论文汇报

1. Is ChatGPT a Good Causal Reasoner? A Comprehensive Evaluation.

2. Using ChatGPT for Entity Matching

汇报人：陈鑫宇

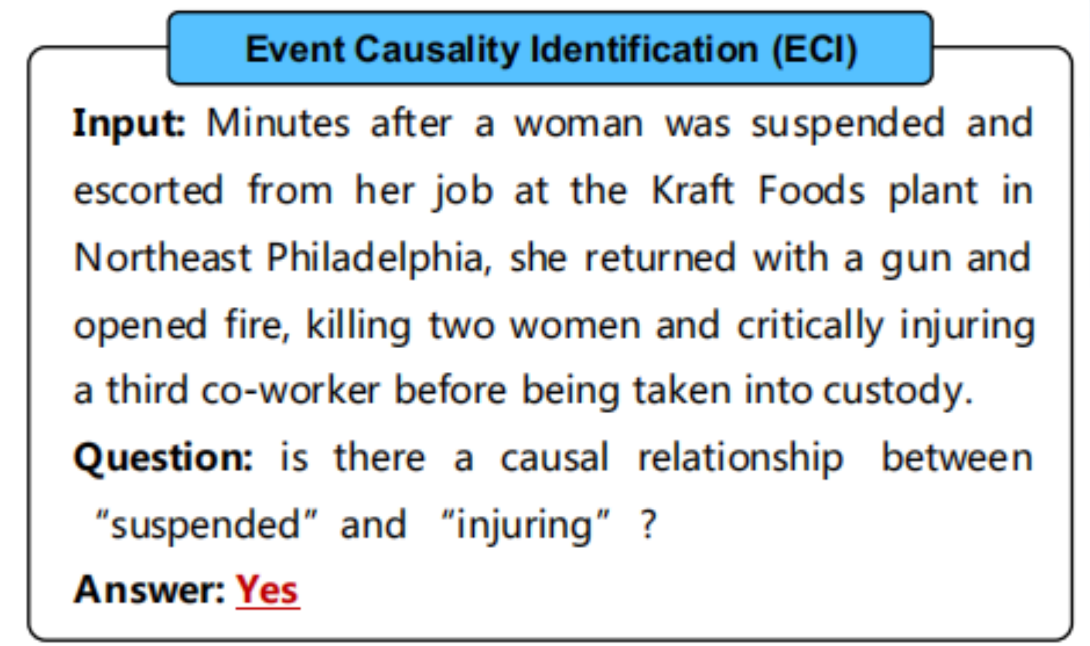
时间：2023年5月17日

Is ChatGPT a Good Causal Reasoner? A Comprehensive Evaluation.

1. 因果解释能力大于因果推理能力

2. 因果推理幻觉

3. 封闭式模板优于开放式模板，显示因果优于隐式因果，短距离性能更佳

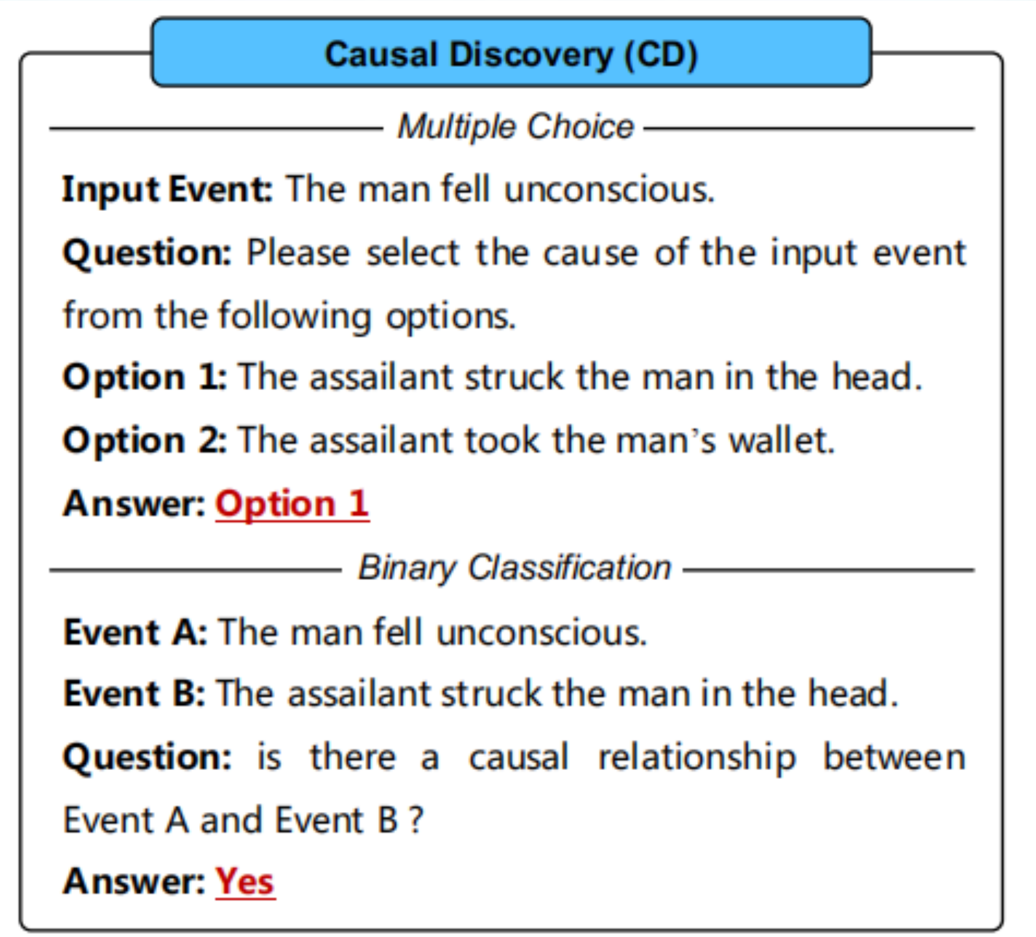


Event Causality Identification (ECI)

Input: Minutes after a woman was suspended and escorted from her job at the Kraft Foods plant in Northeast Philadelphia, she returned with a gun and opened fire, killing two women and critically injuring a third co-worker before being taken into custody.

Question: is there a causal relationship between “suspended” and “injuring” ?

Answer: Yes



Causal Discovery (CD)

Multiple Choice

Input Event: The man fell unconscious.

Question: Please select the cause of the input event from the following options.

Option 1: The assailant struck the man in the head.

Option 2: The assailant took the man’s wallet.

Answer: **Option 1**

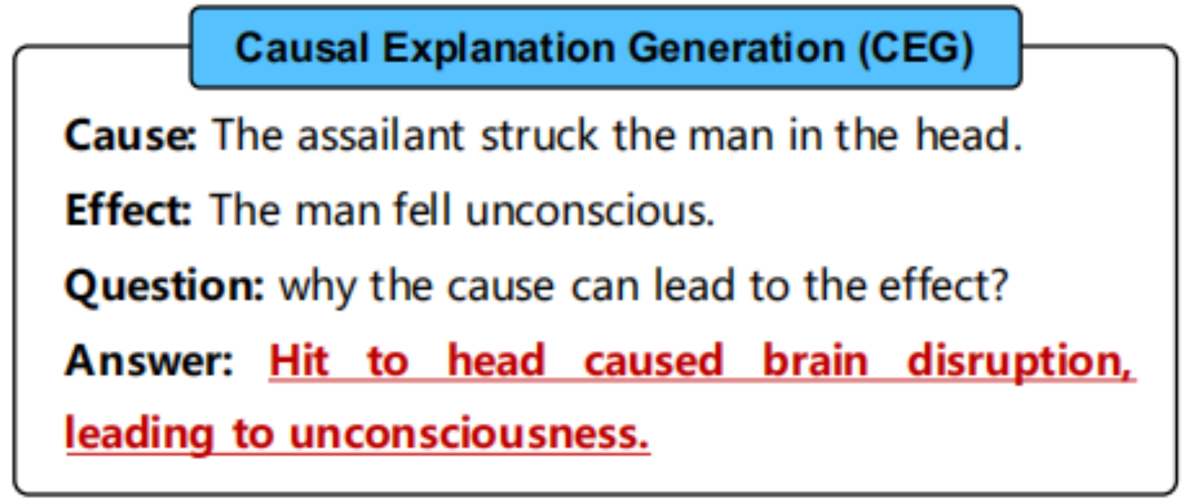
Binary Classification

Event A: The man fell unconscious.

Event B: The assailant struck the man in the head.

Question: is there a causal relationship between Event A and Event B ?

Answer: **Yes**



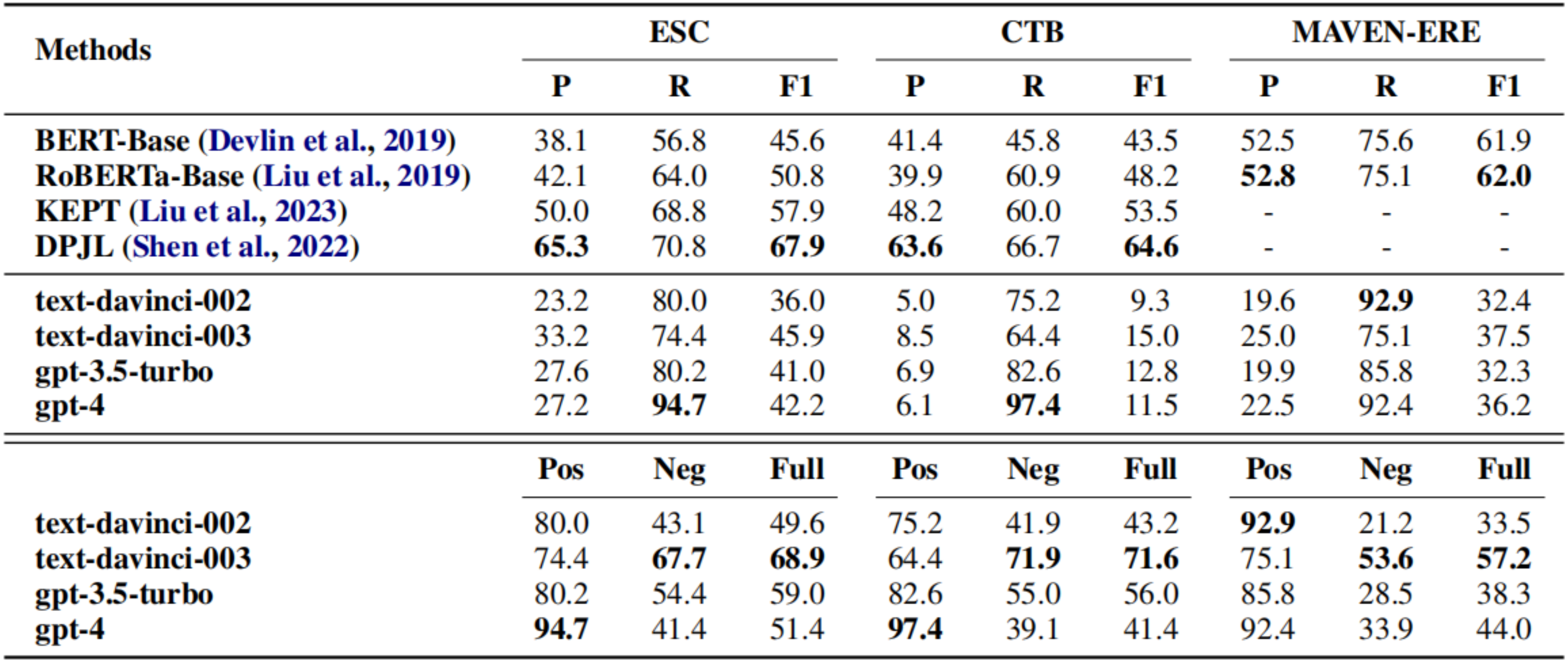
Causal Explanation Generation (CEG)

Cause: The assailant struck the man in the head.

Effect: The man fell unconscious.

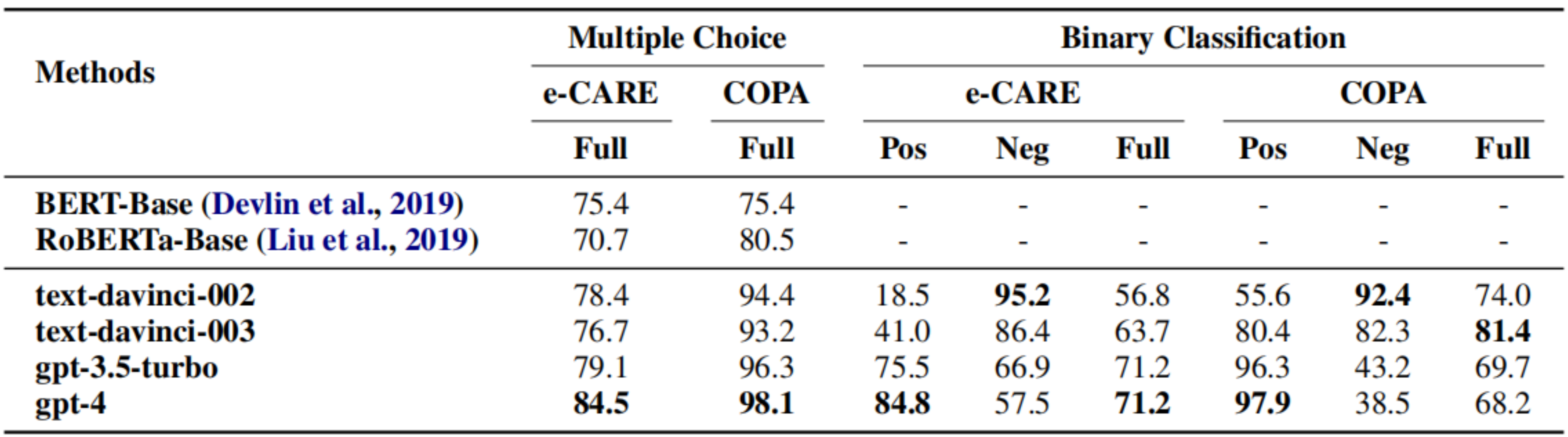
Question: why the cause can lead to the effect?

Answer: **Hit to head caused brain disruption, leading to unconsciousness.**

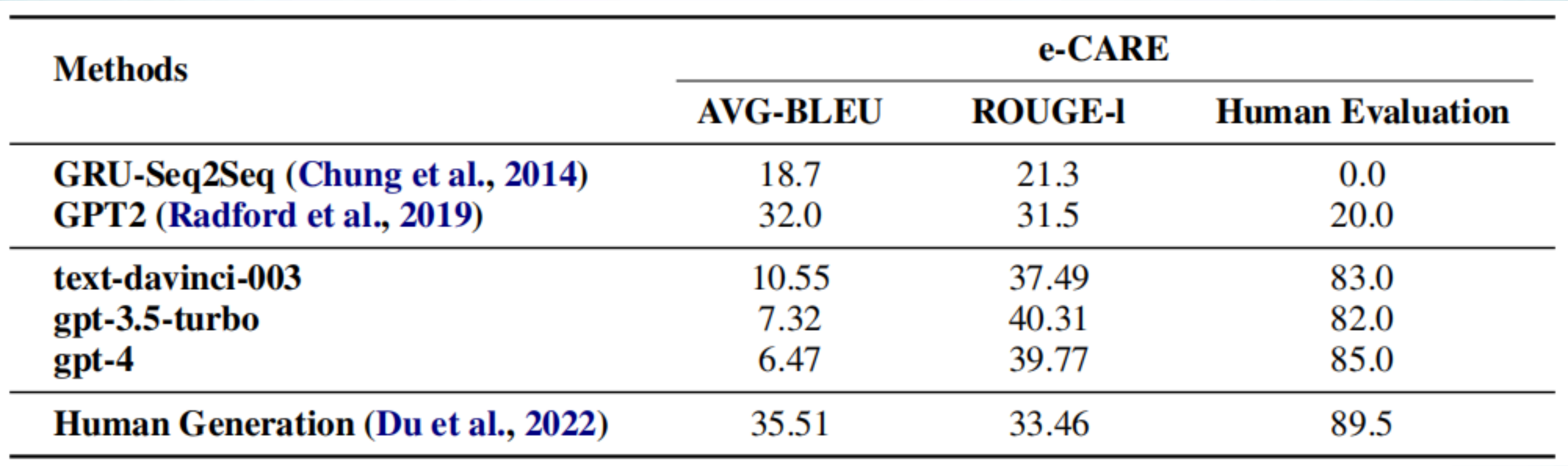


Methods	ESC			CTB			MAVEN-ERE		
	P	R	F1	P	R	F1	P	R	F1
BERT-Base (Devlin et al., 2019)	38.1	56.8	45.6	41.4	45.8	43.5	52.5	75.6	61.9
RoBERTa-Base (Liu et al., 2019)	42.1	64.0	50.8	39.9	60.9	48.2	**52.8**	75.1	**62.0**
KEPT (Liu et al., 2023)	50.0	68.8	57.9	48.2	60.0	53.5	-	-	-
DPJL (Shen et al., 2022)	**65.3**	70.8	**67.9**	**63.6**	66.7	**64.6**	-	-	-
text-davinci-002	23.2	80.0	36.0	5.0	75.2	9.3	19.6	**92.9**	32.4
text-davinci-003	33.2	74.4	45.9	8.5	64.4	15.0	25.0	75.1	37.5
gpt-3.5-turbo	27.6	80.2	41.0	6.9	82.6	12.8	19.9	85.8	32.3
gpt-4	27.2	**94.7**	42.2	6.1	**97.4**	11.5	22.5	92.4	36.2

	Pos	Neg	Full	Pos	Neg	Full	Pos	Neg	Full
text-davinci-002	80.0	43.1	49.6	75.2	41.9	43.2	**92.9**	21.2	33.5
text-davinci-003	74.4	**67.7**	**68.9**	64.4	**71.9**	**71.6**	75.1	**53.6**	**57.2**
gpt-3.5-turbo	80.2	54.4	59.0	82.6	55.0	56.0	85.8	28.5	38.3
gpt-4	**94.7**	41.4	51.4	**97.4**	39.1	41.4	92.4	33.9	44.0

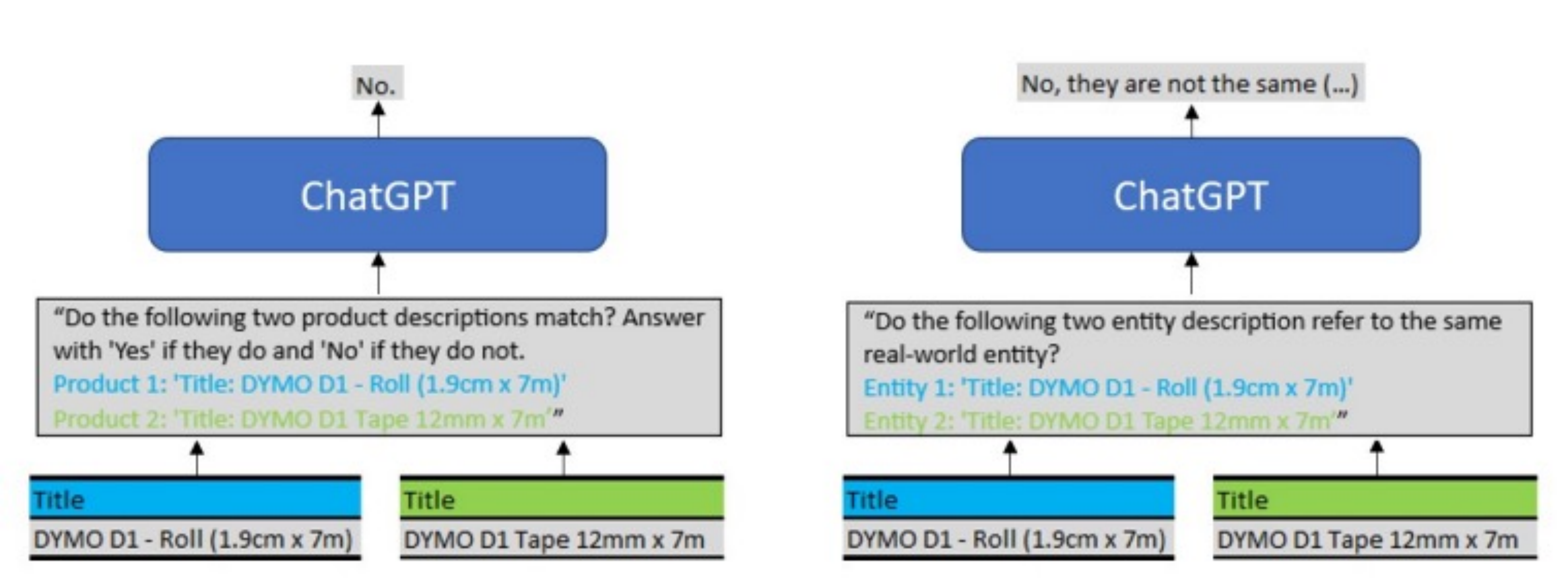


Methods	Multiple Choice		Binary Classification					
	e-CARE	COPA	e-CARE			COPA		
	Full	Full	Pos	Neg	Full	Pos	Neg	Full
BERT-Base (Devlin et al., 2019)	75.4	75.4	-	-	-	-	-	-
RoBERTa-Base (Liu et al., 2019)	70.7	80.5	-	-	-	-	-	-
text-davinci-002	78.4	94.4	18.5	**95.2**	56.8	55.6	**92.4**	74.0
text-davinci-003	76.7	93.2	41.0	86.4	63.7	80.4	82.3	**81.4**
gpt-3.5-turbo	79.1	96.3	75.5	66.9	71.2	96.3	43.2	69.7
gpt-4	**84.5**	**98.1**	**84.8**	57.5	**71.2**	**97.9**	38.5	68.2



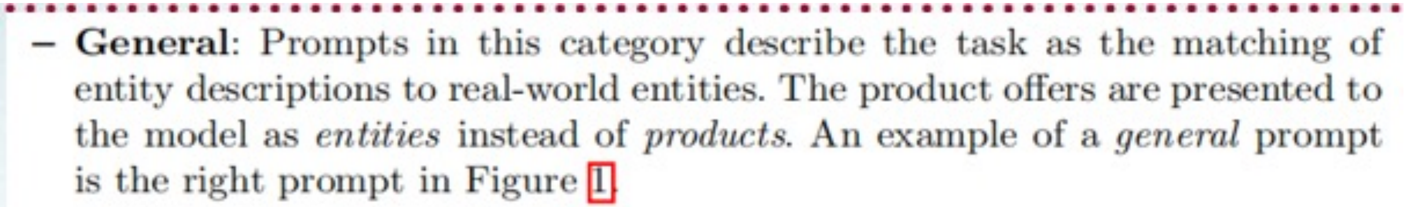
Methods	e-CARE		
	AVG-BLEU	ROUGE-l	Human Evaluation
GRU-Seq2Seq (Chung et al., 2014)	18.7	21.3	0.0
GPT2 (Radford et al., 2019)	32.0	31.5	20.0
text-davinci-003	10.55	37.49	83.0
gpt-3.5-turbo	7.32	40.31	82.0
gpt-4	6.47	39.77	85.0
Human Generation (Du et al., 2022)	35.51	33.46	89.5

Using ChatGPT for Entity Matching

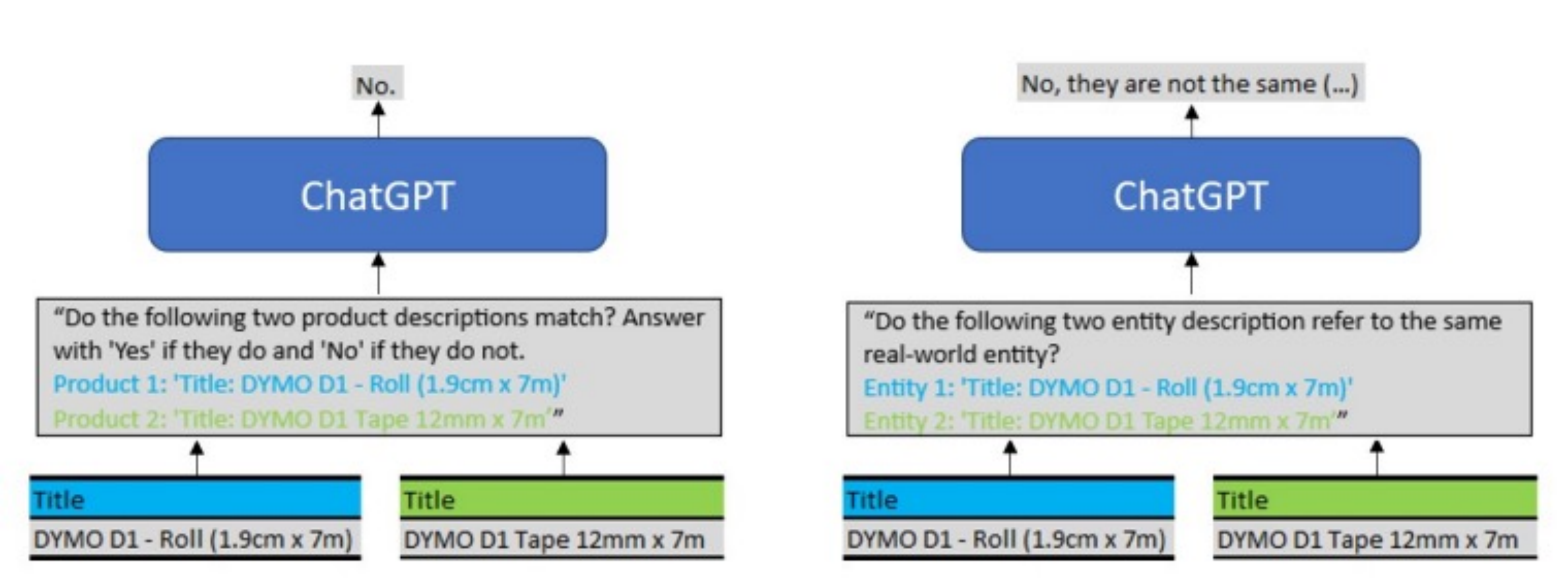

No.
ChatGPT
Title
DYMO D1 - Roll (1.9cm x 7m)
Title
DYMO D1 Tape 12mm x 7m
No, they are not the same (...)
ChatGPT
Title
DYMO D1 - Roll (1.9cm x 7m)
Title
DYMO D1 Tape 12mm x 7m

“Do the following two product descriptions match? Answer with 'Yes' if they do and 'No' if they do not.
Product 1: 'Title: DYMO D1 - Roll (1.9cm x 7m)'
Product 2: 'Title: DYMO D1 Tape 12mm x 7m'”

“Do the following two entity description refer to the same real-world entity?
Entity 1: 'Title: DYMO D1 - Roll (1.9cm x 7m)'
Entity 2: 'Title: DYMO D1 Tape 12mm x 7m'”



- **General**: Prompts in this category describe the task as the matching of entity descriptions to real-world entities. The product offers are presented to the model as *entities* instead of *products*. An example of a *general* prompt is the right prompt in Figure 1.



No.

ChatGPT

“Do the following two product descriptions match? Answer with 'Yes' if they do and 'No' if they do not.
Product 1: 'Title: DYMO D1 - Roll (1.9cm x 7m)'
Product 2: 'Title: DYMO D1 Tape 12mm x 7m'”

Title
DYMO D1 - Roll (1.9cm x 7m)

Title
DYMO D1 Tape 12mm x 7m

No, they are not the same (...)

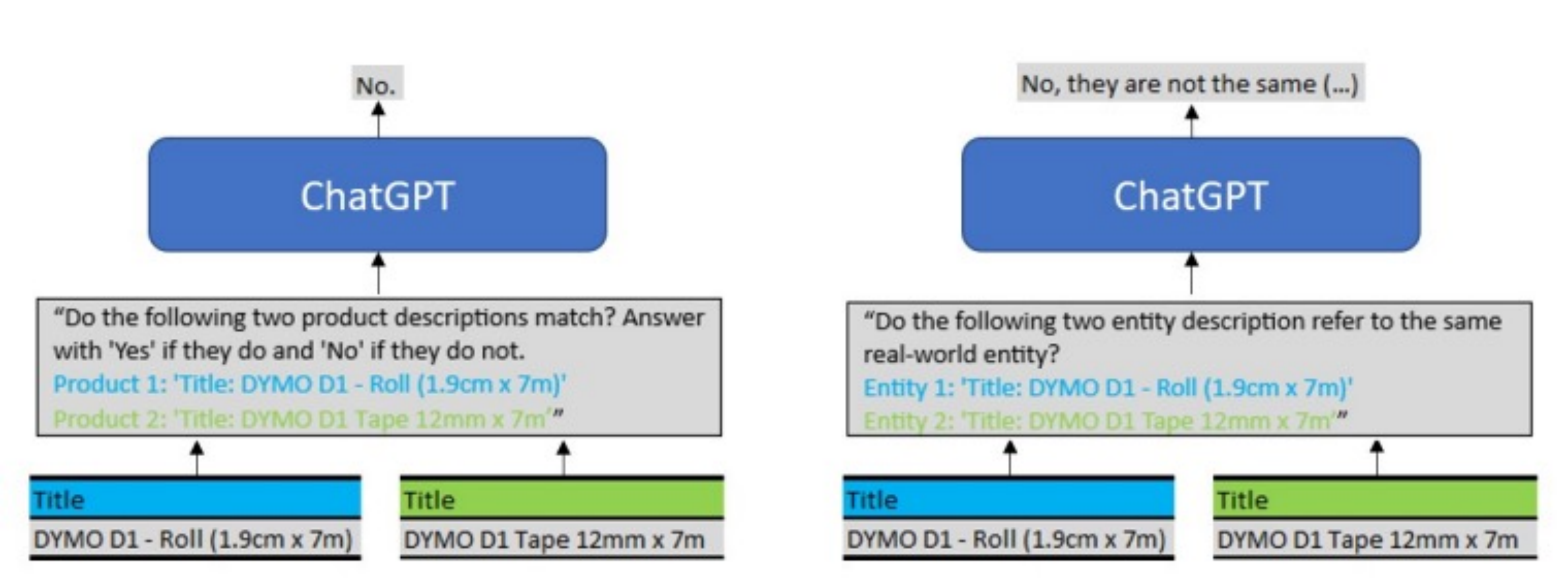
ChatGPT

“Do the following two entity description refer to the same real-world entity?
Entity 1: 'Title: DYMO D1 - Roll (1.9cm x 7m)'
Entity 2: 'Title: DYMO D1 Tape 12mm x 7m'”

Title
DYMO D1 - Roll (1.9cm x 7m)

Title
DYMO D1 Tape 12mm x 7m

– **Domain**: In comparison to *General*, the domain-specific prompt describes the task as matching of product descriptions and refers to the examples as *product offers*. An example of this kind of prompt is the left prompt in Figure 1



No.

ChatGPT

"Do the following two product descriptions match? Answer with 'Yes' if they do and 'No' if they do not.
Product 1: 'Title: DYMO D1 - Roll (1.9cm x 7m)'
Product 2: 'Title: DYMO D1 Tape 12mm x 7m'"

Title
DYMO D1 - Roll (1.9cm x 7m)

Title
DYMO D1 Tape 12mm x 7m

No, they are not the same (...)

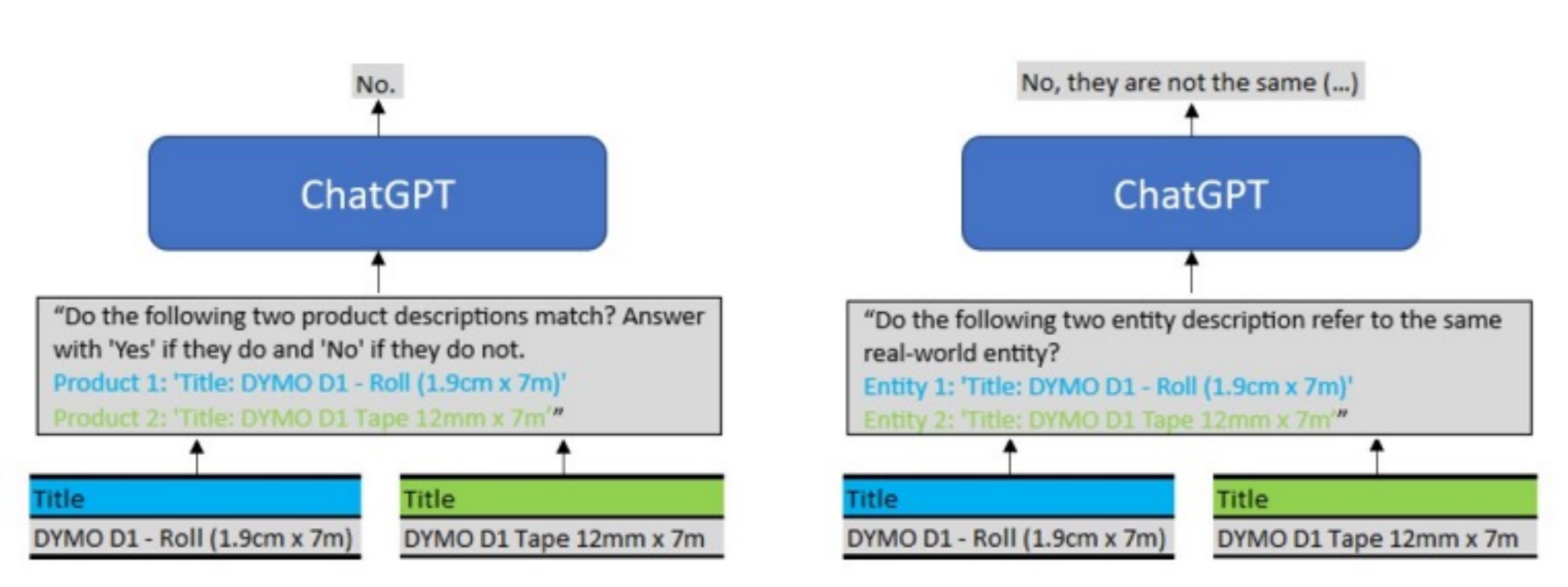
ChatGPT

"Do the following two entity description refer to the same real-world entity?
Entity 1: 'Title: DYMO D1 - Roll (1.9cm x 7m)'
Entity 2: 'Title: DYMO D1 Tape 12mm x 7m'"

Title
DYMO D1 - Roll (1.9cm x 7m)

Title
DYMO D1 Tape 12mm x 7m

- **Complex**: Prompts in this category use more complex language when stating the question, specifically they use the formulations "refer to the same real-world product" or "refer to the same real-world entity". An example is the right prompt in Figure 1.



No.

ChatGPT

“Do the following two product descriptions match? Answer with 'Yes' if they do and 'No' if they do not.
Product 1: 'Title: DYMO D1 - Roll (1.9cm x 7m)'
Product 2: 'Title: DYMO D1 Tape 12mm x 7m'”

Title
DYMO D1 - Roll (1.9cm x 7m)

Title
DYMO D1 Tape 12mm x 7m

No, they are not the same (...)

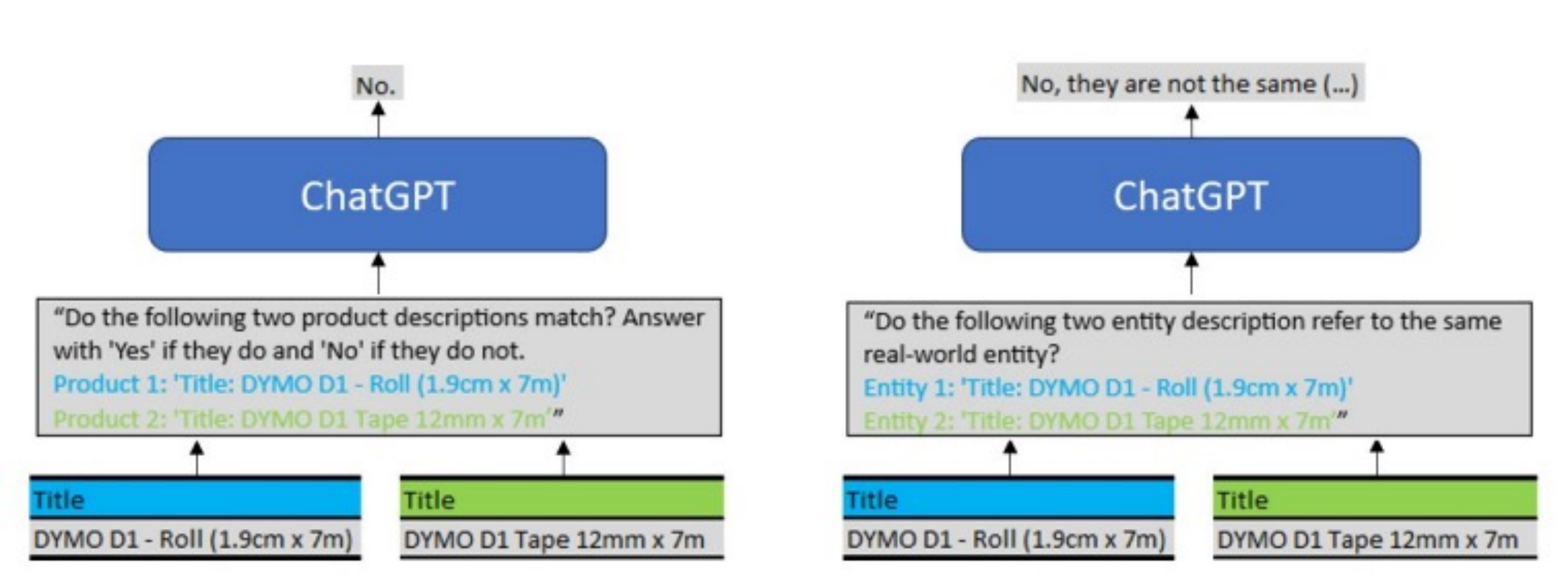
ChatGPT

“Do the following two entity description refer to the same real-world entity?
Entity 1: 'Title: DYMO D1 - Roll (1.9cm x 7m)'
Entity 2: 'Title: DYMO D1 Tape 12mm x 7m'”

Title
DYMO D1 - Roll (1.9cm x 7m)

Title
DYMO D1 Tape 12mm x 7m

– **Simple**: This kind of prompt uses more simple language and replaces the formulations from *Complex* with a simple ”match”. An example is the left prompt in Figure 1.



No.

ChatGPT

“Do the following two product descriptions match? Answer with 'Yes' if they do and 'No' if they do not.
Product 1: 'Title: DYMO D1 - Roll (1.9cm x 7m)'
Product 2: 'Title: DYMO D1 Tape 12mm x 7m'”

Title
DYMO D1 - Roll (1.9cm x 7m)

Title
DYMO D1 Tape 12mm x 7m

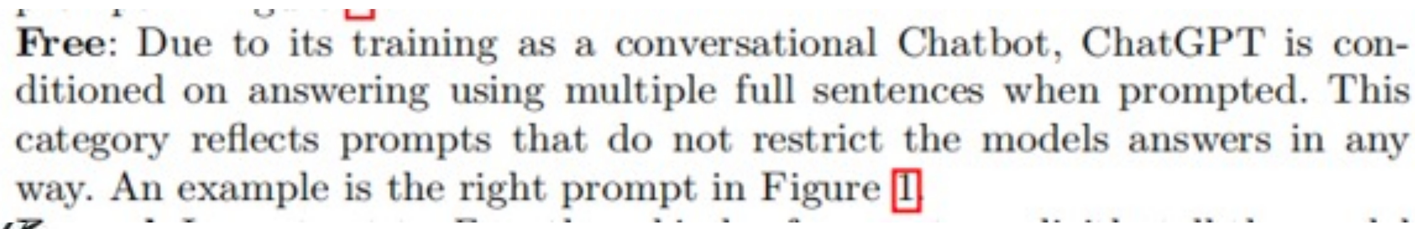
No, they are not the same (...)

ChatGPT

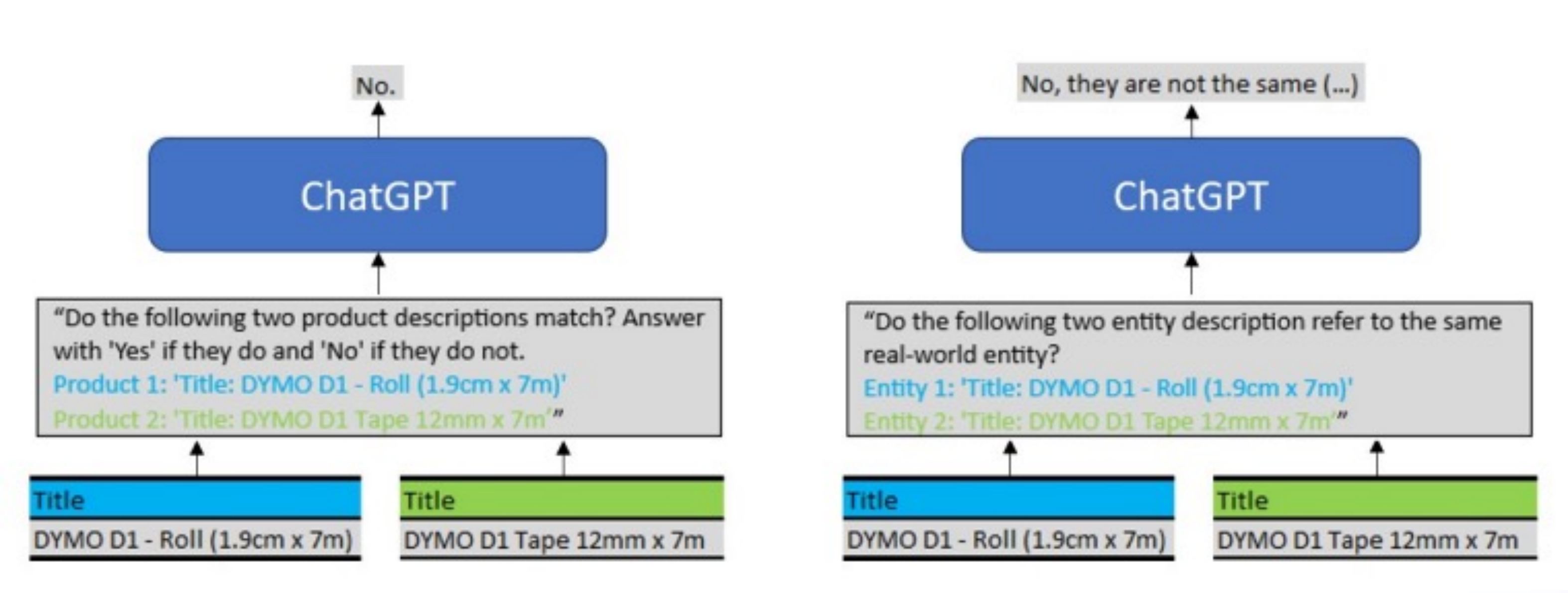
“Do the following two entity description refer to the same real-world entity?
Entity 1: 'Title: DYMO D1 - Roll (1.9cm x 7m)'
Entity 2: 'Title: DYMO D1 Tape 12mm x 7m'”

Title
DYMO D1 - Roll (1.9cm x 7m)

Title
DYMO D1 Tape 12mm x 7m



Free: Due to its training as a conversational Chatbot, ChatGPT is conditioned on answering using multiple full sentences when prompted. This category reflects prompts that do not restrict the models answers in any way. An example is the right prompt in Figure 1.



No.

ChatGPT

“Do the following two product descriptions match? Answer with 'Yes' if they do and 'No' if they do not.
Product 1: 'Title: DYMO D1 - Roll (1.9cm x 7m)'
Product 2: 'Title: DYMO D1 Tape 12mm x 7m'”

Title
DYMO D1 - Roll (1.9cm x 7m)

Title
DYMO D1 Tape 12mm x 7m

No, they are not the same (...)

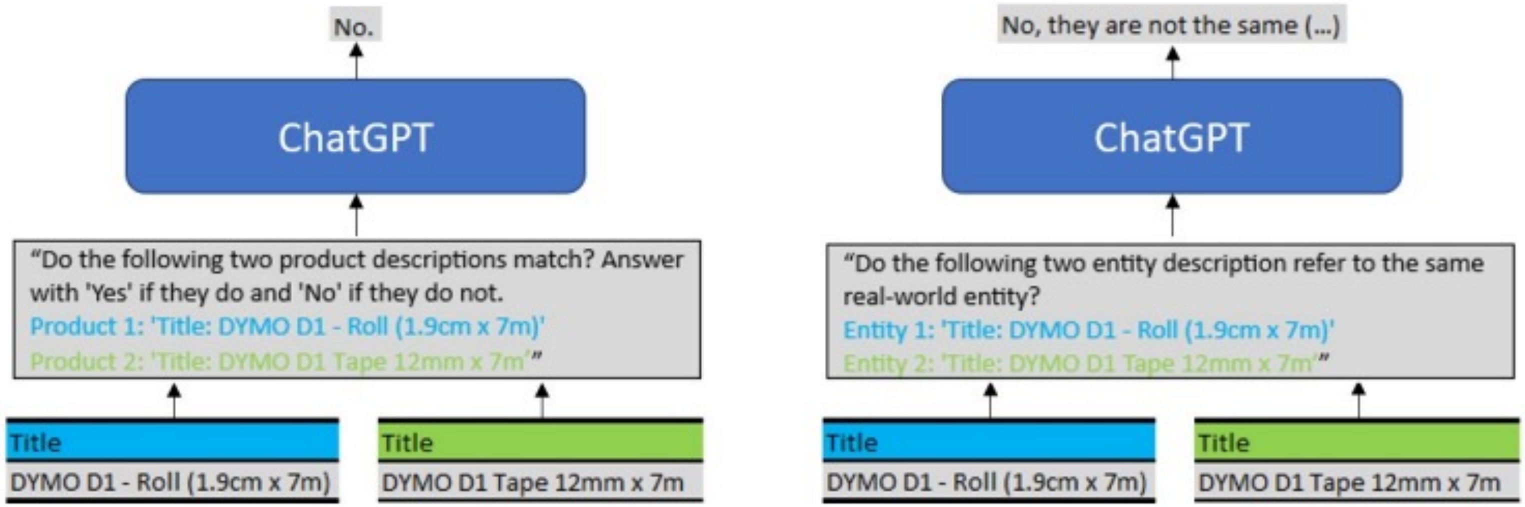
ChatGPT

“Do the following two entity description refer to the same real-world entity?
Entity 1: 'Title: DYMO D1 - Roll (1.9cm x 7m)'
Entity 2: 'Title: DYMO D1 Tape 12mm x 7m'”

Title
DYMO D1 - Roll (1.9cm x 7m)

Title
DYMO D1 Tape 12mm x 7m

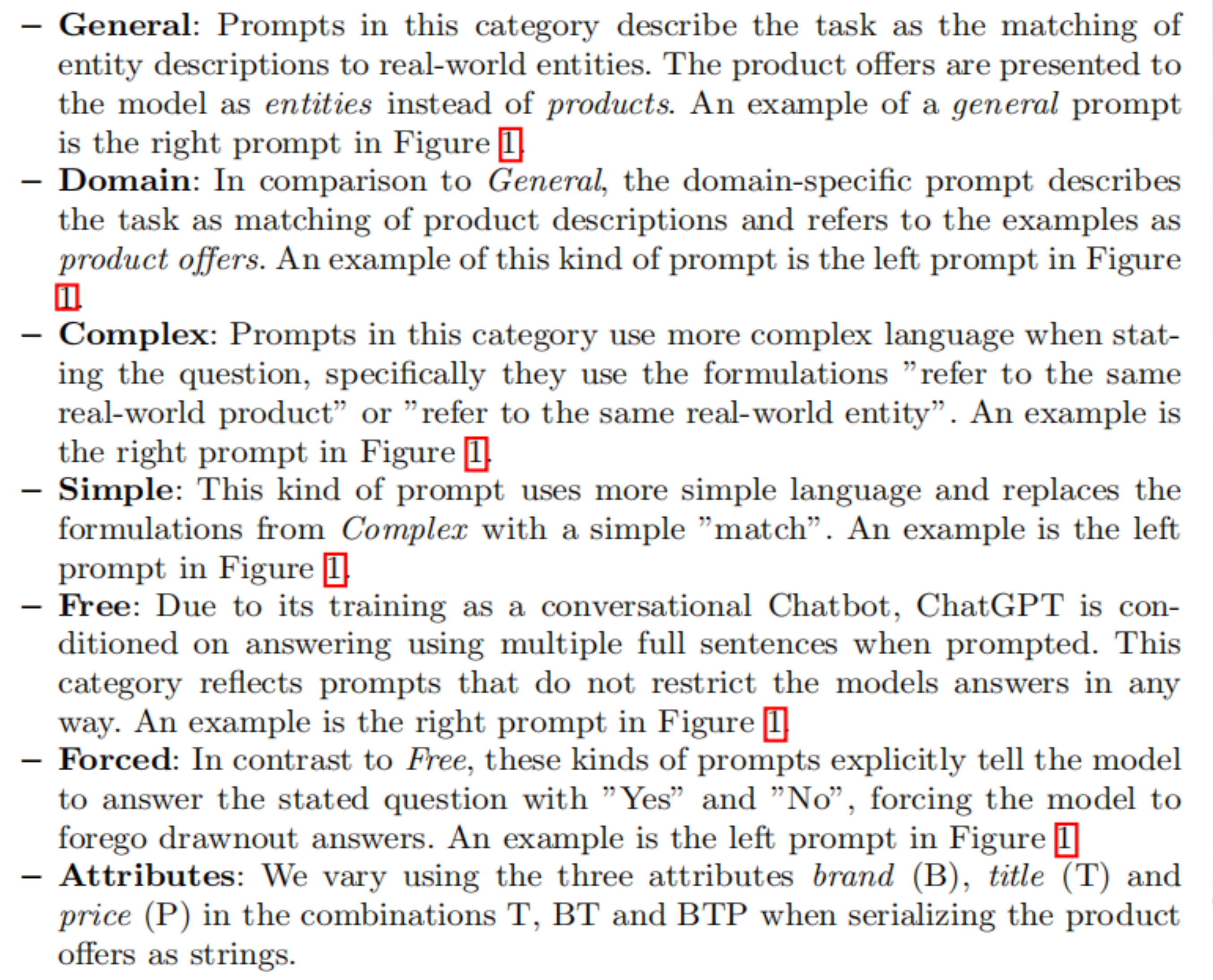
Forced: In contrast to *Free*, these kinds of prompts explicitly tell the model to answer the stated question with "Yes" and "No", forcing the model to forego drawnout answers. An example is the left prompt in Figure 1


No.
ChatGPT
No, they are not the same (...)
ChatGPT
Title
DYMO D1 - Roll (1.9cm x 7m)
Title
DYMO D1 Tape 12mm x 7m
Title
DYMO D1 - Roll (1.9cm x 7m)
Title
DYMO D1 Tape 12mm x 7m

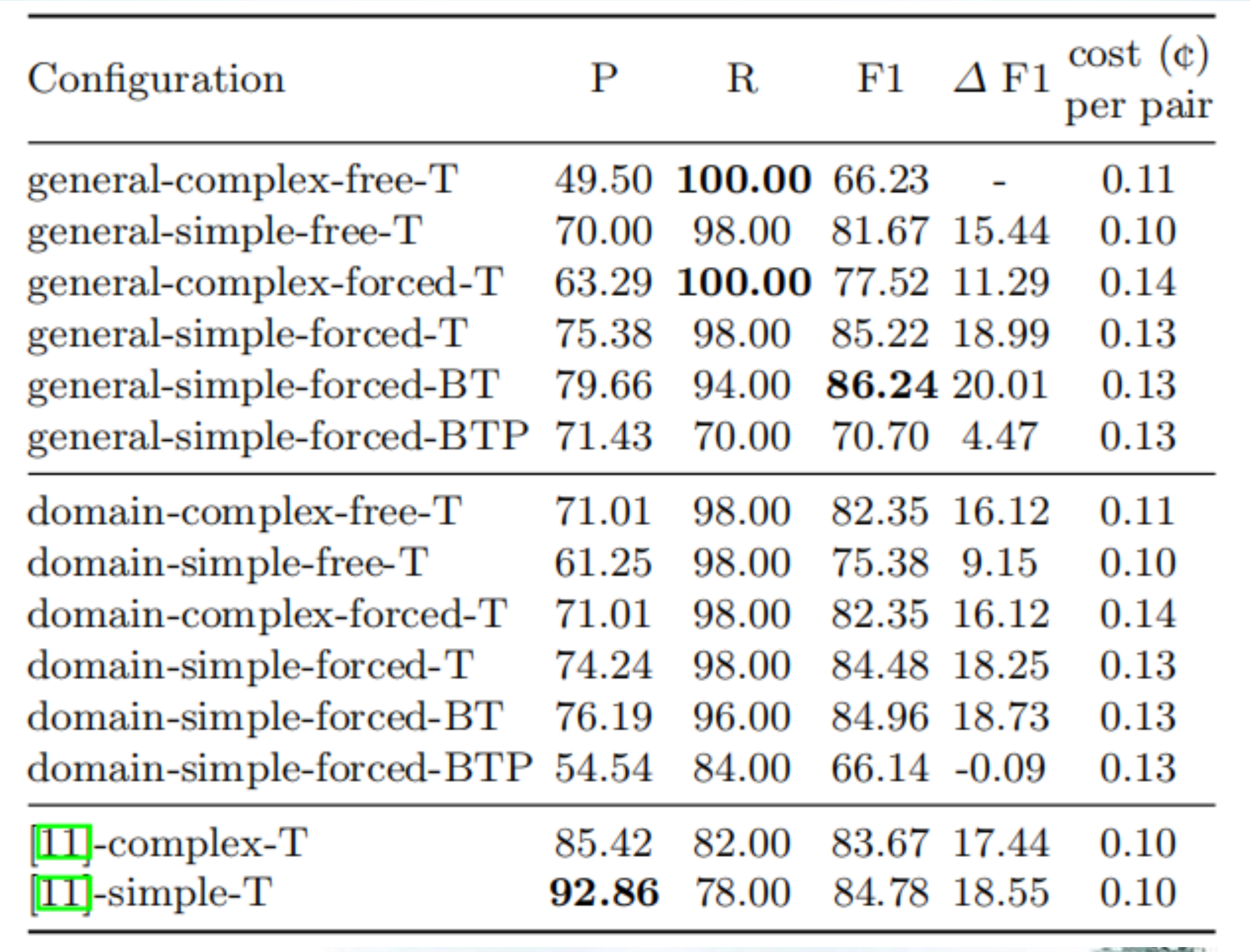
"Do the following two product descriptions match? Answer with 'Yes' if they do and 'No' if they do not.
Product 1: 'Title: DYMO D1 - Roll (1.9cm x 7m)'
Product 2: 'Title: DYMO D1 Tape 12mm x 7m'"

"Do the following two entity description refer to the same real-world entity?
Entity 1: 'Title: DYMO D1 - Roll (1.9cm x 7m)'
Entity 2: 'Title: DYMO D1 Tape 12mm x 7m'"

Attributes: We vary using the three attributes *brand* (B), *title* (T) and *price* (P) in the combinations T, BT and BTP when serializing the product offers as strings.



- **General**: Prompts in this category describe the task as the matching of entity descriptions to real-world entities. The product offers are presented to the model as *entities* instead of *products*. An example of a *general* prompt is the right prompt in Figure 1.
- **Domain**: In comparison to *General*, the domain-specific prompt describes the task as matching of product descriptions and refers to the examples as *product offers*. An example of this kind of prompt is the left prompt in Figure 1.
- **Complex**: Prompts in this category use more complex language when stating the question, specifically they use the formulations ”refer to the same real-world product” or ”refer to the same real-world entity”. An example is the right prompt in Figure 1.
- **Simple**: This kind of prompt uses more simple language and replaces the formulations from *Complex* with a simple ”match”. An example is the left prompt in Figure 1.
- **Free**: Due to its training as a conversational Chatbot, ChatGPT is conditioned on answering using multiple full sentences when prompted. This category reflects prompts that do not restrict the models answers in any way. An example is the right prompt in Figure 1.
- **Forced**: In contrast to *Free*, these kinds of prompts explicitly tell the model to answer the stated question with ”Yes” and ”No”, forcing the model to forego drawnout answers. An example is the left prompt in Figure 1.
- **Attributes**: We vary using the three attributes *brand* (B), *title* (T) and *price* (P) in the combinations T, BT and BTP when serializing the product offers as strings.



Configuration	P	R	F1	Δ F1	cost (¢) per pair
general-complex-free-T	49.50	**100.00**	66.23	-	0.11
general-simple-free-T	70.00	98.00	81.67	15.44	0.10
general-complex-forced-T	63.29	**100.00**	77.52	11.29	0.14
general-simple-forced-T	75.38	98.00	85.22	18.99	0.13
general-simple-forced-BT	79.66	94.00	**86.24**	20.01	0.13
general-simple-forced-BTP	71.43	70.00	70.70	4.47	0.13
domain-complex-free-T	71.01	98.00	82.35	16.12	0.11
domain-simple-free-T	61.25	98.00	75.38	9.15	0.10
domain-complex-forced-T	71.01	98.00	82.35	16.12	0.14
domain-simple-forced-T	74.24	98.00	84.48	18.25	0.13
domain-simple-forced-BT	76.19	96.00	84.96	18.73	0.13
domain-simple-forced-BTP	54.54	84.00	66.14	-0.09	0.13
[11]-complex-T	85.42	82.00	83.67	17.44	0.10
[11]-simple-T	**92.86**	78.00	84.78	18.55	0.10