style篇章关系识别、多源篇章关系识别、对话篇章关系识别、篇章理解应用）

Method	STAC		Molweni	
	Link	Link&Rel	Link	Link&Rel
Afantenos et al. (2015)	68.8	50.4	-	-
Perret et al. (2016)	68.6	52.1	-	-
Shi and Huang (2019)	73.2	55.7	78.1	54.8
ChatGPT$_{\text{zero w/ desc.}}$	20.5	4.3	26.7	5.0
ChatGPT$_{\text{zero w/o desc.}}$	20.0	4.4	28.3	5.4
ChatGPT$_{\text{few (n=1) w/ desc.}}$	21.0	7.1	25.7	6.0
ChatGPT$_{\text{few (n=3) w/ desc.}}$	20.7	7.3	25.1	5.7
ChatGPT$_{\text{few (n=1) w/o desc.}}$	21.2	6.2	27.2	6.8
ChatGPT$_{\text{few (n=3) w/o desc.}}$	21.3	7.4	26.5	6.9

Method	STAC		Molweni	
	Acc	F1	Acc	F1
Random	6.2	4.8	6.3	4.1
ChatGPT$_{\text{Prompt}}$	22.8	8.7	16.5	6.9
ChatGPT$_{\text{PE}}$	25.9	8.6	23.0	7.6
ChatGPT$_{\text{ICL}}$	24.1	13.9	14.7	8.1

1、ChatGPT的表现明显低于监督基线（ChatGPT对对话篇章结构理解不足）
2、添加额外的例子可以提高ChatGPT在关系预测的性能，对链接无用
3、测试“关系分类”下的结果，也没有取得很高的性能

ChatGPT Evaluation on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations

篇章关系（PDTB-Style篇章关系识别、多源篇章关系识别、对话篇章关系识别、篇章理解应用）

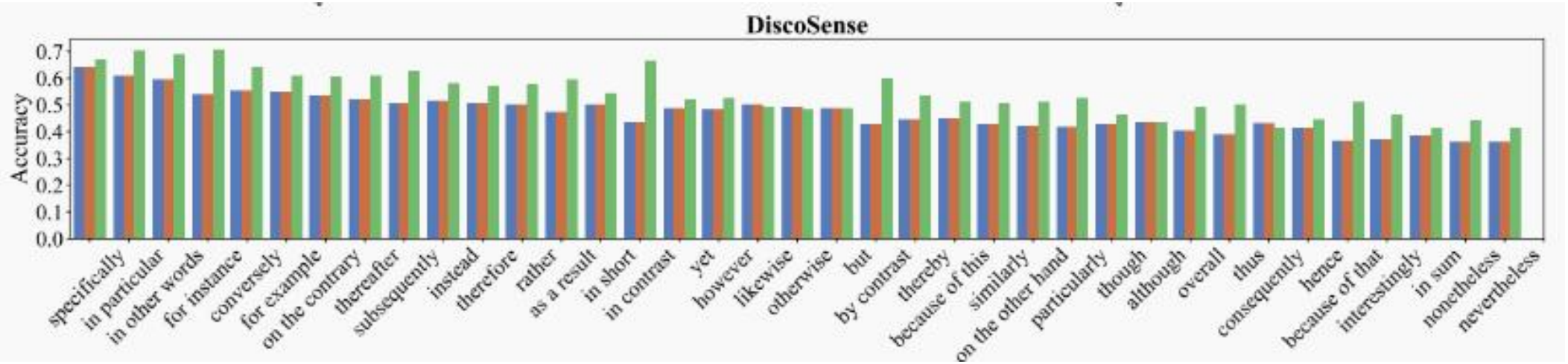
CKBP				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
Prompt Engineering	Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX drinks coffee, as a result, PersonX feels, refreshed.	Yes	Yes	T
In-Context Learning	Answer whether the following statement is plausible. Answer with only Yes or No:If PersonY accept the interview, as a result, PersonY or others will, PersonX give PersonY this opportunity.A: No Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX lead the line, as a result, PersonY or others feel, PersonX support PersonX family.A: No Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX form PersonY conception, as a result, PersonY or others want to, PersonY want to discuss with PersonZ.A: Yes Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX give, PersonX is seen as, PersonX be comunicative.A: Yes Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX be nervous, as a result, PersonX will, that be important to PeopleX.A: Yes Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX celebrate persony, because PersonX wanted, PersonX feel oneself.A: No Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX learn to ride a bike, but before, PersonX needed, PersonX wear helmet.A: Yes Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX take PersonY time, as a result, PersonX feels, PersonX feel mortified.A: No Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX take PersonY time, as a result, PersonX feels, PersonX feel mortified.A: No Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX want to ask a tought question, as a result, PersonX wants to, PersonX want to throw out PersonX clothes.A: No Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX achieve PersonX end, happens after, PersonX start a small business.A: Yes Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX like the idea, happens before, PersonX call a uber.A: No Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX get injure, because, PersonX feel odd.A: No Answer whether the following statement is plausible. Answer with only Yes or No:If person x be bed ridden with illness, can be hindered by, PersonX find the perfect dog.A: No Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX play violin, includes the event or action, PersonX make noise.A: Yes Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX could not complete something, causes, PeopleX have find it.A: No Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX drinks coffee, as a result, PersonX feels, refreshed.	Yes	Yes	T

1、PE：回答下面的说法是否可信。只回答是或否：如果PersonX喝咖啡，因此，PersonX会感觉神清气爽。
2、上下文学习：手动选择输入输出范例

ChatGPT Evaluation on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations

篇章关系（PDTB-Style篇章关系识别、多源篇章关系识别、对话篇章关系识别、篇章理解应用）

Method	CKBP v2. AUC	CKBP v2. F1	DISCOSENSE Acc
Fine-tuned SOTA	73.70	46.70	65.87
$\text{ChatGPT}_{\text{PE}}$	65.77	45.93	47.25
$\text{ChatGPT}_{\text{ICL}}$	66.20	46.42	54.67



1、缺乏微妙的推理能力来区分不同的话语关系
2、ChatGPT在CKBP v2中获得了类似的F1分数，但在AUC方面仍然表现不佳
3、Discosense数据集，ChatGPT还有很长的路要走(人类的性能（95.40）)

ChatGPT Evaluation on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations

ChatGPT Evaluation on Sentence Level Relations: A Focus on Temporal,Causal, and Discourse Relations

结论：

ChatGPT擅长检测和推理因果关系，但却难以确定事件的时间顺序。

虽然它可以通过连接词来识别大多数语篇关系，但隐式关系仍然具有挑战性。

ChatGPT在对话篇章解析任务中也表现出较弱的性能。

对ChatGPT句子层次关系的评估（时间关系、因果关系和篇章关系）

Uncovering the Potential of ChatGPT for Discourse Analysis in Dialogue: An Empirical Study

对每个任务采用了两种prompt(判别prompt和生成prompt)。
判别prompt主要将任务视为一个多选择问题指导LLM从候选选项中选择正确答案
生成prompt主要将任务视为生成问题，指导LLM以合适的格式生成正确的答案

ChatGPT在对话语篇分析中的潜力（话题分割 篇章关系识别 对话篇章分析）

主题分割

Type	Prompts for Dialogue Topic Segmentation
Discriminative	The following is a dialogue. Give each utterance a binary label, where 1 indicates that the utterance starts a new topic. please output the result of the sequence annotation as a python list. $0 : U_1$ $1 : U_2$ $\cdots$ $n : U_n$
Generative	Please identify several topic boundaries for the following dialogue and each topic consists of several consecutive utterances. please output in the form of {'topic i':[], ... ,'topic j':[]}, where the elements in the list are the index of the consecutive utterances within the topic, and output even if there is only one topic. $0 : U_1$ $1 : U_2$ $\cdots$ $n : U_n$

判别prompt
将ChatGPT作为一个序列标注模型，指示ChatGPT用0/1标记每个话语，1表示该话语开始一个新的主题

生成prompt
告诉ChatGPT抽象描述，以帮助它理解主题分割的本质，指示ChatGPT结构化输出主题索引和相应的连续话语

区别：生成式直接揭示话题切分的本质，而判别式通过抽象的0/1标签间接表示主题关系识别

ChatGPT Evaluation on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations

篇章关系识别

判别prompt
将多分类任务视为一个多选择问题，并指示ChatGPT从候选选项和提示中选择正确的话语关系

生成prompt
将关系识别视为一项生成任务，并根据SDRT理论指导ChatGPT生成篇章关系

区别：生成式直接指导ChatGPT生成篇章关系，而判别式则指导ChatGPT从候选选项中选择正确的篇章关系

Type	Prompts for Dialogue Discourse Relation Recognition
Discriminative	Please select the rhetorical relation of the utterance pair U_i and U_j from the following candidate rhetorical relations Relation 0: Elaboration. Relation 1: Comment. Relation 2: Clarification_question. Relation 3: Acknowledgement. Relation 4: Explanation. Relation 5: Conditional. Relation 6: Question-answer pairs. Relation 7: Alternation. Relation 8: Question-Elaboration. Relation 9: Result. Relation 10: Background. Relation 11: Narration. Relation 12: Correction. Relation 13: Parallel. Relation 14: Contrast. Relation 15: Continuation. Please output the rhetorical relations and do not output anything except the rhetorical relation.

Generative	There are sixteen rhetorical relation according to the segmented discourse representation theory listed as:
	{
	"Comment":"",
	"Clarification_question":"",
	"Elaboration":"",
	"Acknowledgement":"",
	"Explanation":"",
	"Conditional":"",
	"Question-answer pairs":"",
	"Alternation":"",
	"Question-Elaboration":"",
	"Result":"",
	"Background":"",
	"Narration":"",
	"Correction":"",
	"Parallel":"",
	"Contrast":"",
	"Continuation":""
	}
	Please generate the rhetorical relation of the following utterance pair according to the segmented discourse representation theory.
	Utterance pair:
	U_i
	U_j

ChatGPT Evaluation on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations

对话篇章分析

判别prompt
告诉ChatGPT修辞关系的类型，并指示它为包含n个话语的对话注释n-1条边，其中每个边连接两个话语，指示ChatGPT以稀疏矩阵的形式输出

生成prompt
直接告诉ChatGPT对话的修辞结构可以用有向无环图表示，指示ChatGPT对对话的修辞结构进行注释，并将其以稀疏矩阵的形式表示

区别：生成式直接指导ChatGPT生成对话的整个修辞结构，而判别式则指导ChatGPT对每个环节及其对应关系进行局部判断

Type	Prompts for Dialogue Discourse Parsing

Discriminative	There are sixteen rhetorical relations and each relation can connect any two utterances according to the Segmented Discourse Rhetorical Theory. The relations are as follows:
	{
	'Comment':'',
	"Clarification_question":"",
	"Elaboration":"",
	"Acknowledgement":"",
	"Explanation":"",
	"Conditional":"",
	"Question-answer pairs":"",
	"Alternation":"",
	"Question-Elaboration":"",
	"Result":"",
	"Background":"",
	"Narration":"",
	"Correction":"",
	"Parallel":"",
	"Contrast":"",
	"Continuation":""
	}
	Giving you a dialogue Consisting N utterances, you need to annotate N-1 rhetorical relations and each of them connects two different utterances. Please annotate the following dialogue and output as follows "[i , j, relationtype] \n [i, j, realtiontype]", where i, j are the index of utterances.

Generative	According to the Segmented Discourse Rhetorical Theory, the rhetorical structure of a dialogue can be represented by a directed acyclic graph, where nodes are utterances and edges are the following 16 relations: {
	'Comment':'',
	"Clarification_question":"",
	"Elaboration":"",
	"Acknowledgement":"",
	"Explanation":"",
	"Conditional":"",
	"Question-answer pairs":"",
	"Alternation":"",
	"Question-Elaboration":"",
	"Result":"",
	"Background":"",
	"Narration":"",
	"Correction":"",
	"Parallel":"",
	"Contrast":"",
	"Continuation":""
	}
	please annotate the rhetorical structure of the following dialogue and represent it in the form of [index1, index2, 'relation'], where index1 and index2 are the index of two utterances, and the 'relation' is one of the above relations to connect the two utterances.

ChatGPT Evaluation on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations

主题分割

Domain		Chitchat DialSeg_711		Finance ZYS		Chitchat TIAGE		Chitchat CNTD	
Methods		$P_k(\downarrow)$	$F_1(\uparrow)$	$P_k(\downarrow)$	$F_1(\uparrow)$	$P_k(\downarrow)$	$F_1(\uparrow)$	$P_k(\downarrow)$	$F_1(\uparrow)$
unsupervised	TextTiling	40.44	60.80	45.86	48.50	47.27	45.57	51.36	46.84
	GreedySeg	50.95	40.10	44.12	50.20	52.63	49.47	53.81	53.36
	TeT+CLS	40.49	61.00	43.01	50.20	40.49	61.00	43.01	50.20
	UPCS	26.80	77.60	**40.99**	**52.10**	47.19	58.63	46.11	58.18
supervised	BERT	-	-	-	-	-	66.60	-	80.80
	T5	-	-	-	-	-	73.90	-	81.10
	MGP	-	-	-	-	-	**76.20**	-	**84.70**
ChatGPT	DP	$49.68_{\pm 0.11}$	$51.95_{\pm 0.11}$	$42.55_{\pm 0.33}$	$40.67_{\pm 0.06}$	$50.33_{\pm 0.25}$	$53.27_{\pm 0.69}$	$48.93_{\pm 0.84}$	$59.05_{\pm 0.35}$
	GP	$\mathbf{10.56}_{\pm 0.18}$	$\mathbf{89.42}_{\pm 0.08}$	56.19 ± 0.29	$49.60_{\pm 0.04}$	$\mathbf{42.35}_{\pm 2.31}$	$61.31_{\pm 1.87}$	$\mathbf{27.08}_{\pm 1.00}$	$77.36_{\pm 0.35}$

1、生成比判别能达到更高的性能，说明ChatGPT直接表示比间接使用0/1标签更有效
2、生成prompt的ChatGPT在Dialog_711、TIAGE和CNTD上可以显著超越所有无监督方法，在ZYS上也可以达到相当的性能，ChatGPT对一般领域主题有很好的理解，但缺乏特定领域的知识来识别特定领域的主题
3、在TIAGE和CNTD数据集的F1指标上，它可以分别达到SOTA监督基线MGP的80%(61.31/76.20)和90%(77.36/84.70)的性能，这表明ChatPGT对相对简单的主题结构具有较强的理解能力

ChatGPT Evaluation on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations

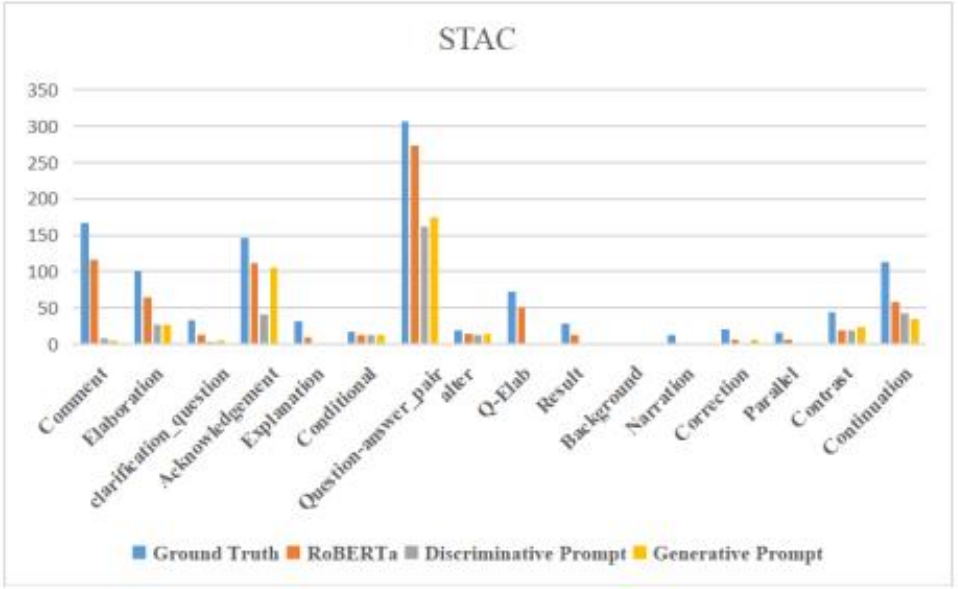
篇章关系识别

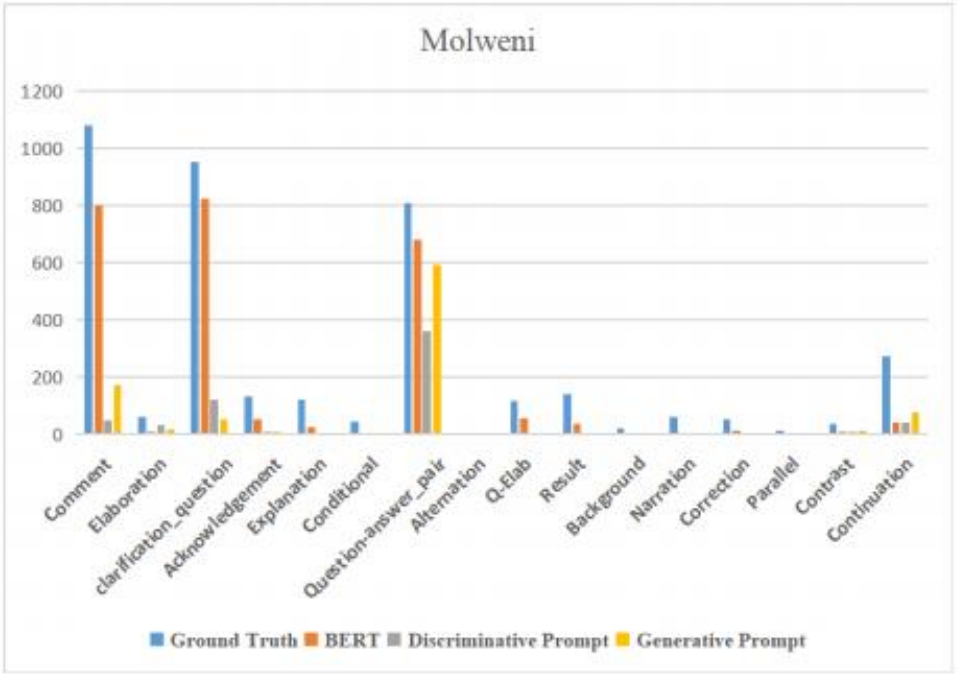
Methods		STAC		Molweni	
		Micro-F1	Macro-F1	Micro-F1	Macro-F1
Supervised	XLNet	63.92	46.28	65.52	25.90
	BERT	64.80	46.22	65.16	31.07
	RoBERTa	68.09	51.97	65.36	26.78
ChatGPT	DP	$28.43_{\pm 1.09}$	$17.31_{\pm 0.45}$	$16.38_{\pm 0.16}$	$7.47_{\pm 0.02}$
	GP	$36.12_{\pm 0.06}$	$19.95_{\pm 0.65}$	$24.70_{\pm 0.72}$	$9.11_{\pm 0.11}$

1、生成式也比判别式有更高的性能，这表明ChatGPT直接生成篇章关系比间接使用多选项更有效
2、ChatGPT在STAC数据集上的性能分别为RoBERTa基线的53%(36.12/68.09)和38%(19.95/51.97)，在Molweni数据集上的性能分别为BERT基线的38%(24.70/65.16)和29%(9.11/31.07)
3、在关系识别上的表现远不如对话主题分割，表明ChatGPT对更复杂的篇章关系的理解能力下降

篇章关系识别

1、ChatGPT在一些有特征的关系类型上，如Acknowledgement、Continuation、Questionanswer_pair等，与监督基线相比，可以取得具有竞争力的性能
2、ChatGPT很难识别评论关系。这可能与ChatGPT不允许主观感受或评论，不能识别涉及主观感受的评论关系有关
3、ChatGPT倾向于将具有Clarification_question关系的话语对识别为问答对关系，表明ChatGPT在识别语篇关系时忽略了问题的顺序
4、ChatGPT对无法确定篇章关系的话语对给出合理的解释，人类很难确定这些话语对的关系，ChatGPT可以帮助人类纠正这些标注了不合理关系的话语对





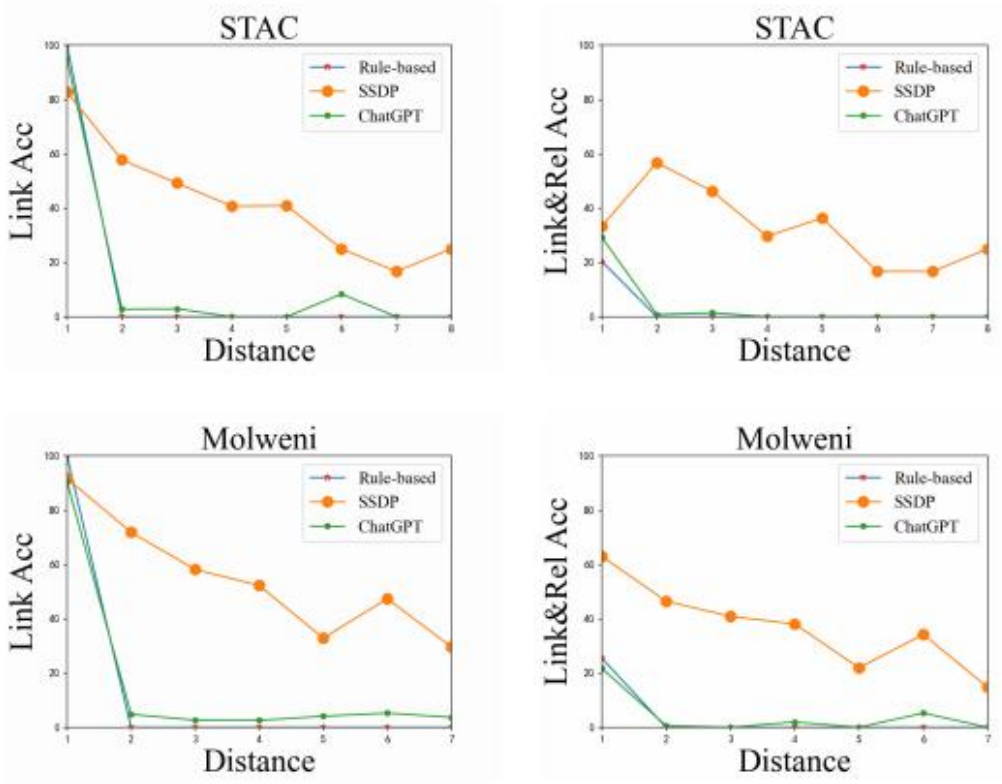
ChatGPT Evaluation on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations

对话篇章分析

Methods		STAC		Molweni	
		Link	Link&Rel	Link	Link&Rel
	rule-based	60.57	20.11	67.56	25.60
Supervised	DSM	71.99	53.62	76.94	53.49
	SSAM	73.48	57.31	81.63	58.54
	SSP	73.00	57.40	83.70	59.40
	DAMT	73.64	57.42	82.50	58.91
	SDDP	74.40	59.60	83.50	59.90
ChatGPT	DP	$60.67_{\pm 0.13}$	$21.82_{\pm 0.71}$	$65.22_{\pm 0.28}$	$20.69_{\pm 0.01}$
	GP	$59.91_{\pm 0.13}$	$25.25_{\pm 0.88}$	$63.75_{\pm 0.04}$	$23.85_{\pm 0.06}$

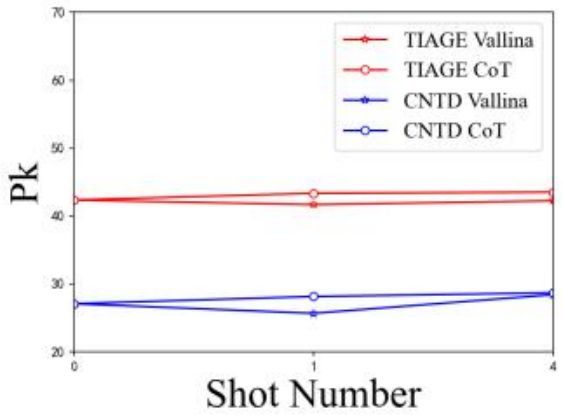
1、生成比判别更好，ChatGPT对对话篇章结构的理解是一个很大的挑战
2、ChatGPT在STAC和Molweni上的Link&Rel指标分别达到了SDDP的42% (25.25 vs 59.60)和40% (23.85 vs 59.90)的性能，ChatGPT和监督SOTA基线之间的差距表明，仍然有很大的改进空间

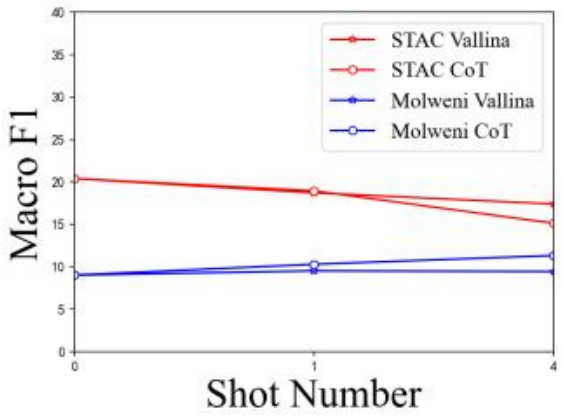
对话篇章分析

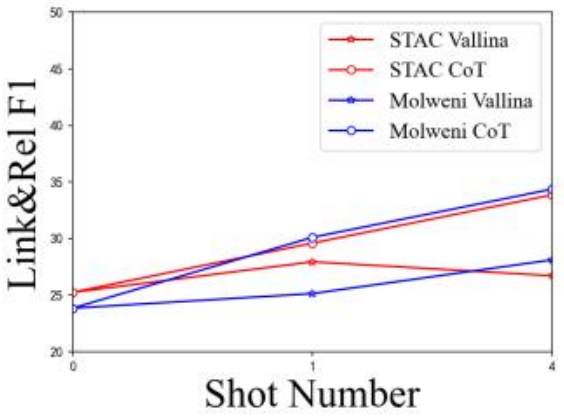


ChatGPT和几个基线在不同距离下的性能比较，ChatGPT在距离1上表现良好，并且与Rule-based方法有相似的趋势，表明ChatGPT只能以线性形式表示对话的话语结构，很难理解对话的篇章结构

语境学习(ICL)







分析语境学习(ICL)对三个任务的影响
1、Vallina方法，它直接将示例提供给ChatGPT
2、思想链(CoT)，它将范例和人类推理过程提供给ChatGPT

结果：
无论是Vallina方法还是CoT方法在主题分割和关系识别方面的性能都不随shot number的增加而线性增加。
但在对话解析方面有了显著的改进，尤其是CoT方法
表明对话语篇解析是一项复杂的任务，CoT可以帮助ChatGPT更好地理解对话篇章结构

Uncovering the Potential of ChatGPT for Discourse Analysis in Dialogue: An Empirical Study

结论：

ChatGPT在主题分割方面表现突出，有时甚至优于人工标注。
还可以作为识别篇章关系任务中的可靠助手。
但对于最具挑战性的篇章分析任务，表现有限。

ChatGPT在对话语篇分析中的潜力（话题分割 篇章关系识别 对话篇章分析）